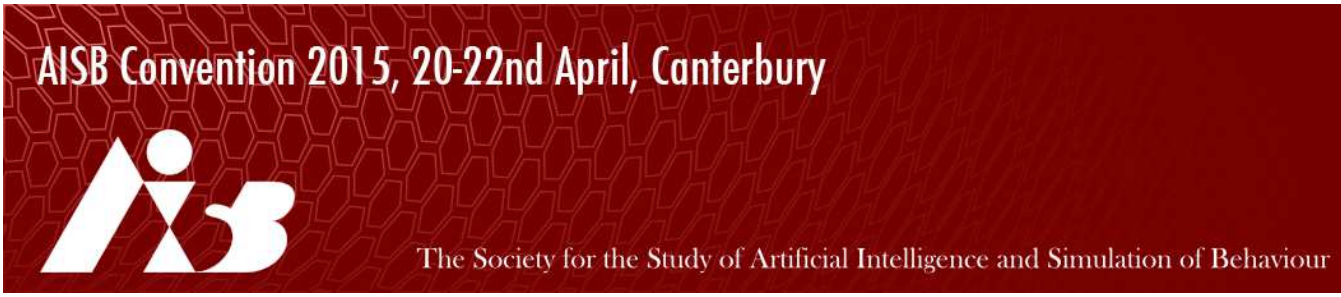


AISB Convention 2015

University of Kent, Canterbury

Proceedings of the AISB 2015 Symposium on AI and Games

Edited by Daniela Romano and David Moffat

ISBN: 978-1-908187-47-5

Introduction to the Convention

The AISB Convention 2015—the latest in a series of events that have been happening since 1964—was held at the University of Kent, Canterbury, UK in April 2015. Over 120 delegates attended and enjoyed three days of interesting talks and discussions covering a wide range of topics across artificial intelligence and the simulation of behaviour. This proceedings volume contains the papers from the *Symposium on AI and Games*, one of eight symposia held as part of the conference. Many thanks to the convention organisers, the AISB committee, convention delegates, and the many Kent staff and students whose hard work went into making this event a success.

—Colin Johnson, Convention Chair

Contents

Modelling Cultural, Religious and Political Affiliation in Artificial Intelligence Decision-Making

Mark R Johnson[1]

Abstract. This paper examines cutting-edge work in the generation of individual AI actors who behave according to procedurally-generated social, cultural, political and religious norms. Based on the author's ongoing development of the game *Ultima Ratio Regum* (URR) – built with a hand-made game engine in Python – the paper explores three core aspects of URR's AI actors. Firstly, the generation of a full world population of AI actors and ensuring that they are distributed appropriately and logically for a culturally-varied world; secondly the procedural generation of densely complex religious, political, cultural and socially normative values to assign to these AI actors, and how their decision-making processes are determined by these allegiances; and thirdly and lastly how this game, among other objectives, seeks to forward what I term "qualitative AI" where culture and society, not pathfinding and "optimal" decision-making, are the primary determinants of behaviour. The paper concludes with a summary of both these three points and the future plans for the game's AI systems.

1 INTRODUCTION

This exploratory paper is based on the author's own work, having been for the last three years the sole developer of the roguelike game "*Ultima Ratio Regum"* (URR). Set during the Scientific Revolution, almost everything within the game is procedurally generated – this ranges from the "macro" level of 2000+ years of detailed history, historical figures, empires and nations, religions, wars, dozens of vast cities and a vast world population of procedurally-generated non-player characters (NPCs) unique to each playthrough, to the "micro" level of individual towns and cities, individual NPCs, specific buildings and items, and flora and fauna. Inspired by the works of Umberto Eco and Jorge Luis Borges, the game is an exploration of a number of themes including historiography and the writing of the historical record, metanarrative and political ideology, and the philosophical idealism of George Berkeley. Most crucially the work aims to specifically integrate this "thematic" content with the game's mechanics, rather than leaving such content as "background" or "lore" that the player can take or leave. Much of this will be achieved through the use of innovative AI actors currently being developed at time of writing. Every aspect of the behaviour of these actors – their greetings, their insults, their dress, their farewells, their behaviour in challenging situations, their reaction to those from other nations, and much else – is procedurally generated, and fore-grounded in their decision-making algorithms. It is these actors and the roles they play which this paper focuses upon, and the break they represent from much traditional AI research into decision-making optimization [1] and pathfinding [2].

[1] Science & Technology Studies Unit, Dept. of Sociology, Univ. of York, YO10 5DD, UK. Email: mrj503@york.ac.uk.

2 PROCEDURAL GENERATION AND ARTIFICIAL INTELLIGENCE

Firstly, the paper explores the procedural content generation of the game, and how this affects the AI actors. An "average" generated URR world has a population of approximately ten million NPCs. Naturally for such numbers, the management of these NPCs takes place at a number of different "levels" in the game depending on the player's activities – many of the NPCs are "abstracted out" at any given time. The game also contains a system which identifies the most "important" NPCs and ensures that their actions and decisions are always simulated regardless of the player's location (roughly 500-700 NPCs on average are considered "important" by the game's algorithm at any one time, and their actions are carried out constantly, unless one falls below the metric for "importance", at which point that actor is then abstracted out once more). NPCs within the game vary according to a significant range of variables: according to their race, language, cultural background and cultural norms, sex and gender, age, political alignment, religious beliefs (if any), national citizenship, and interests and agendas. A rough calculation currently suggests that there are over 1 trillion possible AI actors that may be procedurally generated within the game world who will, crucially, behave differently according to the social and cultural context within which they are generated. The agendas of these actors (returned to later in this abstract) are largely dependent on their cultural and religious backgrounds, leading to a densely complex world within which the player will uncover information about religious feuds, cultural differences, long-standing war bitterness, language difficulties, and many similar concepts of a sort not normally explored in games. The paper will therefore examine the generation of the AI actors from a creative standpoint; the management of so many AIs from a technical standpoint; and the integration of the two into a culturally and socially variegated and dense world.

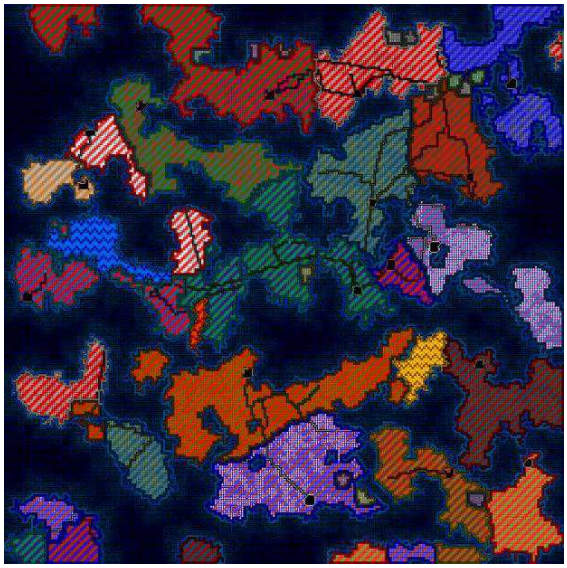


Figure 1. Example of Generated Political Divisions

Secondly, the paper explores how this content generation creates a deeply complex environmental simulation, arguably one of the most detailed and dense generated worlds ever created in a game. As above, the agendas of these AIs are dependent on the procedural generation of their *origins*. A generated URR world contains approximately forty civilizations (Figure 1) designed to emulate the massive variety in real-world civilizations from this historical era. Some may be nomadic desert peoples who travel in lengthy caravan routes across the world, or hunter-gatherer tribes in close to the Arctic Circle who construct their buildings from ice and stone and have limited trading relations with a nearby civilization, or feudal civilizations who range from the imperialist and the expansionist to the protectionist or isolationist, and have widely differing cultural preferences on issues such as aesthetics, slavery, gladiatorial sport, ethics and morality, and so forth. This variety extends into other areas, such as religion, where a complex algorithm can procedurally create over a million detailed religions with information about their beliefs, their god(s), what festivals or special events are on their religious calendar, their relationships with other religions, their presence in civilizations, eschatological and creation beliefs, the appearance of their altars, expectations from worshippers, etc. All of these "cultural actors" inform the creation of the AI actors who exist *within these contexts*. Crucially, therefore, rather than presenting this civilizational/cultural/religious detail as "background" or "lore" as many games do, they are foregrounded in the AI actors, whose motivations, interests and agendas can only be understood via a detailed understanding of the generated cultural backgrounds from which they originate. In turn, this affects their willingness to interact with the player, to assist or communicate with the player, and to potentially oppose the player if the player has aligned themselves with religions or cultures inimical to those of other NPCs. At the same time, it is a game of incomplete information [cf 3] where both the player, and NPCs, must make judgements about the opinions of others based on the data they possess. The actions of AIs are dependent upon the social conditions and expectations into which they are "born", and therefore strongly differentiate between all the procedurally-generated AI actors in a given instance of the game. Equally, the greater the knowledge the player has attained about the world's culture, the more able the player is to make their wishes felt within the game world.

3 TOWARDS QUALITATIVE AI

Thirdly, the paper brings these together to explore the use of this integration of procedural generation and sociological concepts as a method for game-based learning in the fields of philosophy, sociology, and the humanities more generally. AIs respond and behave according to their political, cultural, social and religious affiliations, and this transforms these concepts in the social sciences into gameplay mechanics that affect the behaviour of AI and the world the player explores, rather than simply a method for *constructing* a game world which then has no further impact upon the player's experience. This is in part akin to the world by Mateas on "expressive AI" [4] and Gruenworldt and Katchabaw's "Realistic Reaction System" [5] but develops it into further qualitative and social science domains, and integrates far broader "relationship" structures of religions and cultures into the interpersonal dimension previous focused upon. The paper therefore explores how the game depicts the influence of these many factors on social interaction, and how these influences are represented in the actions, decisions and interests of the game's AI. In turn, this leads to game-based learning where understanding the cultural, political and religious motivations of AI actors is actually essential to success or failure within the game world. Lastly, this also serves to illustrate the potential for the development of 'qualitative' game mechanics in video games more generally, and highlights the potential for the use of complex AI actors in moving away from the ubiquitous stat-based gameplay of levels, items, rewards, and so forth, and towards developing "AI" that can be understood in terms of their as full actors with a range of interest and agendas, rather than as only actors in combat or strategy situations.

4 CONCLUSIONS

The paper explores three central components to the game's AI – the emphasis on procedural content generation and the integration between that and artificial intelligence; the emphasis within this on creating cultures, societies and religions, and having these directly influence AI decisions; and thirdly the potential for this game to develop "qualitative AI" and to create gameplay mechanics based on political, sociological and humanist concepts rarely explored in interactive media. It notes the potential educational and pedagogic value of these, the potential for new forms of gameplay rarely explored in computer games, and the paper lastly notes the planned future developments of the game's in-development system.

REFERENCES

[1] G. Chaslot, S. Bakkes, I. Szita and P. Spronck. Monte-Carlo Tree Search: A New Framework for Game AI. In: *Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference*. AAAI (2008).

[2] T. Standley. Finding Optimal Solutions to Cooperative Pathfinding Problems. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*. AAAI (2010).
[3] D. Billings, A. Davison, J. Schaeffer, D. Szafron. The Challenge of Poker. In *Artificial Intelligence*, 134:1-2:201-240 (2002).
[4] M. Mateas. Expressive AI: Games and Artificial Intelligence. *edings of Level Up: Digital Games Research Conference* (2003).
[5] L. Gruenwoldt and M. Katchabaw. Creating Reactive Non Player Character Artificial Intelligence in Modern Video Games. In *Proceedings of the 2005 GameOn North America Conference* (2005).

Discerning Human and Procedurally Crafted Content for Video Games

Tommy Thompson[1] and **Rob Watling**[2]

Abstract. We discuss the results of a preliminary study where participants discern between human and computationally crafted content for a video game. Participants were tasked with completing a portion of the game with the knowledge that segments were created either by a procedural generation algorithm or by a game designer. When asked to discern which segments were built by humans and vice versa, overall accuracy of participant guesses is relatively low. However, rationale reached by participants in making these conclusions leads to some interesting discussion about expectations of procedural generation systems and requirements for future studies.

1 Introduction

Procedural Content Generation (PCG) is a popular design paradigm found in video game development. While the origins of this method can be found in the likes of *Elite* [3] to overcome hardware limitations, the emphasis has shifted towards experimentation and challenge. This is typified by the *Borderlands* series [5]: where weapons and tools are presented for the player to discover, adopt or discard based upon personal preference. Meanwhile, *Diablo* [4] and *Spelunky* [21] adopt PCG for map generation in an effort to retain variety, novelty and challenge for even the most seasoned of players

If we consider this transition of the role of PCG systems, what is most interesting is that players perception of in-game content is becoming of greater focus. As problem scope increases, developers place a stronger emphasis on ensuring content is as interesting as it is varied. This has resulted in significant work in Artificial Intelligence (AI) to create intelligent PCG processes [19], with efforts to create 'custom' and more bespoke content [6, 20] and tools to aid the development process [8].

In this paper, we discuss preliminary work in generating content for an 'endless runner' game entitled *Sure Footing*[3]. The game tasks players with navigating a hazardous environment for as long as possible. Players are presented an early build of the game that carries content designed both by the developers and an early build of a PCG system. The task for participants was to identify the human-built and PCG samples and give a rationale for why they reached their conclusion. Our hypothesis was that if we were to base our PCG system on a meta-creative approach; adopting principles from a human designer, that players by-and-large would struggle to identify any key differences.

[1] University of Derby, UK, email: tommy@t2thompson.com
[2] University of Derby, UK, email: therobwatling@gmail.com
[3] A game being developed by Table Flip Games Ltd.: `http://www.tableflipgames.co.uk`

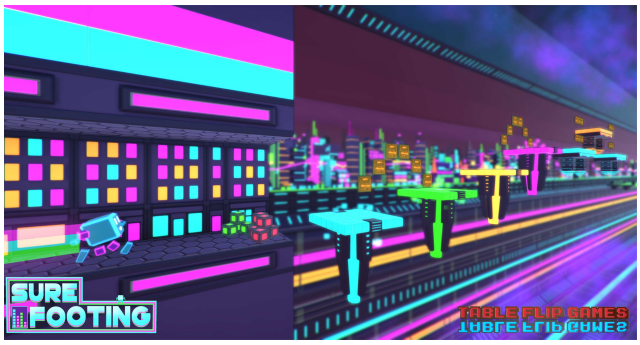


Figure 1. A screenshot of the *Sure Footing* video game, where the player, represented by a blue cube, must navigate a series of platforms and environmental hazards.

2 Sure Footing & Endless Runner Games

Sure Footing, shown in Figure 1 is an 'endless runner', where the player must navigate through a hazardous environment for as long as possible. Player's must traverse a collection of platforms and avoid obstacles placed upon them whilst evading an enemy that is following them throughout. Should the player fail a jump between platforms or be captured by their pursuer, the game will restart from the beginning of the current segment of play.

The endless runner genre is an effective platform for experimenting in PCG given that players are seldom aware of what is ahead of them. This allows for sudden change to the world that the player must adapt to. This is part of the novelty and charm that drove the popularity of seminal endless runner *Canabalt* [13] and subsequently titles such as *Flappy Bird* [10], and *Temple Run* [7].

Endless runners have a difficult balance to attain due to their unpredictable nature: should changes prove too sudden, players may subsequently lose interest. Ultimately, it is crucial that players feel the challenge of the game comes from their own ability to master game mechanics, rather than unfair design of the game. Equally players should be able to understand how to proceed through the game, irrespective of whether particular 'chunks' of level design have previously been seen in play. As discussed in Section 5, we place an emphasis on difficulty and progression in each participant's playthrough.

3 Related Work

Arguably the most established research in PCG for platforming games can be found in the *Mario AI Competition* which ran from 2009 to 2012 and has since been succeeded by the *Platformer AI Competition*[4]. The competition is dependent upon participants adopting a clone of the popular *Super Mario Bros.* [11] series. While originally intended to focus on gameplay, a level generation track was introduced in 2010 [18], with each entrant required to adopt player data from the an initial test level [14]. While the emphasis is to generate an intelligent and customised level generator, the focus of the competition is to find levels that judges deem 'interesting', rather than accurately reflect the designs of the *Super Mario Bros.* series. As such, the competition refrains from having judges compare PCG levels to original *Super Mario* levels built by human designers.

This work, among others in the AI field, focusses on search-based procedural generation. While this is an intelligent process that aims to create customised and unique content, there is seldom any emphasis on modelling the creative processes adopted by human designers in game development [2]. There have been notable exceptions to this, with one of the most prominent examples being the 'Sentient Sketchbook' project. As detailed in [8, 12, 9], this project carries a stronger emphasis on the use of PCG for human-designers as a tool; allowing for intelligent and useful content to be created in line with a designers expectations and habits.

The inspiration for this project is the *Tanagra* project detailed in [17]: a mixed-initiative design tool that aids in the creation of levels for 2D platformer games. The system allows for a designer to establish a timeline of 'beats': setting the pace of gameplay. The first phase of this work detailed in [15] is adopted in this project, where levels are built courtesy of rhythm groups which establish activities that take place.

4 System Design

In this section we give a brief overview of the PCG system adopted for this experiment. As we continue to discuss the design behind this system, we adhere to the taxonomy for PCG techniques defined in [19].

As noted in Section 3, our level generator adopts the rhythm approach discussed in [15]. The generator adopts a generate and test approach: creating and refining the rhythm of play followed by the geometry. The rhythm generator is comprised of a grammar representing player actions. This is encompassed by what is referred to as a *sprint*, a vector of game actions that lasts no longer than 60-90 seconds in-game. Actions are constrained to particular durations, denoted as *short* (≤ 1 second), *normal* ($1-3$ seconds) or *long* ($3-5$ seconds). A full list of all available actions can be found in Table 1.

Once a full sprint vector is established, a critic will briefly evaluate to ensure a sense of flow is retained: the critic may swap pairs of activities, or add segments to give players a brief respite. This vector is passed into the geometry generator to create the level for play. This geometry generator is responsible not only for the selection of geometry but its subsequent placement within the game scene.

Each of the activities identified in Table 1 have one or more prefabricated pieces of geometry, hereby referred to as *prefabs*, that effectively represent the intended behaviour from the player. An example of this can be seen in Figure 2, which is one of the 'hopscotch' prefabs. The geometry generator places these items into the scene, aligning them such that a complete level is constructed. Once a sprint is completed, a 'rest' prefab is placed into the world. Typically this whole procedure is an online process and takes place during play. However, as discussed in Section 5, this process is made offline for the duration of this experiment.

[4] `http://www.platformersai.com/`

Figure 2. One of the prefab geometry pieces adopted by the geometry generator for the 'hopscotch' activity in Table 1.

5 Experiment Design

Our experiment was conducted during the *GameCity* festival in Nottingham, UK[5]. The focus of the experiment was to determine whether users could differentiate between levels crafted by a prototype PCG system, versus levels designed by one of the authors. In an effort to prepare for the festival, we exported six levels from the PCG system and stored them for later use. In addition to the PCG levels, six levels of equivalent length were crafted in the game engine by one of the authors.

While each level that was designed was unique, there are similarities that can be seen throughout. This is in part due to the prefabs discussed in Section 4 which were adopted in all level creation. In addition, given that the PCG system detailed in Section 4 was written by one author, with the other responsible for building the human levels, there is an argument to be made in that design habits of the authors have been injected, albeit rigidly, into the rhythm system. We return to these points in Section 6.1 and note the limitations they present as well as future steps for improvement.

Table 2. A breakdown of the percentage of participants who guessed either human or PCG-crafted level after each stage of completion. Followed by the success rates of those guesses at that particular stage.

Breakdown of Designer Guesses			
Level	Human Level	PCG-Level	Unsure
1	63.15%	23.7%	13.15%
2	50%	23.7%	26.3%
3	28.9%	42.1%	29%
Success Rates			
1	71.43%	25%	N/A
2	92.86%	12.5%	N/A
3	28.57%	37.5%	N/A

Each play-through of *Sure Footing* comprised of three 'levels'. With a minimum of one human and one PCG-crafted level per playthrough. The third and final level was selected at random from the

[5] The festival took place during 25th October to 1st November 2014: `http://www.gamecity.org`

Action	Duration	Description
Run	Short, Normal, Long	A flat section of terrain which the player must run across.
Jump	Short	A gap between platforms which may carry a variation in height, such that can either jump or fall depending upon the context.
Incline	Normal	A series of short platforms closely placed to one another or a ramp that gradually increases in height.
Decline	Normal	A series of short platforms closely placed to one another or a ramp that gradually decrease in height.
Hopscotch	Normal	A series of short platforms with one in the middle that is higher than the others, forcing the player to hop atop or over it.
Fall	Normal, Long	Two platforms with separated by a significant vertical drop. Players are expected to fall or jump down to the lower platform.
Spring	Normal, Long	A long platform with a spring attached to the end that will launch the player to a much higher platform.

Table 1. The collection of actions that can take place in a given 'sprint' of play.

Table 3. A table showing the frequency of reasons left by participants. Including the percentage of responses that left a given reason, followed by a breakdown with respect to whether they guessed a level was human or PCG-crafted.

	Reasons For Decision						
	Difficulty	**Pace**	**Variety**	**Length**	**Item Placement**	**Don't Know**	**Other**
All Responses	35.09%	36.84%	29.82%	14.91%	29.82%	7.89%	8.77%
No Vote	0.88%	2.63%	0.88%	0.88%	1.75%	4.39%	3.51%
Decided Human-Crafted Level							
All Guessed Human	18.42%	23.68%	14.04%	10.53%	19.30%	0.88%	0.88%
Correctly Guessed Human	10.53%	7.89%	7.02%	5.26%	7.89%	0.88%	0%
Decided PCG-Crafted Level							
All Guessed PCG	15.79%	10.53%	14.91%	3.51%	8.77%	2.63%	4.39%
Correctly Guessed PCG	7.89%	4.39%	7.02%	2.63%	3.51%	0.88%	1.75%

PCG and human-designed sets, thus certain users would be exposed to each type of content, with one type more-so than the other.

At the beginning of the play-through, players were briefed that they would play at minimum one of each kind of level and that their task was to discern between the two types. Upon completion, the next level was immediately loaded into the game for the player to complete. In the event that players found these levels too challenging, the option was given to allow for a level to be skipped. Players were given as many tries as was necessary to complete the set of three levels. Upon completion, participants we asked if they could identify PCG and human samples; identifying whether level difficulty, pace, variety of rhythm, length and placement of items informed their decision. In addition, players were also given the option to express in detail additional elements that helped cement their opinion. Only after this questionnaire was completed and the game saved performance data was it revealed to users whether a given level was indeed crafted by a human or PCG system.

6 Results & Discussion

The results from 45 participants can be seen in Table 2, showing the breakdown of guesses at each stage of the process. In addition, we provide a breakdown of the frequency that particular reasons were given and their success in Table 3.

There are a number of interesting results, noting not only gradual trends in guessing patterns, but also the reasons given in certain circumstances. Firstly, we note that players were more likely to correctly denote a level as being crafted by a human than by the PCG system. This is perhaps not surprising, given that players would assume by default that content was man-made if they found it fun or engaging. Another interesting element is that not only is the success rate for voting PCG-levels less accurate, but players are more likely to be left unsure in their decision. Despite the level of accuracy behind human guesses, players became less confident over time in voting for a human-designed level, arguably due to not discovering a significant difference in the content that was being shown during gameplay. We believe this could be a limitation of the current generator, given PCG levels may appear remarkably similar to human-crafted content.

If we look further at the feedback from Table 3, it is interesting to note that that pace and difficulty followed by variety and item placement are deemed the biggest factors for making a given decision. Despite this, in certain circumstances this proved to be an incorrect assertion. For example, less than half of all participants who blamed pace for a human-designed level were proven correct. Overall, there does not appear to be a real consensus from this study for understanding whether a level was human or PCG-crafted.

In addition to the provided reasons, there was written feedback that was provided through the 'Other' column of the questionnaire. This yield some equally interesting yet contradictory reasons for participants decisions. Specific written feedback from participants noted that levels were "very good" or "intriguing", with several participants noting "flow" as one of the reasons for human-crafted samples, only to be proven wrong. One participant went so far as to criticise the design of one level, noting that "no human would place" a particular

segment of prefabs together and was correct in that assertion.

We note that the average success rate was 25%, with 29% of participants failing to recognise *any* level successfully. Meanwhile 13% were capable of scoring 100% accuracy, identifying all PCG and human-crafted levels. It is arguably their written feedback or experience that proved most valuable. One participant was an independent game developer who could 'see' the patterns at play. Meanwhile another noted that item placement in particular showed an emphasis on human design. Given blocks and power-ups would be dropped in what they deemed "easier" segments of play. One fact that is not made visible in Table 2 is that in two cases, participants completely ignored the briefing given to them and stated that all levels were man-made. We would argue that part of this challenge in the eyes of players originates in the problem domain. As discussed in Section 2, the endless runners constrain the amount of change available to the designer. In addition, there are still numerous limitations in our system which we will now discuss.

6.1 Study Limitations

While this study does yield some interesting results, there are some notable limitations both with the study as well as the current generation system that we aim to address in future studies.

Firstly, the *Sure Footing* generator is a weak computationally creative system [1]: given it is largely reliant upon the pre-conceived notions of the human authors. Art assets are stored in pre-built chunks the system is reliant upon and the generator is not overly flexible. As such, any level built will carry heavy influences from human designers. More importantly, this generator was not particularly expressive, with only differing configurations of one base level 'template' that could be achieved. While the range of expression permitted to the generator must be improved, relating back to our previous point, future studies must also focus on measuring the full expressivity of the system. This notion, as discussed in [16], can help us identify the range of content the generator can establish and subsequently what impact this has on player perceptions. In addition, this would allow for assessment of whether current generators can build the same range of content as a human designer.

Furthermore, future studies would benefit from multiple generators for players to consider: ranging from humans, to intelligent procedural generations systems, with a variety of purely random generators in between. Lastly, future studies would benefit from testers being able to identify particular areas of gameplay where their suspicions of PCG or human-driven design are raised.

7 Conclusion

In this paper we highlighted a short study assessing players perceptions of procedurally generated versus human-crafted content for an endless-runner game. Players proved more successful in identifying human-crafted content than one by a PCG system, which in some respects is a positive step for the level generator; given that the majority of players could not find any patterns or trends that identified a given sample as procedurally generated. Given that this generator is influenced by a human creative process, it is perhaps to be expected that players find it harder to identify PCG-crafted levels. However, when we consider that the PCG system is rather rigid in this current version, it is surprising that the majority of users do not identify any real differences.

The feedback from this process has been adopted by the *Sure Footing* team who aim to build an improved level generator. Future work is focussed on building a more intelligent solution, in addition to addressing the issues raised in Section 6.1, such that a second study may be conducted over a longer period. This would allow for richer discussion of players perceptions of procedurally generated content as the generator becomes more expressive and their restrictions lifted.

ACKNOWLEDGEMENTS

The authors would like to thank the *Sure Footing* development team: Jonathan Boorman, Neall Dewsbury, Charlotte Sutherland, Matthew Syrett and James Tatum, for their assistance with this study.

REFERENCES

[1] Mohammad Majid al Rifaie and Mark Bishop, 'Weak and strong computational creativity', in *Computational Creativity Research: Towards Creative Machines*, 37–49, Springer, (2015).
[2] Nuno Barreto, Amílcar Cardoso, and Licínio Roque, 'Computational creativity in procedural content generation: A state of the art survey', in *Proceedings of the 2014 Conference of Science and Art of Video Games)*, (2014).
[3] Bell, I. and Braben, D. Elite. Acornsoft, 1984.
[4] Blizzard North. Diablo. Blizzard Entertainment, 2009.
[5] Gearbox Software. Borderlands. 2K Games, 2009.
[6] Erin J Hastings, Ratan K Guha, and Kenneth O Stanley, 'Evolving content in the galactic arms race video game', in *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, pp. 241–248. IEEE, (2009).
[7] Imangi Studios. Temple Run, 2011.
[8] Antonios Liapis, Georgios N Yannakakis, and Julian Togelius, 'Sentient sketchbook: Computer-aided game level authoring.', in *Proceedings of the 8th International Conference on the Foundations of Digital Games*, pp. 213–220, (2013).
[9] Antonios Liapis, Georgios N Yannakakis, and Julian Togelius, 'Designer modeling for sentient sketchbook', in *Computational Intelligence and Games (CIG), 2014 IEEE Conference on*, pp. 1–8. IEEE, (2014).
[10] Nguyen, D. Flappy Bird. .GEARS Studios, 2013.
[11] Nintendo EAD. Super Mario Bros. Nintendo, 1985.
[12] Mike Preuss, Antonios Liapis, and Julian Togelius, 'Searching for good and diverse game levels', in *Computational Intelligence and Games (CIG), 2014 IEEE Conference on*, pp. 1–8. IEEE, (2014).
[13] Saltsman, A. Canabalt. Semi-Secret Software, 2009.
[14] Noor Shaker, Julian Togelius, Georgios N Yannakakis, Ben Weber, Tomoyuki Shimizu, Tomonori Hashiyama, Nathan Sorenson, Philippe Pasquier, Peter Mawhorter, Glen Takahashi, et al., 'The 2010 mario ai championship: Level generation track', *Computational Intelligence and AI in Games, IEEE Transactions on*, **3**(4), 332–347, (2011).
[15] Gillian Smith, Mike Treanor, Jim Whitehead, and Michael Mateas, 'Rhythm-based level generation for 2d platformers', in *Proceedings of the 4th International Conference on Foundations of Digital Games*, pp. 175–182. ACM, (2009).
[16] Gillian Smith and Jim Whitehead, 'Analyzing the expressive range of a level generator', in *Proceedings of the 2010 Workshop on Procedural Content Generation in Games*, p. 4. ACM, (2010).
[17] Gillian Smith, Jim Whitehead, and Michael Mateas, 'Tanagra: A mixed-initiative level design tool', in *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, pp. 209–216. ACM, (2010).
[18] Julian Togelius, Noor Shaker, Sergey Karakovskiy, and Georgios N Yannakakis, 'The mario ai championship 2009-2012.', *AI Magazine*, **34**(3), 89–92, (2013).
[19] Julian Togelius, Georgios N Yannakakis, Kenneth O Stanley, and Cameron Browne, 'Search-based procedural content generation: A taxonomy and survey', *Computational Intelligence and AI in Games, IEEE Transactions on*, **3**(3), 172–186, (2011).
[20] Geogios N Yannakakis, 'Game ai revisited', in *Proceedings of the 9th conference on Computing Frontiers*, pp. 285–292. ACM, (2012).
[21] Yu, Derek. Spelunky. Mossmouth, 2009.

Hybrid Procedural Content Generation: A Proposal

Michael Cook and Simon Colton[1]

Abstract. Procedural content generation in games tends to target content that is abstract, dry and devoid of connection with the game's meaning. This paper proposes merging user-driven content generation approaches with procedural content generation to create a new paradigm which we call *Hybrid Procedural Content Generation*. By replacing aspects of existing procedural generation techniques with humans, we can give rise to new kinds of game experiences.

1 Introduction

Procedural content generation and user-generated content (PCG and UGC respectively) are two concepts which are familiar to anyone who has played or made games in the past decade. The idea that content for a game can be created after it has shipped enables many new kinds of game experience, as well as engaging players in new kinds of activities, including creative involvement in the game. They also provide interesting research platforms to ask new questions and build intelligent systems to help shape these new ideas about games.

In this paper we introduce the concept of *Hybrid* Procedural Content Generation (HPCG), a fusion of user-generated and procedurally-generated content that similarly offers new kinds of game design and also new opportunities for artificial intelligence in games. By incorporating players into procedural content generation systems we can produce hybrid systems that are much stronger than standard procedural or user-driven generative approaches.

We illustrate the concept of HPCG by giving three examples of prototype games which incorporate some kind of HPCG system into their game design. *Murder* is an assassination game set in a Cluedo-esque mansion at a dinner party, in which the player must perform several narrative actions and then kill another character at the party. *Mystery* is a Poirot-style detective game in which the player must solve a murder using deduction and exploration. *The Book Of A Thousand Tales* is a roleplaying game in which the player leads a band of heroes through a branching narrative.

The remainder of the paper is organised as follows: in *Background* we discuss both PCG and UGC and their relative weaknesses. In *Hybrid PCG* we briefly introduce the concept of HPCG, its motivating factors and how we see it being used within games. We then describe two simple game designs that comprise a HPCG system. Finally in *Opportunities for Computational Intelligence* we talk about the longer-term impact of such approaches and the potential for new research directions HPCG could give rise to. We then sum up our proposal in *Conclusions*

2 Background

According to [6], most PCG systems can be categorised as either *constructive* or *generate-and-test* systems. In the former, content is gradually built up out of successive passes at generation, and each layer of generation is "*guaranteed to never produce broken content*" [6]. Spelunky [2] is a good example of this style of generation, where dungeon levels are built out of several different layers of content which are hand-crafted to some extent to guard against failure [7]. Generate-and-test approaches employ a generative step that produces content, and then an evaluative step which assess what was generated and either triggers further generation/alteration (such as an evolutionary system which will run many times to evolve a result, as in [2]) or simply reject the generated content and begin again from scratch. *Dwarf Fortress* [3] employs a generate-and-test approach during its world generation.

PCG has been applied very effectively to many kinds of content generation, particularly level design [3] and general game content such as item generation in roleplaying games. However, many types of content are hard to generate using either of the above approaches. In particular, content which requires an understanding of context of the real world is hard to generate, such as game narratives or replicating human-like qualities in NPC actions such as deception or fallibility. These dynamic kinds of content rely on an understanding of the real-world, from cultural knowledge (like understanding symbolism when constructing a narrative) to common-sense reasoning (when deciding how a character should react to a particular situation, for instance). As a result, most content generation focuses on abstract data that is detached from the game's setting and theme (the levels in Spelunky are simply arrays of numbers, for instance – the system does not need to understand what a cave looks like or what an explorer does).

User-generated content (UGC) is also a common feature in many modern games. Allowing the player to create content for a game both increases the amount of content available at no extra cost to the developer, and gives players a sense of engagement and investment in the game world by allowing them to contribute to it. *Spore* [4] is a prominent example of user-generated content – players designed animal species for inclusion in their games using an assortment of body parts and customisations. These animal species propagated not only throughout the player's world but also to their friends' worlds via cloud sharing online.

UGC is one of the biggest recent trends in the mainstream industry thanks to the enormous success of *Minecraft* [5], which merged user-generated content with the core mechanics of the game. In Minecraft, generating content is how one plays the game: building structures, artworks and shaping the world as the player sees fit. UGC has drawbacks, however. In the case of generators like Spore's, which present themselves as tasks outside of gameplay, the user is consciously

[1] Computational Creativity Group, Goldsmiths, University of London

[2] Mossmouth Games, 2009
[3] 2006, Bay Twelve Games
[4] Maxis, 2008
[5] Mojang, 2011

aware that they are generating content. As a result they are thinking about how the content will be perceived by others, which has an impact on how and what they create. This can be seen somewhat in the comedic nature of many of Spore's creatures – players know they are creating things which will amuse or confuse other people. While this may be seen as a positive for some tasks (in Spore's case the objective is specific content generation) because the player is consciously considering the design of their content, for other tasks it may be less good – particularly those that take place in a fictional context. For example, in Minecraft it is possible to construct floating houses, which may break the suspension of disbelief for other players. It is preferable here that all players construct buildings in a similar way, so that they can maintain the narrative fiction for everyone equally.

The second drawback is that players tend not to be designers, and UGC systems rarely have any kind of feedback mechanism or assistive aspect to them. Content is either used wholesale or not used at all, and frequently even this decision is made by players rather than an intelligent software system. Creatures in Spore are uploaded and shared online, structures in a Minecraft world exist for all players in that world and can't be edited or changed by the game. UGC is all-or-nothing and thus lives or dies on the skill and appreciation of the players using these systems. In some cases this can be worked around – ratings systems in games such as LittleBigPlanet [6] simply filter the best creations and downplay the rest. In this case, however, UGC simply becomes a means by which to discover talented people and get them to produce content, rather than allowing everyone to contribute equally.

3 Hybrid PCG

We propose that PCG and UGC approaches can be combined in a single approach that solves some of the problems mentioned in the previous section while opening up new challenges and research questions for computational intelligence research to tackle. We call this combined approach *Hybrid PCG* because it synthesises software-driven content generation with player activity. The underlying premise is to replace generative systems or parts of systems with playable games, resulting in new ways of generating, evaluating and filtering content, not just for single games by potentially for many different games at once.

To illustrate this approach, we will describe in this section two in-development game prototypes, *Murder* and *Mystery*, which utilise a HPCG approach to generate a large corpus of content and filter it. These games not only supply content to one another: by generating content that is transferred between games, they also produce a corpus that can be used by other games or intelligent systems. After describing the games we will discuss the new affordances such a setup offers and then lead into a discussion of the opportunities for computational intelligence they represent.

3.1 Illustrative Example - Murder/Mystery

In *Murder* the player takes on the role of a character attending a dinner party at a mansion, as either a guest, a family member, or an employee of the host. Like most of the people present they have a motive to kill the host, and must do so at some point during the evening. In addition, they must also complete one or more objectives relating to their motive (such as confronting the host in an argument, or breaking into a room and stealing something). The game operates in a 'sandbox' style, where the player can explore the house freely and approach their objectives in many different ways. However, the game simulates player action carefully and records things like fingerprints left on surfaces, sightings by other people in the house, and so on.

At the end of the game, once their tasks are completed, the player can choose to 'discover' the body themselves or wait for it to be discovered by someone else. They are then asked to provide an account of their whereabouts for the evening by being shown their actual movements and then editing them to change their version of events – for example, by claiming they were never in a particular room at a certain time, and so on. The game then assesses how quietly and quickly they completed the game, as well as how well their alibi compares to the evidence they left behind, and gives them a rating.

In *Mystery* the player takes on the role of a detective tasked with solving a murder at a dinner party. They play a point-and-click adventure in which they can examine the alibis and backgrounds of the characters present, ask for accounts of events, and walk around the house looking for clues or analysing parts of the crime scene. The case files are built from case descriptions produced by *Murder*, potentially converted using an automated system that can filter the case to make it harder or easier (by making certain evidence more or less conclusive or adjusting the memories of other characters, for example) or simply presented to players unaltered – we discuss this further in section 4.

There is a time and resource limit on solving a case - if the player takes too long or uses up all of their investigative resources (such as sending objects for fingerprinting) the case remains unsolved. Whatever the result, the case file data gets sent back to a central server which both affects the value of a case (repeatedly unsolved cases rise in value to detectives) and the reputation of the player who created the case file in *Murder*.

3.2 HPCG in Murder/Mystery

Both *Murder* and *Mystery* are standalone games that are effectively separate from one another. If the data format for case files is open, anyone could design a game which retrieved case files produced by *Murder* players and use them in their game. Similarly, several games might produce case files with the right format that could be used by *Mystery* as game content for the player to investigate and solve. The games are not intrinsically linked except through the exchange of information about the case files and whether or not they are solvable by players.

In the language of PCG, players of *Murder* are acting as a generator of case files, in the first step of a *generate-and-test* system. There are two important consequences of this. Firstly, unlike UGC approaches, the players are *engaged in a game* while generating content, pursuing objectives in whatever way they see fit. We argue that this leads to more natural behaviour by players and therefore a more human-like kind of content generated than if players had been asked to manually design case files as authors. Secondly, the content being generated is complex - it involves creative problem-solving and asks the player to respond to social situations (such as confronting someone about a personal relationship, or making small-talk at a dinner party). Such content is difficult to generate automatically without a lot of involvement from a designer, and even with such involvement the content is likely to be lacking in variety over a long period of play. By using players to generate it, we make this difficult generative task easier.

To continue the PCG metaphor, players of *Mystery* act as evaluators of the content generated by *Murder* players. Let us assume that

[6] Media Molecule, 2008

Mystery either does not edit the case files at all, or at most edits them in order to ensure that they can be solved by some process of deduction (by ensuring that at least one piece of incriminating evidence exists, for instance). Players solving, or attempting to solve, cases are providing data about how easy a case is to solve. The routes players took, the order in which they examined evidence or questioned people, and their ultimate success at solving the murder can all be recorded as additional metadata attached to the original case file. In the same way that people can be used to generate content that requires complex understanding of the real world, people can also be used to provide evaluation metrics that would be difficult to encode into a system by hand (and too subjective to source from a single designer).

3.3 Desirable Properties of HPCG Scenarios

While this remains a preliminary proposal for HPCG, and the idea still needs much exploration, we posit that certain game designs or scenarios are better suited for the application of HPCG. We discuss them briefly here, and hope to clarify this in future work after more experimentation and prototype development.

3.3.1 Asynchronous Activity

The most important property for employing HPCG is that the games involved deal with asynchronous activity. Murder/Mystery work well because the two game phases are chronologically non-overlapping: one player commits a crime, then after they are finished the second player can arrive and solve it. This means that no player is left waiting for action to be completed in real-time, which could affect the experience of either player and slow down gameplay, and it also means that any PCG systems have complete information from the other game or games when they begin generating content.

3.3.2 Well-Defined And Decoupled Interfaces

Keeping the interfaces between games as simple as possible is a good feature if the designer intends for other systems to feed data into the HPCG besides their own. For Murder/Mystery we noted that in theory it is possible for other games to generate crimes for Mystery to solve, or to design other games which use Murder case files as input content. In order to enable this, it's important that the interfaces between the games are very well-defined and public so that other developers can take advantage of them. Making sure the games can export data as well (such as putting Murder's case files in external text documents) also makes this easier.

3.3.3 Guided Player Activity

Depending on the kind of content being generated or the roles the players are taking on in the larger HPCG system, it may be desirable for the gameplay to be very directed or guided. The reason for this is that the HPCG system is making assumptions that the data they collect represents a certain kind of behaviour from the player - for example, committing a crime, not wanting to leave evidence behind, acting in order to blend in. It's important to be able to encourage and motivate the player to work towards certain objectives so that these assumptions carry through into the data they generate, and can then be relied upon to generate good quality content in other areas of the HPCG system. If a player begins acting differently, or isn't sufficiently motivated to play properly, the HPCG system will still proceed with the data and this can generate undesirable outcomes in other games.

4 Opportunities for Computational Intelligence

On the surface, HPCG appears to replace software-driven PCG systems with players that perform the same tasks, therefore resulting in systems that involve *less* computational intelligence, rather than more. However, HPCG systems open up new research questions that demand answers, and also create opportunities to build even more complex generative software. In this section we discuss several possibilities in brief.

4.1 Learning From Human Generators

One possible outcome from HPCG systems is that they eventually transition back into being PCG systems which use a player's in-game activity as a source of training data. In [4] Orkin and Roy describe *The Restaurant Game* (TRG), an experiment in which participants played through an interactive scenario in pairs and their behaviour was then recorded and later analysed using machine learning to build behaviour models of characters in those situations. TRG suffers from some of the same problems that we mentioned in the context of UGC earlier in the sense that players are aware they are generating content as they play. Nevertheless, the authors' argument is that automatic content generation (in this case speech and behaviour patterns) can be mined from large-scale data corpora [5].

By employing HPCG to tackle complex generative tasks, like the generation of creative behaviour in *Murder*, such systems produce special cases of the kinds of corpora Orkin and Roy present with The Restaurant Game. They are special cases in the sense that they are obtained through observing players at a time when their primary concern is *completing* a game rather than performing for another observer (whether that observer is a human or a data-mining program). The player is not participating in an experiment, nor is their ultimate goal to provide good data. Instead, they are focused on achieving objectives and are immersed in a ludic task. As a result, we argue that their behaviour is more natural and as a result more valuable, resulting in useful corpora of data that can be mined, as with TRG, to obtain behaviour. In the case of games such as *Murder*, the available information is particularly valuable because the player is providing information that an ordinary PCG system would not have access to - such as solving problems in creative or innovative ways, as well as failing at tasks in a natural, humanlike way.

4.2 The Computer As Curator

The game *Murder* can be seen as a generator of content for the game *Mystery*, but raw generated case files from the game may not be interesting, fun to solve or, indeed, solvable at all. Building *Murder* as a HPCG system provides us with a wealth of generated murder cases for players to solve, but it doesn't guarantee their quality or difficulty level. If a player plays a perfect game, it will be fairly unsatisfying for players of *Mystery* to repeatedly fail to solve. Similarly, the player may make an obvious mistake that renders a case trivial. This poses an interesting problem: how can software curate, tweak and improve raw HPCG output to ensure consistently entertaining content for another player?

There are many factors to tweak in a case file produced by *Murder* – both the actions of the players and the other characters, the evidence left behind, the ordering of events. Altering this information requires

an understanding of how people's behaviour is interpreted by others, to assess whether a change will make a case easier or harder to solve for a player detective. HPCG systems leverage human players to solve creative, complex problems that are hard to solve using generative software alone. It follows, therefore, that curating and improving the results of a HPCG problem requires an understanding of how these players reason about problems and act in certain situations. The task of curating complex creative content sourced from humans may have parallels with the problem of curating and evaluating in Computational Creativity [1].

Recall that in section 2 we discussed the problems with existing UGC and PCG paradigms. One problem with UGC approaches is that players are not designers, and expecting them to be able to produce quality game content, either knowingly or not, is unreasonable and often results in a large volume of low-quality content that no-one wants to use. HPCG offers an opportunity to leverage the output of users and improve it using computational intelligence, obtaining content that has its foundations in the creativity of real players, but has been curated and refined by software to be of higher quality.

5 Acknowledgements

The authors would like to thank the reviewers who provided helpful feedback which improved this paper. This work was sponsored in part by EPSRC grant EP/L00206X.

6 Conclusions

In this paper we briefly outlined a proposal for *Hybrid Procedural Content Generation* or HPCG, a synthesis of user-generated content and procedural content generation where subsystems in a content generation pipeline are replaced with players playing games achieving similar tasks. We illustrated the idea with two connected games – *Murder* and *Mystery* – in which players of the former acted as a generator of content which was then filtered and evaluated by players of the latter. We discussed what new avenues of research such an approach might offer and how it solves some of the problems that procedural content generation and user-generated content can have.

This paper is an early proposal for such games and systems to be designed, but we hope that it will spark discussion and potentially lead to interesting new kinds of games and intelligent software. We believe that working with game developers may be of essence here, to leverage good game design alongside new kinds of computational intelligence. Collaboration is difficult, but we believe this is a promising avenue to explore.

REFERENCES

[1] Simon Colton, Michael Cook, Rose Hepworth, and Alison Pease, 'On acid drops and teardrops: Observer issues in computational creativity', in *Proceedings of the 7th AISB Symposium on Computing and Philosophy*, (2014).

[2] Erin J Hastings, Ratan K Guha, and Kenneth O Stanley, 'Evolving content in the galactic arms race video game', in *IEEE Symposium on Computational Intelligence and Games*, (2009).

[3] Britton Horn, Steve Dahlskog, Noor Shaker, Gillian Smith, and Julian Togelius, 'A comparative evaluation of procedural level generators in the mario ai framework', (2014).

[4] Jeff Orkin and Deb Roy, 'The restaurant game: Learning social behavior and language from thousands of players online', *Journal of Game Development*, (2007).

[5] Jeff Orkin and Deb K. Roy, 'Understanding speech in interactive narratives with crowdsourced data.', in *AIIDE*. The AAAI Press, (2012).

[6] Julian Togelius, Georgios N. Yannakakis, Kenneth O. Stanley, and Cameron Browne, 'Search-based procedural content generation: A taxonomy and survey', *IEEE Transactions on Computational Intelligence and AI in Games*, (2011).

[7] Derek Yu. The full spelunky on spelunky, 2012.

Revealing Social Identity Phenomena in Videogames with Archetypal Analysis

Chong-U Lim[1] and **D. Fox Harrell**[2]

Abstract. In this paper, we present a novel approach toward revealing social identity phenomena in videogames using archetypal analysis (AA). Conventionally used as a dimensionality reduction technique for multivariate data, we demonstrate how AA can reveal social phenomena and inequity such as gender/race-related steretoyping and marginalization in videogame designs. We analyze characters and default attribute distributions of two critically acclaimed and commercially successful videogames (*The Elder Scrolls IV: Oblivion* and *Ultima IV*) together with 190 characters created by players in a user-study using a third system of our own design. We show that AA can computationally 1) reveal implicit categorization of characters in videogames (e.g., base player roles and hybrid roles), 2) model real world racial stereotypes and stigma using character attributes (e.g., physically dominant attributes for *Oblivion*'s ostensibly African-American "Redguard" race) and 3) model gender marginalization and bias (e.g., males characterized as more archetypal representations of each race than females across attributes.) We highlight how AA is an effective approach for computationally modeling identity representations and how it provides a systematic way for the critical assessment of social identity phenomena in videogames.

1 INTRODUCTION

Videogames often construct virtual environments and worlds that are populated with virtual characters. Both these worlds and characters may be represented in a multitude of ways. Graphical 2-dimensional (2D) or 3-dimensional (3D) assets grant visual appearances, textual descriptions provide intriguing narrative, backstories, and characteristics, while numerical statistical attributes provide quantifiable measurements defining character skills and capabilities for a variety of interactions, from dealing damage against a mighty adversary, to charming a non-playable character into handing over an elusive item.

Though often considered to be purely virtual, these representations are in fact blended real/virtual identities that are both affected by, and capable of influencing, aspects of real world identities. Even in the case of a fairly rudimentary character such as Pac-Man, in action we have a blend of a real users control with a 2D animated sprite. Recent studies have shown how representations of race and gender within videogames have deep social implications [8]. In the commercially successful and critically acclaimed role-playing game (RPG) *The Elder Scrolls IV: Oblivion*, some character designs "implement and amplify many disempowering social identity constructions" [9]. "Females of some races are more intelligent than their male counterparts and individuals of the ostensibly French 'race' (Bretons) are twenty points more intelligent than their ostensibly Norwegian (Nords) counterparts, regardless of gender" [9]. It highlights the *importance of the underlying implementations and data structures used to construct these representations*. If developed without due consideration, undesirable social implications related to identity such as marginalization and stereotyping may be further perpetuated. Research has shown that peoples' performances are impacted by stereotypes [20] and behaviors in the physical worlds are altered by their avatar use [22].

However, it is important to recognize that these issues are not simply technical in nature. We adopt a *critical computing* [9] approach, using algorithmic processing and data structuring for critically assessing and providing commentary about the real world and related social phenomena. In this paper, we demonstrate an how archetypal analysis can be used as an Artificial Intelligence (AI) tool for such critical assessments of computational identity-related social phenomena in two commercial videogames, as well as a character creation system of our own design. The upshot is that we found that AA is a robust method for computationally modeling underlying social identity phenomena grounded in cognitive science. We use AA to model social phenomena within games, such as male characters being favored over female characters based on statistical attribute distributions or in-game races having real world stereotypes imparted upon them (e.g., ostensibly African-American "Redguard" characters in *Oblivion* given better physical but lower mental stats.) To the best of our knowledge, this application of cognitive science (apart from some notable exceptions like Santa Ana's work on discrimination and racism [17] and Lakoff's work on political affiliations [12]) and AI has not often been applied to analyze nuances of identity. Most computational systems, like videogames, are built with classical models and categories explicitly built into software. Hence, they provide a good venue to critically assess such cognitively grounded AI approaches to studying digital identity.

2 BACKGROUND

In this section, we present the theoretical framework for our work and provide an overview of the videogames used for our analysis.

2.1 Cognitive Categorization and the Sociology of Classification

Our view of categorization is based upon cognitive scientist George Lakoff's work in cognitive categorization [11] termed *category gradience* and psychologist Eleanor Rosch's *protoypes* [16]. As opposed to outmoded classical or "folk" approaches, which character-

[1] Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA, email: culim@mit.edu

[2] Computer Science & Artificial Intelligence Laboratory, Comparative Media Studies Program, Massachusetts Institute of Technology, USA, email: fox.harrell@mit.edu

ize category membership to be defined by a fixed set of characteristics, centrality gradience recognizes that some members are typically deemed "better examples" of a category than others. Extending upon this, we use the following concepts from the sociology of classification by Geoffrey Bowker and Susan Leigh Star [3] for describing categorization-related social phenomena. **Membership** is the experience of encountering and interacting with objects within certain social groups, and increasingly engaging in naturalized relationships with them. **Naturalization** is the deepening familiarity of such interactions within a given social group. **Marginalization** is a result of enforced naturalization occurring where members of a marginal category exist outside of social groups, or are less prototypical members of communities. It is also characterized by exclusion from a social group or an individual having *multiple memberships* and often refers to *exclusion or difference from normative behaviors* (Stigma) [7, 9]. **Markedness** indicates that, unlike normative categories, marginal categories are demarcated visually and linguistically.

To reconcile these concepts with the systems in this paper, we use a cognitively-grounded model for critically assessing computing systems for social analysis [9]. It suggests that category gradience enables semantic relations to be structured or ranked according to how constitutive they are of the category. Naturalization may be assessed by *user actions and attributes* that reinforce category semantics, resulting in a higher degree of membership. Marginalization may be implemented through enabling *degrees of membership* and represented as being *further away from the prototypes*. Normative groups that are often unnamed and unmarked may possess *implicitly assumed* normative privileges that may be identified and modeled. This theoretical framework forms the basis for using archetypal analysis as an approach for social analysis and empowerment through critically assessing the statistical attributes of characters within videogames for revealing implicitly-derived social phenomena such as gender-related marginalization and stereotyping.

2.2 Archetypal Analysis

Archetypal Analysis (AA), introduced by Cutler and Breiman [5], is a method for reducing the dimensionality of multivariate data [1]. Given a set of multivariate data points, the aim of AA is to be able to represent each data point as a *convex combination* of a set of key data points called **archetypes**. For example, applying AA on a dataset of basketball players and their statistics [6] computationally revealed and represented the following four archetypes – "benchwarmer," "rebounder," "three-point shooter," and "offensive." Every individual player in the entire data set could then be represented as a hybrid mixture of these archetypes [18]. Formally, given a data set of points $\{x_1, x_2, ..., x_n\}$, AA seeks to find a set of archetypes $\{z_1, z_2, ..., z_k\}$, where $z_j = \sum_{i=1}^{n} \beta_{ij} x_i$, and enables each data point x_i to be represented in terms of the k archetypes as $x_i = \sum_{j=1}^{k} \alpha_{ji} z_j$. The objective function minimizes the residual sum of squares $RSS = ||x_i - \sum_{j=1}^{k} \beta_{ij} z_j||^2$ under the constraints that the weights $\sum \beta_{ij} = 1$ $\beta_{ij} \geq 0$ and coefficients $\sum \alpha ji = 1$ $\alpha_{ji} \geq 0$. These ensure the archetypes *meaningfully resemble* and are *convex mixtures* of the data. These archetypes are located on the data convex hull [5] and are represented are combinations of individual points, making them more easily interpretable [1], unlike other dimensionality reduction techniques like Principal Component Analysis [10] and Non-negative Matrix Factorization [13]. AA has been shown to be effective compared to other techniques for various AI-related problems. Compared to other recommender models (nearest neighbor, two popularity, random baseline) AA provided the highest recall rates for archetypal recommender systems [19] in games, demonstrating robustness for finding relevant recommendations. Here, AA is an appropriate approach given our aim to computationally model individuals that are more "prototypical" than others (archetypes) and being able to measure the "centrality gradience" of each individual with respect to these archetypes. As described in Section 2.1, we believe that such models would enable us to begin critically assess social phenomena such as marginalization and stereotyping computationally.

2.3 Overview of Videogames

We provide an overview of the two commercially successful videogames used in this paper. Both are important open-world single-player RPGs with strong customization. *Ultima IV: Quest of the Avatar* is arguably the most influential game on the open world RPG genre and *The Elder Scrolls IV: Oblivion* is a stunning recent success with a strong customization system and diversity. Even in excellent games, there is the potential for implicit stereotypes and inequity. Our observations are meant to be useful for improvement in this regard.

The Elder Scrolls IV: Oblivion is the fourth installment of the popular *Elder Scrolls* computer role-playing game series, developed by *Bethesda*. In the lore of the game designed by game designers there are several races, each with their own fictional background stories and histories. Three basic player roles exist in the game – "Fighter", "Mage" and "Thief" [15], which are derived from common roles across most RPGs stemming from old table-top RPGs like Dungeons and Dragons. Each race is associated with the three basic roles in varying degrees (hybrid roles), which compliment the game's lore about its people and races. Players choose to play as one of the ten different races available, customizing characters over 7 basic attributes (strength, intelligence, willpower, agility, speed, endurance, and personality,) together with their height and weight.

Ultima IV: Quest of the Avatar is the fourth installment of the *Ultima* series of role-playing games, and the first in the "Age of Enlightenment" trilogy, *Ultima IV* was first released in 1985 by *Origin Systems*. The player is assigned one of eight classes to play and does not directly choose or assign values to attributes. Instead, the user is posed several questions embedded within the games narrative at the beginning, resulting in the players ranking of eight **virtues** in the game based on the game's three **principles** of Truth, Love, and Courage. There are seven companions that the player may choose to form a party with. Each character has a particular class, each associated with a virtue, and possesses seven numerical attributes (strength, dexterity, intelligence, hit points (HP), magic points (MP), level, and experience,) an armor type, a weapon type, and their gender.

3 APPROACH

1. Analyzing existing systems for designer-centered phenomena. In order to assess the kinds of categorization and social identity phenomena that arise as a result of designer choices (top-down), we applied archetypal analysis to the statistical attribute allocation for new characters in both *Oblivion* and *Ultima IV*. For *Oblivion*, the variables included the races, gender, and eight attributes. For *Ultima IV*, the variables included the character classes and seven attributes.

2. Analyzing emergent phenomena with a system of our own creation. For the purpose of assessing the kinds of categorization and social phenomena that may be *implicitly-derived* from players (bottom-up), we conducted a user-study with 190 players where they constructed avatars in an avatar constructor of our own creation. Players customized both their character's visual appearance and statistical attributes values of six commonly used videogame attributes (strength, endurance, dexterity, intelligence, charisma, and wisdom) on a 7-point Likert scale with a total of 27 allocatable points. The avatar constructor used our avatar game data-mining system called *AIRvatar* [14], that stores each created avatar, the statistical attribute allocations, and textual descriptions made by the players.

3. Determining the number of archetypes During AA, we varied the number of archetypes k in the range $1 \leq k \leq 10$. We adopt the convention of the Cattell scree test [4] for using the residual sum-of-squares (RSS) to determine the optimal number of archetypes by picking the value of k matching the first point of the "elbow" of a screeplot with corresponding to the biggest change in RSS. This balances the trade off between minimizing RSS and overfitting.

4 RESULTS

We present results describing the archetypes obtained from analyzing the statistical attributes of each system using archetypal analysis.

4.1 Oblivion

In *Oblivion* we found $k = 3$ to be optimal. Both Archetypes 2 and 3 were pure archetypes ($\alpha_j = 1$). The ternary plot in Figure 2(a) of the Appendix shows a visualization of the α coefficients of these archetypes. We also observed the following characteristics:

- Archetype 1 had the highest "Strength" and "Endurance", but lowest "Intelligence". Archetype 1 had the biggest "Size".
- Archetype 2 was relatively balanced across the attributes, with highest "Willpower" and "Personality".
- Archetype 3 had highest "Intelligence", "Agility" and "Speed", but lowest "Willpower". Archetype 3 had a relatively small "Size".

4.2 Ultima IV

In *Ultima IV*, we found $k = 3$ to be optimal. All three were pure archetypes. The ternary plot in Figure 2(b) of the Appendix visualizes the α coefficients of these archetypes. We also observed that :

- Archetype 1 had the lowest values across all attributes.
- Archetype 2 had the highest values across all attributes, except for "Intelligence" and "Magic Points".
- Archetype 3 had the highest "Intelligence" and "Magic Points".

4.3 AIRvatar

For characters created using *AIRvatar*, we found $k = 3$ to be optimal. The bar plot in Figure 1 shows the three archetypes obtained, represented with the same six RPG attributes. We observed the following:

- Archetype 1 had highest "Intelligence" and "Wisdom" attributes, but lowest "Strength" and "Endurance".
- Archetype 2 had the highest "Strength", "Endurance", and "Dexterity" attributes, but the lowest "Wisdom".
- Archetype 3 had the highest "Charm" but lowest "Dexterity".

5 FINDINGS

5.1 Classes, Roles, and Category Gradience

In *Oblivion*, we found that each **archetype corresponded with the primary roles of the game**, namely "Fighter" (Archetype 1), "Mage" (Archetype 2), and "Thief" (Archetype 3). We used descriptions in the *Unofficial Elder Scrolls Pages* [15], to help identify these roles from obtained archetypes. "Fighters" '*rely heavily upon melee combat to attack enemies, expect to receive a lot of damage rely upon high health...*', "Mages" '*avoid combat, use decoys, and rely upon magical attacks.*' *Magicka*, used for spells and magic, is affected by both "Intelligence" (Capacity) and "Willpower" (Regeneration). A "Thief" '*relies upon sneak attacks and avoids face-to-face combat, uses a poisoned bow as a primary means of attack,*' corresponding to the high "Speed" and "Dexterity" (Bow Accuracy) attributes.

Likewise, in *Ultima IV*, we observed from our results that each **archetype corresponded with characters of primary roles in the game**. Katrina the Shephard is Archetype 1 as her description in the *Unofficial Ultima IV Strategy Wiki* [21] states "...she has the lowest attributes, no magic power and a limited selection of equipment; start the game with her if you're looking for a challenge". Archetype 2 corresponds to "Iolo the Bard", who has the highest "Dexterity" described as "probably the most important attribute because it rules the probability of hitting enemies, avoiding traps and dodging enemies."Archetype 3 corresponds to "Mariah the Mage", with highest "Intelligence" (determines maximum "Magic Points".)

For characters created by players in *AIRvatar*, we observed from our results that the archetypes corresponded with **traditional RPG roles used in games**, which we term "Intelligent/Wise-Cleric" (Archetype 1), "Physical-Fighter," (Archetype 2) and "Charming-Thief" (Archetype 3). We the descriptions of traditional *Dungeons and Dragons* classes to match against the highest-scoring attributes of each archetype to identify these roles. Magic using "Mages/Clerics" focus on magic, and generally have lower strength. "Fighters" are usually strong in attack and defense, but usually have little to no magic capabilities, while "Thieves" often are in-between, but have high capabilities in social skills, cunning and stealth.

We validate this based on the free-text responses that players provided for their avatars, in addition to customizing their characters. We provide selected responses from the highest scoring players for each archetype to highlight this behavior:

1. **Archetype 1 (Intelligent/Wise-Cleric):** "*Stephanie is a wandering wolf **mage**. She was born to a poor family, but her parents did their best to support her academic ventures. She studied hard and was eventually admitted to the nation's most prestigious **arcane** academy.*"
2. (**Archetype 2 (Physical-Fighter)**: "*Gerald ... is a veteran of many wars in Elibca, serving as a **knight** and later as a **general** for the kingdom of Calmenia ... living the remainder of his life in modesty as he nurses **old scars**.*" & *"Saya is an independent **Mercenary** selling her contract not to the highest bidder, but to those she deems in the most need of her services. Secretly, she dreams of becoming a Paladin some day but believes that she has **far too candor in her speech** and methodology to fit in ... "*
3. **Archetype 3 (Charming-Thief):** "*She is **friendly** and ready to reach out to the other villages. She **prefers talking to fighting**, but is tough enough to fight if she needs to.* "

These results shows that AA can effectively model implicit categories, such as intended player roles and relationships between attributes from analyzing raw statistical attribute data. For example, in

both *Oblivion* and *AIRvatar*, "Strength" and "Intelligence" attributes are always maximized on different archetypes, while "Strength" and "Endurance" were be maximized on archetypes together. Additionally, with archetypes corresponding to prototypical player roles, we observed that **each individuals could meaningful represented as mixtures of these archetypes**, corresponding to hybrid roles intended by most designers.

5.2 Revealing Stereotypes, Marginalization, and Inequity

5.2.1 Race-related Stereotyping

From the archetypal analysis results on characters in *Oblivion*, we were able to observe that **some of the in-game races were deemed more "prototypical" with respect to player roles and that we could observe how these in-game races reflected real world stereotypes**. To visualize this, we make use of the ternary plot of results shown in Figure 2(a) of the Appendix. This is best visualized using a ternary plot, as shown in Figure 2(a) of the Appendix. We observe that the ostensibly Norwegian "Nords" are viewed as archetypal Fighters, the ostensibly French "Bretons" as archetypal Mages, and ostensibly South American "Bosmers" as archetypal Thieves. Additionally, the obstensibly African-American "Redguards" stereotypically close to the physical-fighter archetype with no characteristics of the intelligence-mage archetype, though exhibiting some stealth-thief archetype characteristics. This corresponds with findings by Harrell in his assessment of racial stereotypes in *Oblivion* [9].

5.2.2 Gender-related Inequity & Marginalized Characters

In *Oblivion* we also note that **for each race, male characters are consistently deemed more prototypical than their female counterparts than their female counterparts.** This is illustrated in Figure 2(a), where for each archetype, the male characters are always at least as close, or closer to the archetypes, than their counterpart female characters. Insight into the significance of characters being closer to the centers (i.e., further away from archetypes) is highlighted in the design choices made in *Ultima IV*, wherein the NES version of *Ultima IV*, "Julia" was replaced by a male character "Julius", with no modification to the stats. From the ternary plot in Figure 2(b) of the Appendix, it can be seen that "Julia" is the character with negligible "Intelligence" and "MP" attributes and located between the overall lowest and highest-performing archetypes, possessing multiple memberships. This computational modeling of a **less prototypical individual would, by Lakoff's definitions [11], represent the marginalization of that individual**. We hypothesize that the implications of this made it seem "low-stakes" to swap her gender within the game and that it might have been more difficult to swap the genders of an archetype instead (i.e., making a Katrina a male to have the lowest stats or Iolo a female while having the highest stats.) To validate the effects of marginalization (being further away from archetypes), we sampled characters created with *AIRvatar* that had coefficient values $.3 \leq \alpha_k \leq 0.6$ for *all three archetypes*. These reflected characters that players created to be less prototypical.

- **Character #41:** "*Pinkie is a girl with a **unique gift** for magic, ... works **best in a team but can hold her own** when needed.*"
- **Character #102:** "*A spellcaster ... a love for **forbidden magics**. **Chaotic good**, generally **tries to do the right thing but isn't afraid** to crack a few eggs to make an omlette.*"

5.2.3 Gender-related Stereotyping

In the results of characters created using *AIRvatar*, we observed that **players constructed characters with more homogeneous gender distributions between archetypes and also when close to the archetypes**. We define close as individuals with coefficient values $\alpha_k >= 0.80$. In Table 1 of the Appendix, we observe that all three archetypes had a mixture of male and female avatars close to each of them. Both Archetype 1 ("Intelligent/Wise-Cleric") and Archetype 3 ("Charming-Thief") had more female avatars closer to the the archetypes than male avatars, while Archetype 2 ("Physical-Fighter") had more male avatars closer to it. These results share similarities with those of *Oblivion*, *Ultima IV*, as well as our precious analyses in [14] where males avatars were associated with more physical roles, and female avatars with magic-related roles. For the "Charming-Thief" role, neither females nor males were closely associated with it – showing that "Thief"-like roles have less gender stereotyping associated with them. These results appears to suggest that taken collectively, players seek to reduce the degree of marginalization or privilege of either gender relative to what designers commonly portray. We hypothesize that perhaps, in the absence of a well-known game series, people relied more on real-world gender stereotypes. Thus, these results may reveal what people do without being restricted to canonical classes and roles – an observation perhaps useful for developers incorporating race and gender into their designs.

6 LIMITATIONS & FUTURE WORK

Here we discuss several limitations of our approach and describe potential avenues for overcoming them with future work and directions.
1. Determining the number of archetypes The approach we outlined in Section 3 adopts Occam's Razor [2] in that we pick the lowest number of archetypes k from the minimization of the residual sum-of-squares (RSS). However, this may not always be effective, with the result possibly being that the archetypes discovered are not sufficient to *adequately represent* the rest of the data points. For example, with $k = 3$ archetypes applied to results from *AIRvatar*, we discovered that no close individuals ($\alpha_j >= .9$) for one of the archetypes. It is possible that other metrics for determining k could be employed (e.g., choosing higher values of a scree plot's elbow.)
2. Normalizing Statistical Attributes While there are similarities between the statistical attributes used for defining characters in various videogames, there are issues with standardizing the number, the descriptions, and the effects that each attribute has. Also, there is a tension between the gaming use of these terms like "Intelligence" or "Wisdom." and their real meanings. Additionally, different games use different numerical scales (e.g., upon-100 in *Oblivion* but upon-7 in *AIRvatar*) for these attributes. It is difficult to translate the significance of each point due to different granularities. A standardized list and scale would be useful for such cross-platform comparisons.
3. Representation Beyond Statistical Attributes Representation in computing systems spans across several other technical components of the system, including graphical assets and textual descriptions [8]. Our next step is to analyze additional data collected using *AIRvatar*, which include the images of the constructed avatars, textual descriptions made by players, and other behavioral data obtained using the analytical capabilities of *AIRvatar*. We believe that these additional sources of information will enable further insight into the types of social phenomena that players experience and encounter through virtual representations in videogames and other computing systems.

7 CONCLUSION

We have demonstrated a novel approach to computationally model cognitively grounded social identity phenomena in videogames using archetypal analysis (AA). Previous work in this area has relied on qualitative methods (e.g., self-reported surveys) to identity and assess the presence of social identity-related issues such as marginalization, stereotyping, and discrimination. We demonstrated AA's effectiveness for modeling gender-related marginalization and biases like males being represented as closer archetypes than females and race-related stereotypes like in-game races possessing attributes that reflect characteristics of real-world stereotypes. AA was also able to reveal implicit categories like prototypical RPG roles used in videogames, which had implications to such race and gender-related phenomena. Being able to reveal such emergent phenomena through analyzing the data structures and designs of systems mean that computing systems can be analyzed in a systematic way, enabling quantifiable insight to be gained while minimizing the common effects of subjective evaluations such as survey bias. We believe that these findings contribute towards substantiating the use of AI to better understand the effects of virtual characters on players behaviors.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1064495.

REFERENCES

[1] Christian Bauckhage and Christian Thurau, 'Making archetypal analysis practical', in *Pattern Recognition*, 272–281, Springer, (2009).
[2] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth, 'Occam's razor', *Information processing letters*, **24**(6), 377–380, (1987).
[3] Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and its Consequences*, MIT Press, 1999.
[4] Raymond B Cattell, 'The scree test for the number of factors', *Multivariate behavioral research*, **1**(2), 245–276, (1966).
[5] Adele Cutler and Leo Breiman, 'Archetypal analysis', *Technometrics*, **36**(4), 338–347, (1994).
[6] Manuel JA Eugster, 'Archetypal athletes', *arXiv preprint arXiv:1110.1972*, (2011).
[7] Erving Goffman, *Stigma: Notes on the Management of Spoiled Identity*, 1963.
[8] D Fox Harrell, 'Computational and cognitive infrastructures of stigma: Empowering identity in social computing and gaming', *Proceedings of the 7th ACM Conference on Cognition and Creativity*, 49–58, (2009).
[9] D. Fox Harrell. Toward a theory of critical computing. CTheory, ctheory.net/articles.aspx?id=641, 2010.
[10] Ian Jolliffe, *Principal component analysis*, Wiley Online Library, 2005.
[11] George Lakoff, *Women, Fire, and Dangerous Things: What categories reveal about the mind*, 1990.
[12] George Lakoff, *Moral politics: How liberals and conservatives think*, University of Chicago Press, 2010.
[13] Daniel D Lee and H Sebastian Seung, 'Learning the parts of objects by non-negative matrix factorization', *Nature*, **401**(6755), 788–791, (1999).
[14] Chong-U Lim and D Fox Harrell, 'Toward telemetry-driven analytics for understanding players and their avatars in videogames', in *In CHI'15 Extended Abstracts on Human Factors in Computing Systems*, (2015).
[15] Oblivion: Character Creation. The Unofficial Elder Scrolls Pages, 1995.
[16] Eleanor Rosch, 'Principles of categorization', *Concepts: Core readings*, 189–206, (1999).
[17] Otto Santa Ana, *Brown tide rising: Metaphors of Latinos in contemporary American public discourse*, University of Texas Press, 2002.
[18] Sohan Seth and Manuel J. A. Eugster, 'Probabilistic archetypal analysis', Technical report, arXiv.org, (2014).
[19] Rafet Sifa, Christian Bauckhage, and Anders Drachen, 'Archetypal game recommender systems', *Proc. KDML-LWA*, (2014).
[20] Claude M Steele and Joshua Aronson, 'Stereotype threat and the intellectual test performance of african americans.', *Journal of personality and social psychology*, **69**(5), 797, (1995).
[21] Ultima IV: Quest of the Avatar - Companions. StrategyWiki, 2013.
[22] Nick Yee and Jeremy Bailenson, 'The proteus effect: The effect of transformed self-representation on behavior', *Human communication research*, **33**(3), 271–290, (2007).

A COEFFICIENT TABLES

Description	α_1	α_2	α_3	Player Gender	Avatar Gender
Archetype 1					
("Intelligent/Wise-Cleric")	***1.00**	0.00	0.00	Female	Female
	0.90	0.00	0.10	Female	Female
	0.82	0.17	0.01	Male	Male
Archetype 2	0.00	***1.00**	0.00	Male	Male
("Physical-Fighter")	0.00	***1.00**	0.00	Female	Female
	0.00	***1.00**	0.00	Male	Male
	0.00	***1.00**	0.00	Male	Male
	0.00	***1.00**	0.00	Male	Male
	0.00	0.88	0.12	Male	Female
	0.00	0.87	0.13	Male	Male
	0.14	0.86	0.00	Male	Male
	0.15	0.85	0.00	Male	Male
	0.14	0.85	0.02	Female	Female
	0.00	0.83	0.17	Female	Male
	0.00	0.80	0.20	Male	Male
Archetype 3	0.10	0.00	***0.90**	Male	Male
("Charming-Thief")	0.00	0.11	0.89	Female	Female
	0.15	0.00	0.85	Female	Female

Table 1. Table of characters created with *AIRvatar* with high α coefficients to each archetype. Values ≥ 0.90 are bolded. * marks the closest individual(s) of each archetype.

B BAR PLOTS

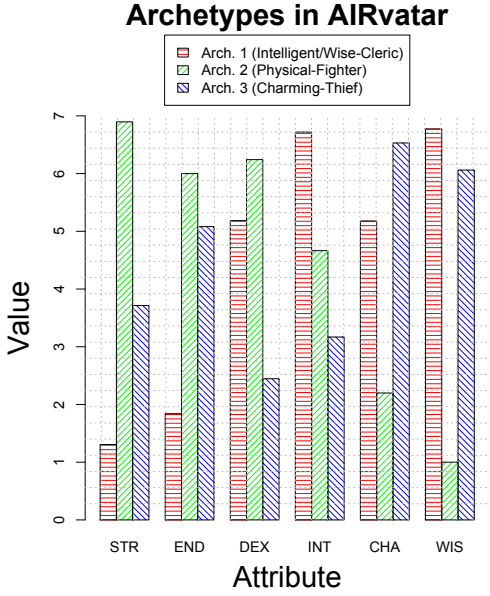
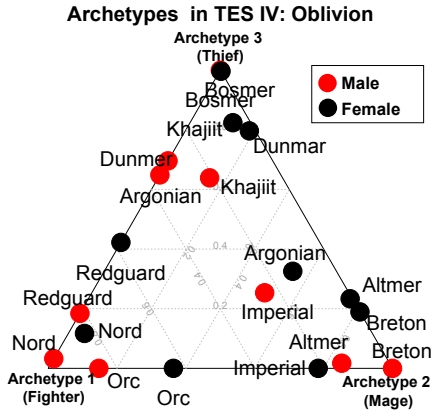
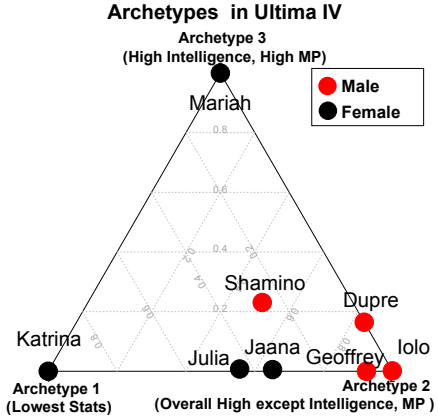


Figure 1. The plot above shows the $k = 3$ archetypes obtained from archetypal analysis on the data set of players and their statistical attribute allocations to each of their avatars. Due to convexity constraints, archetypes can be meaningfully represented with the same features of the original data.

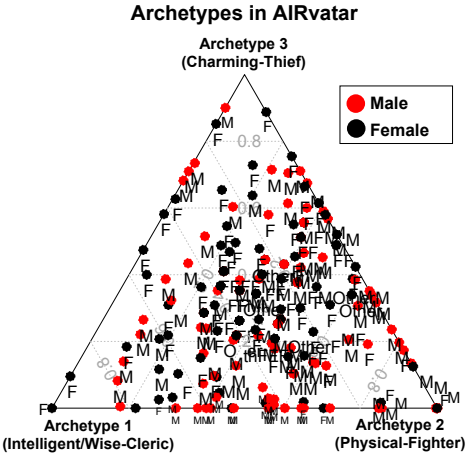
C TERNARY PLOTS



(a) Oblivion



(b) Ultima IV



(c) AIRvatar

Figure 2. Ternary plots representing characters as mixtures of archetypal archetypes in *The Elder Scrolls IV: Oblivion*, *Ultima IV*, and from our *AIRvatar* system. Labels for (a) denote races in *Oblivion*, (b) denote names in *Ultima IV*, and (c) player gender in *AIRvatar*.

PALAIS: A 3D Simulation Environment for Artificial Intelligence in Games

Patrick Schwab and **Helmut Hlavacs**[1]

Abstract. In this paper we present PALAIS — a virtual simulation environment for Artificial Intelligence (AI) in games. The environment provides functionality for prototyping, testing, visualisation and evaluation of game AI. It allows definition and execution of arbitrary, three-dimensional game scenes and behaviors. Additionally, PALAIS incorporates a plugin system that supports swift integration of custom AI algorithms. As a result, PALAIS effectively reduces the effort necessary to research, develop, prototype and showcase behaviors used for non-player characters in games. Finally, we demonstrate the power of the provided plugin system by exemplarily extending the functionality of PALAIS with an external module. PALAIS is available at http://www.palais.io.

1 INTRODUCTION

The development of game AI typically requires a testbed environment to validate and visualise results in a virtual-world scenario. Game developers and researchers frequently employ either game engines or custom-coded game scenes as their testbed environments. Using these environments for simulation has several disadvantages: suboptimal code reuse, significant barriers to entry and increased development time over using a more domain-specific environment. PALAIS attempts to solve these issues by providing commonly required functionality, such as a graphical user interface (GUI), loading required assets, data visualisation, scripting, entity management and rendering, in an existing, accessible framework. Having this framework in place enables the user to focus her efforts on AI-related code.

Moreover, custom-built solutions are often not easily distributed. We propose a container format that stores all scene-related assets in standardised formats. In PALAIS these scene containers are called *scenarios*. Any instance of PALAIS can execute these scenarios. The scenario structure, which is further described in section 3, and its distribution process is depicted in figure 1. The scenario structure allows users to share their scene definitions, graphical assets and game AI. This simplified distribution process gives others the opportunity to learn from, and build on, existing work. Consequently, our tool is also suitable for use in game AI education. Teachers can utilise the provided environment to supply students with interactive demonstrations of game AI techniques. We believe this form of hands-on education, where students can monitor and adapt execution parameters in actual game scenarios, can significantly increase the accessibility of game AI. Similarly, the simulation environment can serve as a demonstration platform for researchers to showcase their algorithms and techniques.

[1] University of Vienna, Faculty of Computer Science, Research Group Entertainment Computing, Austria, email: a0927193@unet.univie.ac.at and helmut.hlavacs@univie.ac.at

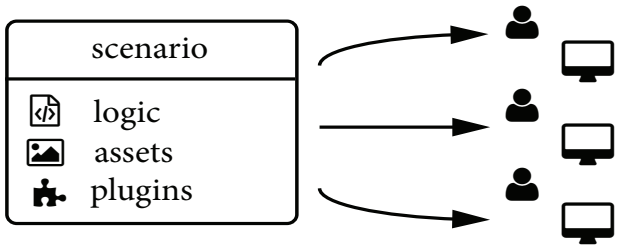


Figure 1. A schematic overview of the scenario structure and its distribution.

2 RELATED WORK

As mentioned, game developers and researchers commonly turn to commercial [15][7], open-source [14] or in-house engines for AI simulation. These general game engines overlap in functionality with PALAIS, particularly in the 3D rendering domain. PALAIS is more suitable for the simulation of game AI, because it provides the domain-specific functionality required for game AI development. Other toolkits, such as MASON [11], BREVE [10] and NetLogo [17], also provide full simulation environments. A significant drawback of some of the listed alternative simulation toolkits is the lack of extensibility via native code. Game developers strive to reach the maximum performance possible with the available computational resources. Thus, time-critical AI code for games is frequently written in native code. Our proposed simulation environment pays tribute to this by offering a plugin system [6] that allows extension through native, dynamically loaded libraries. The plugin system enables developers to test, prototype and evaluate the same native code that they use in their game engine. Ultimately, the ability to interface with native plugins also leads to more independent AI code compared to alternative simulation environments, because only the minimal necessary application programming interface (API) is exposed to plugins. Although the level of abstraction is not as high as it is with realisation-independent approaches. For example, [16] present such an realisation-independent approach.

Additionally, PALAIS provides a scripting API to increase its general accessibility and suitability for rapid prototyping. The scripting API is accessed via ECMAScript [5]. ECMAScript is one of the most widely-understood programming languages. Its most notable implementation is JavaScript, which is used to perform client-side scripting in Internet browsers. As a result of its prevalence, ECMAScript is a natural choice to provide scripting functionality in PALAIS.

To summarise, compared with the mentioned, existing works, the key distinguishing features of PALAIS are domain-specific functionality, interactivity, accessibility and extensibility.

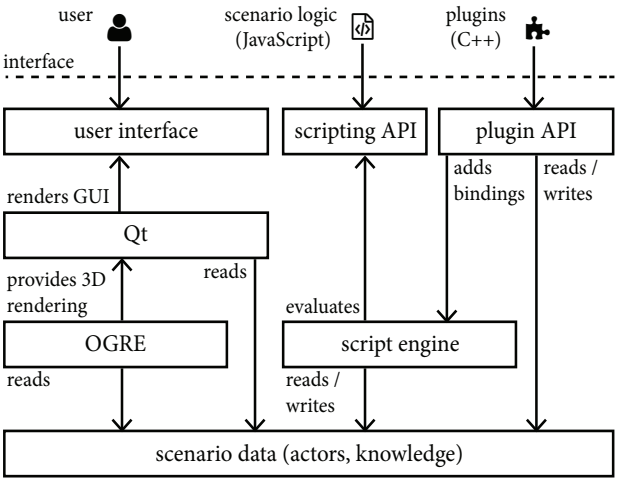


Figure 2. A schematic overview of the most significant interactions between the internal components of the simulation environment and its external accessors.

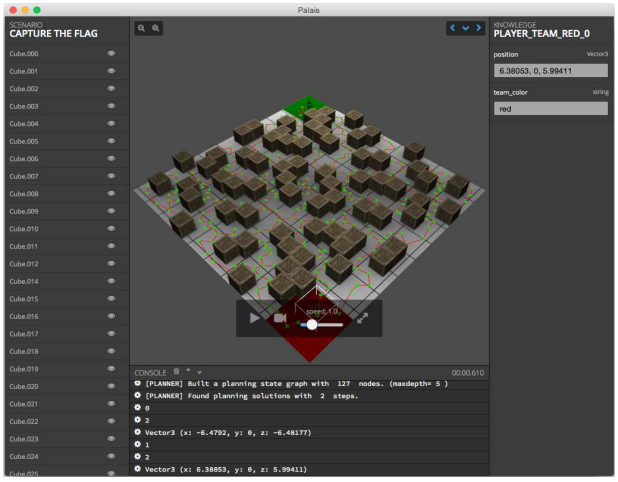


Figure 3. The GUI of PALAIS after loading a scenario. The left panel lists all active actors in the game scene. The right panel shows the knowledge inspector. The center panel displays a rendering of the scene itself.

3 SCENARIO STRUCTURE

Scenarios are the entity corresponding to a given game scene in PALAIS. They encapsulate specific game situations defined by users. The common use case is to define scenarios that provide a minimal environment for evaluation of AI behaviors and algorithms. Essentially, these scenarios are self-contained packages that include the assets, logic scripts and plugins necessary to execute a game scene. The following sections describe the components of a scenario.

3.1 Assets

The term 'assets' in the context of scenarios refers to all scene-related data files that don't contain, native or interpretable, code. Typically, assets mainly consist of the files needed for rendering the scene, such as 3D mesh data, textures and materials. PALAIS can load scene files created with external 3D modelling tools like [1]. However, PALAIS currently only supports the scene and mesh formats native to OGRE.

3.2 Logic Scripts

Logic scripts are the files containing ECMAScript code. PALAIS interprets these files at runtime. Since no compilation is required, the user can simply reload scripts after changes. The ability to reload scripts allows for frictionless development of behaviors, as the results of code changes can be evaluated quickly.

3.3 Plugins

Plugins are the other group of code attached to a game scene. Plugins, unlike logic scripts, contain compiled code. Plugins are standard shared libraries. Their specific file format depends on the operating system (OS) and the processor architecture for which the code was compiled. Relying on platform-specific formats impedes the portability of scenarios across platforms. However, we accept this price to support the integration of precompiled code. In practice, this means that a scenario must contain plugins compiled for every required target platform.

4 PROGRAMMING MODEL

We call programmable entities within a scenario in PALAIS *actors*. A generic key-value store, labeled *blackboard*, represents the individual knowledge of every actor. As the naming suggests, blackboard systems [3] inspired this form of knowledge representation. We chose a blackboard architecture because it offers flexibility and is conceptually easy to grasp and use for developers. To represent global knowledge, the game scene itself incorporates a blackboard as well. For visualisation, all actors must be connected to a rendered object in the 3D game scene. PALAIS implicitly makes all rendered objects within a game scene available as actors. Additionally, native or interpreted code can instantiate new actors at runtime.

4.1 Time Simulation

All code instances, native and interpreted alike, receive notifications of time advances. These tick events are independent of the frame rate of the simulation and represent fixed, simulated time steps. PALAIS adjusts the simulation speed by adapting the rate at which it emits these tick events relative to the passed time. This ensures the simulation results are the same, regardless of simulation speed.

5 INTERFACES

Figure 2 depicts a general overview of the interfaces of PALAIS. PALAIS exposes several external interfaces to fulfil the previously mentioned requirements.

5.1 Graphical User Interface

For users, the main external interface is the graphical user interface (GUI) provided by the runtime of PALAIS. Its main purpose is to display the data related to the currently active scenario. Most importantly, it displays the current state of the scenario in a 3D game scene. We integrated the open-source rendering engine OGRE [14] with the Qt framework [4] to provide a cross-platform GUI and 3D view. The GUI (figure 3) allows the user to configure certain rendering parameters, such as the camera's 3D orientation, zoom level and viewing

direction. The user can also view blackboards of the scenario and actors in the knowledge inspector panel of the GUI.

5.2 Scripting API

The scripting API is another external interface of PALAIS. The scripting layer is primarily meant to enable definition of arbitrary scenario logic as well as to facilitate rapid prototyping of algorithms and behaviors. PALAIS integrates a scripting engine to interpret ECMAScript code. The scripting API provides access to the currently loaded scenario and its actors. Scripts are able to read and write knowledge to the blackboards of the scenario and the actors. Lastly, scripts can consume core functionality provided by the runtime environment, e.g. dynamic actor instantiation, destruction and ray casting.

5.3 Plugin API

The last external interface to access PALAIS is the plugin API. The plugin system allows dynamic loading of third-party code. This core feature makes PALAIS suitable for integration of existing, custom AI code. The plugin API offers the same functionality as the scripting API, plus some more advanced features. Also, plugins are able to expose their functionality to the scripting layer by installing custom bindings. Custom bindings allow the use of arbitrary interaction patterns between native code in plugins and interpreted code in scripts.

5.4 Using Interpreted or Native Code in PALAIS

In essence, either scripting or plugins can be used to implement the same resulting scene logic. In fact, internally, the scripting interface is simply another layer on top of the same functionality. There is a performance overhead associated with the use of the the scripting layer, due to the additional code interpretation. Practically, that overhead means that computationally intensive tasks and tasks that run multiple times per time tick are more suited for implementation as plugins. Thus, the suggested workflow is to make all computationally intensive tasks available to the scripting layer via bindings. The extended scripting API can then be used to orchestrate the scene-specific logic.

6 INTEGRATING AN EXTERNAL MODULE

To demonstrate the power of its extension system we extended PALAIS with an external pathfinding module. The module is based on the A* search algorithm [9]. Our implementation of the pathfinding system follows the one described in [12]. A* pathfinding is a technique for determining shortest paths. It allows non-player characters (NPCs) to navigate game worlds. In this role, A* pathfinding is part of the standard repertoire of AI in games. Therefore, it is well-suited to serve as an example for exhibiting the potential of PALAIS. In particular, adding the functionality of the pathfinding module to PALAIS shows how easily existing AI code can be integrated with its environment.

6.1 Pathfinding Module

The pathfinding module provides methods for constructing and searching shortest paths on navigation graphs. As is typical for game middleware, the module is implemented in C++. The compiled, executable code is in binary form. It contains native code that depends on the processor architecture. Consequently, to integrate the module, we must exploit the ability of PALAIS to load native code as plugins.

6.2 Plugin Integration Workflow

A shared library must conform to a simple, well-defined interface to be loadable in the plugin system of PALAIS. In the current version of PALAIS, said interface consists of just 5 methods. Specifically, it consists of two methods corresponding to the loading and tear-down of the plugin, two methods corresponding to the loading and tear-down of a scenario and one method realising the time tick notification. The methods for the loading and tear-down of plugins give plugins an opportunity to initialise and destroy any general setup structures they require. Similarly, the methods for the loading and tear-down of scenarios can be used to initialise and destroy per-scenario bookkeeping information and to install script bindings with the script engine of the scenario. Finally, the time tick event initiates all time-dependent or regularly scheduled functionality. As a complementary measure, the user can register script bindings to define additional entry points.

6.2.1 Example

As is the case with most custom AI code, our pathfinding module does not conform to the plugin interface. Adapting existing code to the defined interface is the integration effort required to make the functionality of a plugin available to PALAIS. We employ the adaptor design pattern [8] to adapt the interface of our pathfinding module to the interface required by the plugin system of PALAIS. The following steps are necessary to integrate the pathfinding module:

1. First, we use the method corresponding to the initialisation of a scenario to load the navigation mesh of the currently active scenario. A navigation mesh [12] is a continuous representations of the walkable area in a game scene. After loading, the pathfinding module constructs a navigation graph from this navigation mesh. The resulting navigation graph can be searched in response to navigation requests. Furthermore, we install a script binding to make the pathfinding functionality available to scripts. These are the per-scenario steps necessary to provide a pathfinding service.
2. Next, we implement the process of searching a path. The first step in this process is initiated by script code calling the plugin via the binding registered previously. In response, the pathfinding system writes the shortest path to the blackboard of the actor that requested the shortest path.
3. Lastly, we add the actual actor movement according to the plans stored in their blackboards. For this, we use the time tick event: We sequentially check the blackboard of every actor for remaining paths to determine which actors in the current scenario must be moved. Finally, we remove a path node from the blackboard, once the actor that it belongs to reaches it.

This example demonstrates the potency of the blackboard architecture used in PALAIS. Due to the blackboard architecture the plugin system requires only a minimalist plugin interface. As a result, the blackboard architecture effectively decreases the effort required to integrate existing AI code with PALAIS.

6.3 Data Visualisation

Procedures for the in-scene visualisation of data are part of the core functionality of PALAIS. In addition to providing rendering

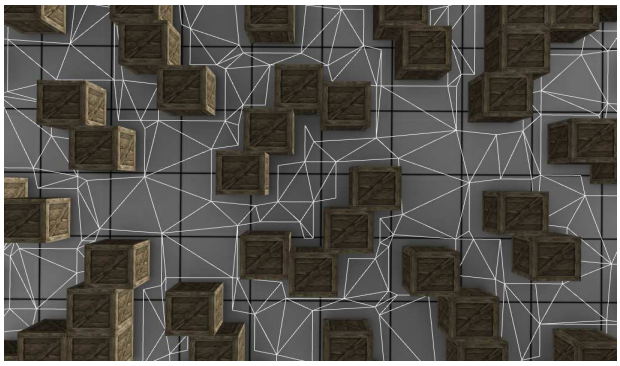


Figure 4. A rendering in PALAIS showing the navigation mesh used by the pathfinding module.

Figure 5. A rendering in PALAIS showing the navigation graph constructed from the navigation mesh in figure 4.

of arbitrary textured meshes, PALAIS provides means for rendering coloured primitives, such as lines, circles, quads, cuboids and spheres. As an example, the pathfinding module renders the navigation graph using the visualisation primitives of PALAIS. Figure 4 and figure 5 depict renderings of the navigation mesh and the navigation graph in PALAIS.

6.4 Accessing the Pathfinding Module

The plugin installs its script bindings when a scene is loaded. In our pathfinding example, all scripts in a scenario, that includes the pathfinding plugin, can invoke the process to navigate an actor to a goal along a shortest path. The script delegates the computation and handling of the movement to the plugin. This abstraction provided by plugins also allows the reuse of plugins in different scenarios.

7 CONCLUSION

PALAIS is a powerful environment for the simulation of AI in games. It caters specifically to the needs of game developers by granting access to its programming interface via interpreted and native code. Our exemplary integration of an external pathfinding module demonstrates that PALAIS is an apt choice for the simulation of scenes that depend on third-party AI libraries. Additionally, the ability to extend PALAIS with plugins lowers the barrier to entry for the usage of the simulation environment, since the same native code, that is used for the simulation in PALAIS, can easily be shared with game engines.

8 FUTURE WORK

The work on the simulation environment PALAIS is part of a larger, ongoing project to build a unified framework for game AI development. The framework includes functionality for each of the layers of the game AI model proposed in [12]. Particularly, it encompasses algorithms that facilitate the implementation of movement, decision making and strategy for non-player characters in games. Pathfinding, Behavior Trees [2] and Goal-Oriented Action Planning (GOAP) [13] are among the standard techniques the framework implements. These techniques will be integrated with PALAIS in the form of plugins to provide users with a solid foundation that allows the rapid development of AI behaviors. On the feature side, future work on PALAIS could involve refinement by adding support for physics-based dynamics and statistical evaluation of behaviors.

REFERENCES

[1] Blender Online Community. Blender - a 3D modelling and rendering package. Retrieved from http://www.blender.org.

[2] Alex Champandard, 'Behavior trees for next-gen game AI', in *Game Developers Conference, Audio Lecture*, (2007).

[3] Daniel D Corkill, 'Blackboard systems', *AI expert*, **6**(9), 40–47, (1991).

[4] Digia Plc. Qt: cross-platform application and UI framework, 2012.

[5] ECMA International, *Standard ECMA-262 - ECMAScript Language Specification*, 5.1 edn., June 2011.

[6] Martin Fowler, *Patterns of Enterprise Application Architecture*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

[7] Epic Games. Unity engine documentation. Retrieved from https://www.unrealengine.com/.

[8] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides, *Design patterns: elements of reusable object-oriented software*, Pearson Education, 1994.

[9] Peter E Hart, Nils J Nilsson, and Bertram Raphael, 'A formal basis for the heuristic determination of minimum cost paths', *Systems Science and Cybernetics, IEEE Transactions on*, **4**(2), 100–107, (1968).

[10] Jon Klein, 'Breve: a 3d environment for the simulation of decentralized systems and artificial life', in *Proceedings of the eighth international conference on Artificial life*, pp. 329–334, (2003).

[11] Sean Luke, Claudio Cioffi-Revilla, Liviu Panait, Keith Sullivan, and Gabriel Balan, 'Mason: A multiagent simulation environment', *Simulation*, **81**(7), 517–527, (2005).

[12] Ian Millington and John Funge, *Artificial intelligence for games*, CRC Press, 2009.

[13] Jeff Orkin, 'Applying goal-oriented action planning to games', *AI Game Programming Wisdom*, **2**(2004), 217–227, (2004).

[14] Torus Knot Software. Object-oriented graphics rendering engine (OGRE) Engine documentation. Retrieved from http://www.ogre3d.org/.

[15] Unity Technologies. Unity documentation. Retrieved from http://unity3d.com/.

[16] Marco Vala, Guilherme Raimundo, Pedro Sequeira, Pedro Cuba, Rui Prada, Carlos Martinho, and Ana Paiva, 'ION framework–a simulation environment for worlds with virtual agents', in *Intelligent virtual agents*, pp. 418–424. Springer, (2009).

[17] Uri Wilensky, 'Netlogo', http://ccl.northwestern.edu/netlogo/, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, (1999).

Simulating Autonomous Non-Player Characters in a Capture the Flag Scenario Using PALAIS

Patrick Schwab and **Helmut Hlavacs**[1]

Abstract. PALAIS is a 3D simulation environment for artificial intelligence (AI) in games. It has built-in support for much of the standard functionality required when simulating AI behaviors. Most importantly, PALAIS allows users to define their own arbitrary game scenes with custom game rules. This paper presents the workflow of authoring game scenes in PALAIS by the example of a Capture the Flag scene. In particular, we demonstrate how users can take advantage of the provided scripting layer to rapidly define their simulation logic. This paper also serves as a description of the content of the accompanying demonstration given at the conference.

1 Simulation Environment

Game scenes in PALAIS are defined in packages called *scenarios*. These scenarios contain all code and graphical assets required for the simulation of the game scene. Users define the visual appearance of scenarios in an external 3D modelling tool. At runtime, users can access the functionality of PALAIS via a scripting or a native programming interface. The scripting interface can be accessed from the ECMAScript [2] programming language. Additionally, users can extend the functionality available to scripts by utilising the plugin system [3] incorporated in PALAIS. The combination of plugins and scripts allows for the definition of rich interaction patterns.

PALAIS automatically creates a blackboard [1] for each actor in a scenario. This form of knowledge representation provides a very flexible means of managing the data flow between the different components of a scenario. The contents of the blackboards of each actor can be examined during the simulation of a scenario. Figure 1 shows the knowledge inspector in action.

2 Capture the Flag Scenario

We chose a Capture the Flag Scenario as our exemplary game scenario. The Capture the Flag scenario involves two opposing teams. Each team has to capture the flag of the opposing team to score points. Characters can capture a flag by taking it from the initial spawning point of the opposing team to the initial spawning point of their own team. Implementing AI for non-player characters in a Capture the Flag scenario is a standard problem in game AI. Thus, it is well-suited to showcase the abilities of PALAIS. The arena of the implemented Capture the Flag scenario is shown in figure 2.

[1] University of Vienna, Faculty of Computer Science, Research Group Entertainment Computing, Austria, email: a0927193@unet.univie.ac.at and helmut.hlavacs@univie.ac.at

3 Authoring Workflow

To implement the Capture the Flag scenario we employ plugins that provide standard algorithms of game AI. These plugins allow us to delegate computationally intensive tasks, such as pathfinding, to native code. We use the scripting interface of PALAIS to orchestrate the actors of the scenario and to define the possible actions they can take.

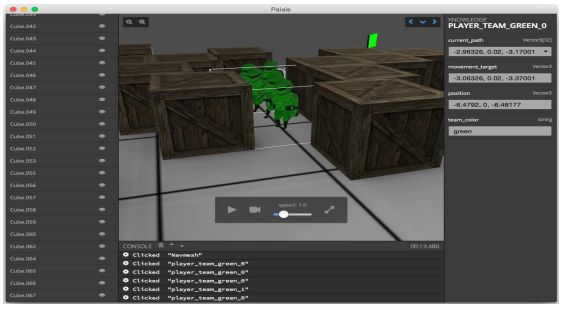


Figure 1. A demonstration of the live inspection of blackboards available in PALAIS. The panel on the right shows the contents of the blackboard of the frontmost actor of the green team.

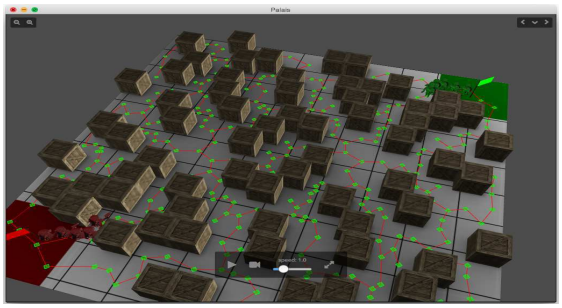


Figure 2. A rendering in PALAIS that shows the arena of the Capture the Flag scenario.

REFERENCES

[1] Daniel D Corkill, 'Blackboard systems', *AI expert*, **6**(9), 40–47, (1991).
[2] ECMA International, *Standard ECMA-262 - ECMAScript Language Specification*, 5.1 edn., June 2011.
[3] Martin Fowler, *Patterns of Enterprise Application Architecture*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

EmohawkVille: Virtual City for Everyone

David Holaň and **Jakub Gemrot** and **Martin Černý** and **Cyril Brom** [1]

Abstract. Despite recent progress, behavior of non-player characters (NPCs) in contemporary games is still kept rather simple. This is an opportunity for the academia to develop novel techniques and tools that would allow for easier creation of complex behaviors that are resilient to the dynamicity implied by the presence of the player. There already exist languages within multiagent community that are thought to be suitable for NPC behaviors creation, but they are usually tested in simplistic environments and our experience indicates that applying them to complex 3D worlds introduces significant obstacles. This is part of the reason why simple reactive techniques are prevalent in game industry practice. Moreover there is no publicly available research-friendly 3D virtual world with sufficient complexity that would allow developers to evaluate their languages and tools in a more realistic setting and improve them toward practical applicability. In this demo we present EmohawkVille: an open-source first-person 3D virtual world that is a candidate for such an environment.

1 Introduction

Many contemporary computer games take a great effort to achieve a high level of believability of their virtual worlds. This is especially true for games with large open worlds, where the user is free to discover the environment on his own and is relatively unconstrained by the game. One of the challenges that arise in this scenario is the problem of choosing the right higher-level action for the NPCs (e.g., move to a point, pick up an item, use an item, ...). Since the game industry relies almost exclusively on simple reactive techniques which make creation of complex behaviors rather time-consuming and costly, non-player characters (NPCs) display complex behaviors only during crucial game events. In between, the NPC behaviors are schematic at best.

The main issue is that going beyond simple behavior and still maintaining the suspension of disbelief introduces significant difficulties to the NPC behavior authoring. There are many possible obstacles to NPC goals and if they are not taken into account, the NPCs are easy for the player to "break" and may provide even worse illusion of a real world than rather static NPCs.

For a truly alive open world, dozens of different and often complex scenarios are needed, which implies that the world needs to be equipped with a rich ontology of items and actions NPCs (as well as the player) can perform.

As the world ontology grows, the number of meaningful NPC action sequences increases and the behavior complexity rises. Not only the means-ends analysis becomes more demanding, new problems emerge such as transitional behaviors, joint behaviors, behaviors ordering or behaviors interleaving [6]. At the same time, game studios usually cannot afford to let an expert AI programmer design such day-to-day behaviors, because that would be cost-prohibitive. Most of the NPC design is thus usually carried out with the aid of some visual tool by scripters with little programming experience.

At this place, academia could provide action selection mechanisms (ASM) and accompanying tools that would help inexpert scripters to create complex behaviors that are *interactively believable*, that is, behaviors that sustain their believability under nondeterminism brought by the player. However, most of the academic research is carried out in environments that either have simple ontologies or are static or discrete. Games on the other hand are dynamic, multi-agent environments that can be for all practical purposes considered continuous in both time and space. There are languages and techniques that can be applied to such worlds: either from the multiagent community or the field of robotics or automated planning. However, to our knowledge, there is currently no 3D virtual world publicly available that would provide rich ontology for NPCs out of the box. This means that in this particular problem area, academia is one step behind the industry — we do not even have an environment to work with.

Note that raw frameworks such as Unity [7] are not sufficient as creating a rich world in a raw framework is a substantial amount of work. An important part of the environment is also the possibility to develop the NPC behavior with a high-level language such as Java since nearly all agent languages of interest can be invoked from Java code. We are not aware of any complex 3D environment that would meet all those requirements. See our paper [4] for a thorough comparison of possible candidates.

Previous research has shown that applying agent languages to 3D environments is neither straightforward nor guaranteed to yield better results than using a general programming language [2, 5]. Common issues with agent languages are incomplete debugging and tool support, some of the architectures are also hard to debug in principle (e.g., because of inherent paralellism). Many agent languages are also declarative in nature, while game worlds feature lots of mechanics that are hard to express declaratively (e.g., determining which object is hit by an arrow). Proper evaluation of agent languages is thus critical.

In this demo, we present an extension of the Pogamut 3 platform [3] called EmohawkVille, the first step towards an open-sourced complex simulation of NPC everyday life in 3D virtual world. We believe that creating a fully working, accessible and polished environment fosters academic progress. The large amount of research work evaluated on Pogamut for Unreal Tournament 2004 supports this view. We have also exerted great effort to make EmohawkVille a mature tool. In practice, there is a long chain of components that are needed to fully connect high-level AI with an NPC: sensors and actuators interface, navigation and pathfinding, character animation support are among the most important, but the list is far from exhaus-

[1] Charles University in Prague, Czech Republic email: {paladin.invictus,jakub.gemrot,cerny.m}@gmail.com, brom@ksvi.mff.cuni.cz

tive. In EmohawkVille, we have resolved large part of those issues on behalf of the researcher. The quality of the EmohawkVille environment was evaluated in a small-scale user study and by use of the environment in our teaching curriculum.

2 General Description

EmohawkVille is a first-person virtual world with detailed interactive elements of day-to-day life. There is a general framework that supports interaction with items, continual actions and processes and inter-agent communication including trade. There is a set of ready-made assets for a cooking scenario. For example, an agent or a human player can pick up a piece of meat, put it on a chopping board and slice it and then fry the slices on a pan (charring the food if he does not add oil or forgets to flip the meat). The cooking scenario was chosen as our first because it features plethora of complex procedures yet it is easy to grasp by programmers and non-programmers alike and is gender-neutral.

EmohawkVille is based on Unreal Development Kit (UDK) [1] and thus is capable of displaying the world in state-of-the-art graphics. UDK is free for educational and non-commercial use and EmohawkVille itself is available under GPLv3[2].

In EmohawkVille the world mechanics are implemented in UnrealScript - a proprietary language deployed with the UDK toolkit. The Pogamut platform provides a high-level Java interface to the UDK for writing the actual AI and takes care of many common tasks (pathfinding with A* and smooth path following, caching sensory data to a blackboard, etc.). Both the UDK and the Java part have been designed with possible further extensions in mind and the basic NPC support is separated from the model of the general EmohawkVille ontology, which is in turn separated from the implementation of the specific mechanics for our cooking scenario. The UDK part also fully supports interaction with a human user through the UDK visual client.

At this moment, EmohawkVille features 20 item types (food, cutlery, cooking tools, ...) and a cooking stove (part of the environment). Interaction is provided by 14 actions, of which nine are instant and four initiate a longer-lasting process, e.g., chopping a vegetable or stirring a broth. An overview of the available items is visible in Figure 1.

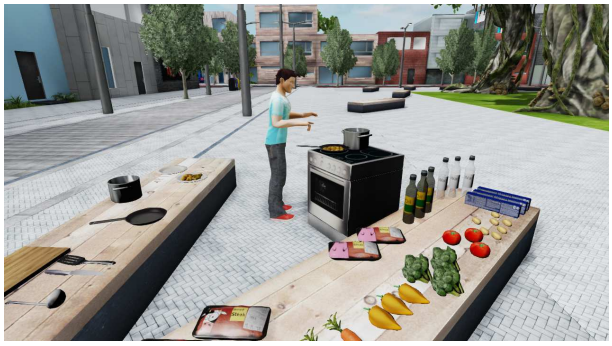


Figure 1. A screenshot of the environment.

The central complexity of the NPC behavior stems from the simulation of cooking. Some ingredients can be boiled, some fried. The speed of cooking is determined by the temperature of respective stoves. Water evaporates from pots and ingredients may burn or char if not stirred or flipped in the pot or the pan. The cooking theme provides important challenges to the NPC behavior creation: cooking a meal may require a long sequence of actions (more than 20), effectivity is increased by performing processes in parallel possibly requiring cooperation of multiple chefs, the player may both support and sabotage the cooking NPC.

Every aspect of the environment and the agents is programmable. EmohawkVille is ready for a researcher to plugin any high-level decision making mechanism (planning, machine learning, ...) without the need to handle low-level details. More detail of the environment is given in our paper [4].

3 Demo Presentation

In our demo presentation we would like to show the environment and its richness, let the spectators interact with the environment themselves, helping a preprogrammed agent to cook a complex meal or sabotaging his effort. We would also like to show that programming the behaviors is easy and EmohawkVille thus lets the researcher focus on the action selection exclusively. This will be demonstrated by a live creation of a cooking NPC and we would enable hands-on programming experience to the spectators.

A video presentation of the environment may be found at `http://www.youtube.com/watch?v=G7lKXkR2Xgg`

ACKNOWLEDGEMENTS

This research was supported by SVV project number 260 224.

REFERENCES

[1] Epic Games Inc. Unreal development kit documentation. `http://www.unrealengine.com/udk/documentation/`, 2009. Last checked: 2015-03-30.

[2] Jakub Gemrot, Zdeněk Hlávka, and Cyril Brom, 'Does high-level behavior specification tool make production of virtual agent behaviors better?', in *Proceedings of CAVE'12*, pp. 167–183, Berlin, Heidelberg, (2013). Springer-Verlag.

[3] Jakub Gemrot, Rudolf Kadlec, Michal Bída, Ondřej Burkert, Radek Píbil, Jan Havlíček, Lukáš Zemčák, Juraj Šimlovič, Radim Vansa, Michal Štolba, Tomáš Plch, and Cyril Brom, 'Pogamut 3 can assist developers in building ai (not only) for their videogame agents', in *Agents for Games and Simulations*, eds., Frank Dignum, Jeff Bradshaw, Barry Silverman, and Willem Doesburg, LNCS 5920, 1–15, Springer-Verlag, (2009).

[4] Gemrot J. Černý M. Holaň, D., 'Emohawkville: Towards complex dynamic virtual worlds', in *Proceedings of GAMEON'2013*, pp. 52–58, (2013).

[5] Radek Píbil, Peter Novák, Cyril Brom, and Jakub Gemrot, 'Notes on pragmatic agent-programming with Jason', in *Programming Multi-Agent Systems*, LNCS 7217, 58–73, Springer, (2012).

[6] Tom Plch, *Action selection for an animat*, Master's thesis, Charles University in Prague, 2009.

[7] Unity Technologies. Unity documentation. `http://docs.unity3d.com/Documentation/Manual/`, 2005. Last checked: 2015-03-30.

[2] EmohawkVille may be downloaded from `http://pogamut.cuni.cz/main/tiki-index.php?page=EmohawkVille`

An interactive, generative Punch and Judy show using institutions, ASP and emotional agents

Matt Thompson [1] and **Julian Padget** and **Steve Battle**

Abstract. Using Punch and Judy as a story domain, we describe an interactive puppet show, where the flow and content of the story can be influenced by the actions of the audience. As the puppet show is acted out, the audience reacts to events by cheering or booing the characters. This changes the emotional state of each agent, potentially causing them to change their actions, altering the course of the narrative. An institutional model is used to ensure that the narrative is constrained to remain consistent with the Punch and Judy canon.

1 Introduction

Agent-based approaches for interactive narrative generation use intelligent agents to model the characters in a story. The agents respond to the interactions of a player with dialogue or actions fitting the shape of a story. However, these agents have little autonomy in their actions, bound as they are to the strict requirements of their role in the narrative.

An institutional model can be used as normative framework for governing the actions of agents in a story. By describing the rules of a narrative in terms of social expectations, the agents are encouraged to perform certain types of actions while still remaining free to break free of these expectations. As in society in the real world, breaking agreed norms comes with consequences, and only generally happens in exceptional circumstances.

One situation where this is desirable is with the use of emotional agents. An agent experiencing an extreme emotion in an emotional model (such as rage or depression) may be allowed to act unusually or uncharacteristically. Allowing characters to break from narrative norms enables them to be 'pushed too far' by circumstances, with results that add an extra dimension of richness to a story.

Through this implementation, we introduce two novel approaches: (i) the use of an institutional model to describe a narrative 'world' or domain, and (ii) how emotional models can give intelligent agents some degree of autonomy to both act in idiosyncratic ways and to react emotionally to input from the audience.

The puppets in the show are each belief-desire-intention (BDI) agents with a valence, arousal, dominance (VAD) emotional model described in section 5. The story is modelled by a set of institutional norms (section 6.1) that describe the Punch and Judy story domain in terms of Propp's 'story moves' [8] (section 3). The agents communicate with their environment using the Bath Sensor Framework, described in section 6.3 [6]. In the final sections, we describe the animation system that functions as the agents' environment (section 6.4), and how the audience interacts with the system (section 7).

[1] University of Bath, United Kingdom, email: m.r.thompson@bath.ac.uk

2 Propp moves and roles

To express story events as an institution, we must look to narrative theory for inspiration. Instead of describing parts of the Punch and Judy story explicitly (such as 'Punch is expected to hit the policeman in this scene'), it is more desirable to describe scenes in a more abstract way ('The villain fights the victim in this scene'). The use of more general story components allows us to reuse them in multiple scenes, or even in other stories.

Narratology, and structuralism in particular, supply such generalised building blocks for stories. Russian formalism is an early movement in narrative theory to formalise the elements of narrative, of which Vladimir Propp is a prominent figure.

In order to direct the course of the narrative, we use a model built upon Propp's 1928 formalism of Russian folktales, *The Morphology of the Folktale* [8]. In this formalism, Propp identifies recurring characters and motifs in Russian folklore, distilling them down to a concise syntax with which to describe stories.

In this formalism, characters have *roles*, such as *hero*, *villain*, *dispatcher*, *false hero*, and more. Characters performing a certain role are able to perform a subset of *story moves*, which are actions that make the narrative progress. For example, the *dispatcher* might send the *hero* on a quest, or the *victim* may issue an *interdiction* to the *villain*, which is then *violated*.

Propp defines a total of 31 distinct story functions, some of which can have subtle variations from story to story. Each function is given a number and symbol in order to create a succinct way of describing entire stories. Examples of such functions are:

- One of the members of a family absents himself from home: *absentation*.
- An interdiction is addressed to the hero: *interdiction*.
- The victim submits to deception and thereby unwittingly helps his enemy: *complicity*.
- The villain causes harm or injury to a member of the family: *villainy*.

Each of these functions can vary to a great degree. For example, the *villainy* function can be realised as one of 19 distinct forms of villainous deed, including *the villain abducts a person*, *the villain seizes the daylight*, and *the villain makes a threat of cannibalism*.

These functions are enacted by characters following certain roles. Each role (or *dramatis personae* in Propp's definition) has a *sphere of action* consisting of the functions that they are able to perform at any point in the story. Propp defines seven roles that have distict spheres of action: *villain*, *donor*, *helper*, *princess*, *dispatcher*, *hero*, and *false hero*.

In a typical story, one story function will follow another as the tale progresses in a sequential series of cause and effect. However, Propp's

formalism also allows for simultaneous story functions to occur at once.

2.1 Propp example: sausages and crocodile scene

The common elements of Punch and Judy are easily described in terms of Propp's story functions. Here we pick one scene from the Punch and Judy show to use as an example: the scene where Punch battles a crocodile in order to safeguard some sausages.

In this scene, Joey the clown (our narrator) asks Punch to guard the sausages. Once Joey has left the stage, a crocodile appears and eats the sausages. Punch fights with the crocodile, but it escapes. Joey then returns to find that his sausages are gone.

The appropriate story functions are:

1. Joey tells Punch to look after the sausages (*interdiction*).
2. Joey has some reservations, but decides to trust Punch (*complicity*).
3. Joey gives the sausages to Punch (*provision or receipt of a magical agent*).
4. Joey leaves the stage (*absentation*).
5. A crocodile enters the stage and eats the sausages (*violation*).
6. Punch fights with the crocodile (*struggle*).
7. Joey returns to find that the sausages are gone (*return*).

3 Institutional model

An institution describes a set of 'social' norms describing the permitted and obligated behaviour of interacting agents. Noriega's 'Fish Market' thesis [7] describes how an institutional model can be used to regiment the actions of agents in a fish market auction. Cliffe [3], Baines and Lee [6] extend this idea to build systems where institutions actively regulate the actions of agents, while still allowing them to decide what to do. Adapting this idea to the world of narrative, we use an institutional model to describe the story world of Punch and Judy in terms of Propp moves and character roles.

Institutional models use deontic logic to describe obligations and permissions that act on interacting agents in an environment. By combining this approach with Propp's concepts of *roles* and *story moves*, we describe a Propp-style formalism of Punch and Judy in terms of what agents are *obligated* and *permitted* to do at certain points in the story.

For example, in one Punch and Judy scene a policeman enters the stage and attempts to apprehend Punch. According to the rules of the Punch and Judy world, Punch has an obligation to kill the policeman by the end of the scene (as this is what the audience expects to happen, having seen other Punch and Judy shows). The policeman has an obligation to try his best to catch Punch. Both agents have permission to be on the stage during the scene. The policeman only has permission to chase Punch if he can see him (Punch is obligated to hide from him at the start of the scene).

The permissions an agent has constrain the choices of actions available to them at any given moment. Obligations affect the goals of an agent. Whether or not an agent actively tries to fulfil an obligation depends on their emotional state.

3.1 Institution example

Here we continue the 'sausages and crocodile' scene example from section 3.1, taking the Propp story functions and describing them as an institutional model.

We define our institution in terms of *fluents*, *events*, *powers*, *permissions* and *obligations*.

3.1.1 Fluents

Fluents are properties that may or may not hold true at some instant in time. *Institutional events* are able to *initiate* or *terminate* fluents at points in time. A fluent could describe whether a character is currently on stage, the current scene of a story, or whether or not the character is happy at that moment in time.

Domain fluents ($\mathcal{D}$) describe domain-specific properties that can hold at a certain point in time. In the Punch and Judy domain, these can be whether or not an agent is on stage, or their role in the narrative (equation 1).

$$\mathcal{D} = \{\text{onstage, hero, villain, victim, donor, item}\} \quad (1)$$

Institutional fluents consist of *institutional powers*, *permissions* and *obligations*.

An **institutional power** ($\mathcal{W}$) describes whether or not an external event has the authority to meaningfully generate an institutional event. Using Propp as an example, an *absentation* event can only be generated by an external event coming from a *donor* character (such as their leaving the stage). Therefore, any characters other than the donor character would not have the institutional power to generate an *absentation* institutional event when they leave the stage.

Equation 2 shows a list of possible empowerments, essentially a list of institutional events.

$$\mathcal{W} = \{\text{pow(introduction, interdiction, give, absentation, violation, return)}\} \quad (2)$$

Permissions ($\mathcal{M}$) are external actions that agents are permitted to do at a certain instant in time. These can be thought of as the set of *socially permitted* actions available to an agent. While it is possible for an agent to perform other actions, societal norms usually prevent them from doing so.

For example, it would make sense in the world of Punch and Judy if Punch were to give the sausages to the Policeman. It is always Joey who gives the sausages to Punch. Also, it would be strange if Joey were to do this in the middle of a scene where Punch and Judy are arguing. We make sure agents' actions are governed so as to allow them only a certain subset of permitted actions at any one time. Equation 3 shows a list of permission fluents.

$$\mathcal{M} = \{\text{perm(leavestage, enterstage, die, kill, hit, give, fight)}\} \quad (3)$$

Obligations ($\mathcal{O}$) are actions that agents *should* do before a certain deadline. If the action is not performed in time, a *violation event* is triggered, which may result in a penalty being incurred. While an agent may be obliged to perform an action, it is entirely their choice whether or not they actually do so. They must weigh up whether or not pursuing other courses of action is worth suffering the penalty that an unfulfilled obligation brings.

Anybody who has seen a Punch and Judy show knows that at some point Joey tells Punch to guard some sausages, before disappearing offstage. Joey's departure is modelled in the institution as the *absentation* event. It could be said that Joey has an obligation to leave the stage as part of the *absentation* event, otherwise the story function is violated. Equation 4 shows how this would be described in the institution.

$$\mathcal{O} = \{\text{obl(leavestage, absentation, viol(absentation))}\} \quad (4)$$

3.1.2 Events

Cliffe's model specifies three types of **event**: *external events* (or 'observed events', $\mathcal{E}_{obs}$), *institutional events* ($\mathcal{E}_{instact}$) and *violation events* ($\mathcal{E}_{viol}$).

External events are observed to have happened in the agents' environment, which can *generate institutional events* which act only within the institional model, *initiating* or *terminating* fluents, permissions, obligations or institutional powers. An external event could be an agent leaving the stage, an agent hitting another, or an agent dying. Internal events include narrative events such as scene changes, or the triggering of Propp story functions such as *absentation* or *interdiction* (described in section 3). Violation events occur when an agent has failed to fulfil an obligation before the specified deadline. These can be implemented in the form of a penalty, by decreasing an agent's health, for example.

$$\mathcal{E}_{obs} = \{\text{startshow, leavestage, enterstage, die, give, harmed, hit, fight, kill, escape}\} \quad (5)$$

$$\mathcal{E}_{instact} = \{\text{introduction, interdiction, give, absentation, violation, return, struggle, defeat, complicity, victory, escape}\} \quad (6)$$

$$\mathcal{E}_{viol} = \{\text{viol(introduction), viol(interdiction), viol(give), viol(absentation), viol(violation), viol(return), viol(struggle), viol(defeat), viol(complicity) viol(victory), viol(escape)}\} \quad (7)$$

3.1.3 Event Generation and Consequences

An **event generation** function, $\mathcal{G}$, describes how events (usually external) can generate other (usually institutional) events. For example, if an agent leaves the stage while the *interdiction* event holds, they trigger the *leavestage* event. This combination generates the *absentation* institutional event (equation 11).

Event generation functions follow a $\langle$preconditions$\rangle \rightarrow \{$postconditions$\}$ format: $\langle \mathcal{G}(\mathcal{X}, \mathcal{E})\rangle \rightarrow \{\mathcal{E}_{out}\}$, where $\mathcal{X}$ is a set of fluents that hold at that time, $\mathcal{E}$ is an event that has occurred, and $\mathcal{E}_{out}$ are the events that are generated. They are generally used to generate internal, institutional events from external events.

Consider the Punch and Judy scenario described in section 3.1. There are seven institutional events (story functions) that occur during this scene: *interdiction*, *complicity*, *receipt* (from Propp's *receipt of a magical agent*) *absentation*, *violation*, *struggle*, *return*. These institutional events are all generated by external events. The *interdiction* is generated when Joey tells Punch to protect the sausages. Punch agreeing amounts to *complicity*. Joey *gives* punch the sausages (*receipt*), then leaves the stage (*absentation*). The crocodile eating the sausages is a *violation* of Punch's oath, the agents fight (*struggle*), then Joey enters the stage again (*return*).

It is desirable that these story function occur in this sequence in order for a satisfying narrative to emerge. Agents may decide to perform actions that diverge from this set of events, but the institution is guiding them towards the most fitting outcome for a *Punch and Judy* world. For this reason, a currently active story function can be the precondition for event generation. For example, the *receipt* event may only be triggered if an agent externally performs a *give* action **and** if the *complicity* event currently holds (equation 10).

Examples of event generation function for this scenario, complete with preconditions, are listed in equations 8 to 14.

$$\mathcal{G}(\mathcal{X}, \mathcal{E}) : \langle \emptyset, \textit{tellprotect}(\text{donor, villain, item})\rangle \rightarrow \{\textit{interdiction}\} \quad (8)$$

$$\langle \{\textit{interdiction}\}, \textit{agree}(\text{villain}))\rangle \rightarrow \{\textit{complicity}\} \quad (9)$$

$$\langle \emptyset, \textit{give}(\text{donor, villain, item}))\rangle \rightarrow \{\textit{receipt}\} \quad (10)$$

$$\langle \{\textit{interdiction}\}, \textit{leavestage}(\text{donor})\rangle \rightarrow \{\textit{absentation}\} \quad (11)$$

$$\langle \{\textit{interdiction}\}, \textit{harmed}(\text{item})\rangle \rightarrow \{\textit{violation}\} \quad (12)$$

$$\langle \{\textit{interdiction}, \textit{absentation}\}, \textit{enterstage}(\text{donor}),\ \text{onstage(villain)}\rangle \rightarrow \{\textit{return}\} \quad (13)$$

$$\langle \emptyset, \textit{hit}(\text{donor, villain})\rangle \rightarrow \{\textit{struggle}\} \quad (14)$$

Consequences consist of fluents, permissions and obligations that are *initiated* ($\mathcal{C}^{\uparrow}$) or *terminated* ($\mathcal{C}^{\downarrow}$) by institutional events. For example, the institutional event *give* could initiate the donor agent's permission to leave the stage, triggering the *absentation* event (equation 16). When the *interdiction* event is currently active and a *violation* event occurs, the interdiction event is terminated (21). Equations 15 to 22 describe the initiation and termination of fluents in the Punch and Judy sausages scenario detailed in section 3.1.

$$\mathcal{C}^{\uparrow}(\mathcal{X}, \mathcal{E}) : \langle \emptyset, \text{interdiction}\rangle \rightarrow \{\text{perm(give(donor, villain, item))}\} \quad (15)$$

$$\langle \emptyset, \text{receipt}\rangle \rightarrow \{\text{perm(leavestage(donor))}\} \quad (16)$$

$$\{\textit{active}(\textit{interdiction})\}, \text{violation}\rangle \rightarrow \{\text{perm(enterstage(dispatcher))}\} \quad (17)$$

$$\{\textit{active}(\textit{absentation}), \textit{active}(\textit{violation})\}, \text{return}\rangle \rightarrow \{\text{perm(hit(donor, villain))}\} \quad (18)$$

$$\mathcal{C}^{\downarrow}(\mathcal{X}, \mathcal{E}) : \langle \emptyset, \text{interdiction}\rangle \rightarrow \{\text{perm(give(donor, villain, item))}\} \quad (19)$$

$$\langle \{\textit{active}(\textit{interdiction})\}, \text{abstention}\rangle \rightarrow \{\text{perm(leavestage(donor))}\} \quad (20)$$

$$\langle \{\textit{active}(\textit{interdiction})\}, \text{violation}\rangle \rightarrow \{\textit{active}(\textit{interdiction})\} \quad (21)$$

$$\langle \{\textit{active}(\textit{absentation}), \textit{active}(\textit{violation})\}, \text{return}\rangle \rightarrow \{\textit{active}(\textit{absentation})\} \quad (22)$$

4 VAD emotional model

In order to make the agents acting out the Punch and Judy show more believable, we apply an emotional model to affect their actions and decisions. For this, we use the valence-arousal (circumplex) model first described by Russell [10].

In order to give each character its own distinct personality, we extend this model with an extra dimension: dominance, as used by

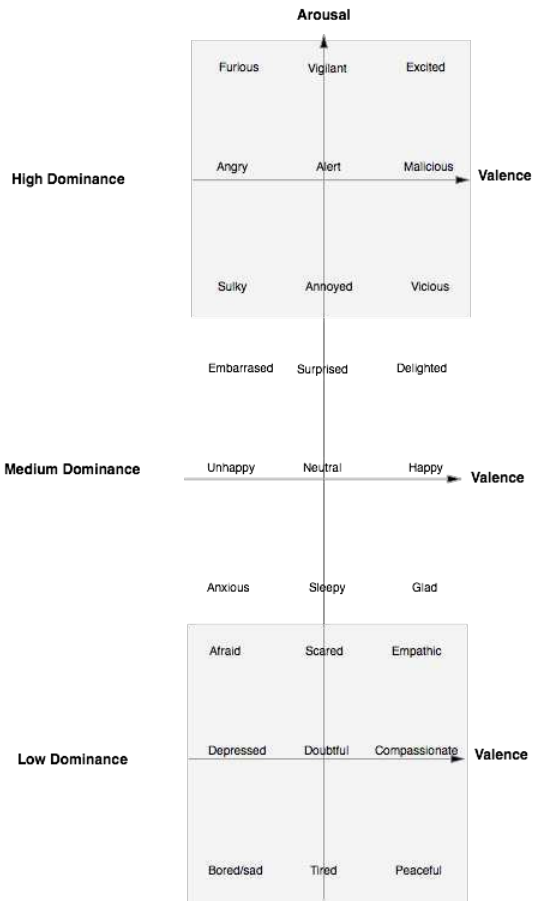


Figure 1. VAD emotional values, adapted from Ahn et al [1]

Ahn et al in their model for conversational virtual humans [1]. This dominance level is affected by the reactions of the audience to the agents' actions. For example, Judy may become more dominant as her suggestions to hit Punch with a stick are cheered on by the audience, emboldening her into acting out her impulses.

Figure 1 shows how valence, arousal and dominance values map to identifiable emotions. Valence, arousal and dominance can each have a value of low, medium or high. This allows the agents to have a total of 27 distinct emotional states.

Valence and arousal levels of each agent are affected by the actions of other agents. For example, a character being chased around the stage by Punch will see their valence level drop while their arousal increases. According to Russell's circumplex model of emotion [10], this would result in them becoming *afraid* (if their dominance level is low).

An agent's emotional state affects its ability to fulfil its institutional obligations. An agent that is *furious* would have no problem carrying out an obligation that requires them to kill another agent. If that same agent is *happy* or *depressed*, however, they might not have the appropriate motivation to perform such a violent action.

5 Architecture

5.1 Multi-Agent System

We use the JASON framework for belief-desire-intention (BDI) agents [2], programming our agents in the AgentSpeak language.

The VAD emotional model is represented inside each agent as a set of beliefs. Each agent has beliefs for its *valence*, *arousal* and *dominance* levels, each of which can take the value of low, medium or high. This combination of VAD values creates one of the 27 emotional states shown in figure 1, affecting whether or not an agent breaks from its permitted or obliged behaviour.

5.2 Institutional Framework

To describe our institutional model, we use instAL [3], a DSL for describing institutions that compiles to AnsProlog, a declarative programming language for Answer Set Programming (ASP). instAL's semantics are based upon the Situation Calculus [9] and the Event Calculus [5]. It is used to describe how external events generate institutional events, which then can initiate or terminate fluents that hold at certain instances in time. These fluents can include the permissions and obligations that describe what an agent is permitted or obligated to do at specific points in time.

For example, if an agent with the role of *dispatcher* leaves the stage, it generates the *absentation* Propp move in the institution:

```
1 leaveStage(X) generates intAbsentation(X) if
      role(X, dispatcher), activeFunction(
      interdiction);
```

The *absentation* institutional event gives the crocodile permission to enter the stage if there are any sausages on the stage. It also terminates the permission of the absented agent to leave the stage, as they have already done so:

```
1 intAbsentation(X) initiates perm(enterStage(
      croc)) if objStage(sausages);
2 intAbsentation(X) terminates onStage(X), perm(
      leaveStage(X));
```

instAL rules like those shown above are compiled into AnsProlog ASP rules. Once the instAL model is compiled to AnsProlog, we use the *clingo* answer set solver [4] to ground the logical variables, and 'solve' queries by finding all permissions and obligations that apply to any agents, given a sequence of events as the query input. The agents' percepts are then updated with their permitted and obliged actions from that moment in time onwards.

5.3 Bath Sensor Framework

The Bath Sensor Framework (BSF) [6] is a framework supporting publish/subscribe-style communication between distributed software components, in this case connecting intelligent agents with their virtual environments. It uses the XMPP publish/subscribe protocol to allow the communication between agents and their environments. Each agent subscribes to receive notifications of environment changes via XMPP server, which relays messages between publishers and subscribers. If any environment change occurs, all subscribed agents are informed of the changes.

This allows agents' environments to be created using entirely different technologies and programming languages from the agents themselves. In our case, BSF is especially useful as the animation engine that acts as the agents' environment is written in Javascript and runs in the browser. This means that the clingo solver and JASON agent

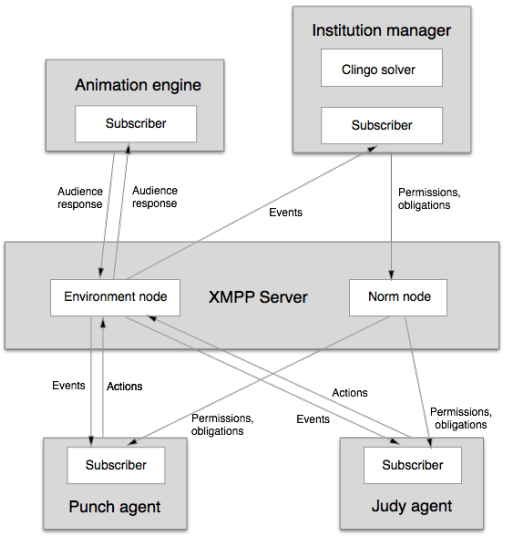


Figure 2. System architecture

framework can run on a central web server and communicate to any connected clients using BSF and XMPP.

Figure 2 shows how BSF is used to coordinate the components of the system. An XMPP server runs two publish/subscribe nodes. One node is for events related to changes in the environment (the *environment* node), the other is for changes in agents' permissions and obligations (the *norm* node).

All agents (in this case, Punch, Judy, the Policeman, etc) are subscribed to both the environment and norm nodes. They can also publish events to the environment node, but not the norm node. Only the institution manager (connected to the *clingo* solver) can publish permissions and obligations to the norm node. This manager (labelled in figure 2 as *institution manager*) is subscribed to the environment node of the XMPP server, watching it for events. These events then get passed to the *clingo* solver with the institutional model, which outputs the new permissions and obligations, publishing them to the norm node.

The animation engine is subscribed to the environment node, watching it for any events that need animating for the puppet show. In addition, it can publish input from the audience ('cheers' or 'boos') as events to the same node.

5.4 Animation

The animation engine that shows the visual output of the agents actions is written in Javascript and the Phaser game framework. It runs entirely in a browser, and communicates with BSF using the Strophe XMPP library.

If the user allows the program access to their microphone, they can cheer or boo the actions of the agents by shouting into the microphone. Otherwise, they can simulate these actions by clicking on 'cheer' or 'boo' buttons at the bottom of the screen.

6 Audience Interaction

The puppet show is designed to be run in front of either a single user's computer, or on a large display in front of an audience. The

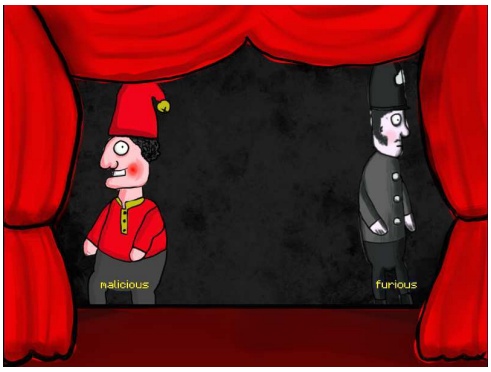


Figure 3. A screenshot of the Punch and Judy show

user/audience is instructed to cheer or boo the actions of the characters of the show, which will be picked up by a microphone and 'heard' by the agents. This will then affect the emotional state of the agents and change the actions they make in the show. Their actions are constrained by the set of 'Punch and Judy' world norms as described in the institutional model.

There are many different ways in which the audience's responses can affect the outcomes of the show. If the audience craves a more 'traditional' Punch and Judy experience, then they can cheer Punch into beating and killing all of his adversaries (including his wife, Judy). Alternatively, a more mischievous audience could goad Judy into killing Punch and then taking over his role as sadist and killer for the rest of the show. The narrative outcomes are dependent on how the audience responds to the action, yet still conform to the rules of the Punch and Judy story world.

7 Conclusion

With our approach to interactive narrative generation, we regulate the rules of the story domain using an institutional model. This model describes what each agent is permitted and obligated to do at any point in the story. This approach alone would be too rigid, however. Though the audience's interactions (cheering or booing) may alter the course of the narrative, the agents would still have to blindly follow a pre-determined set of paths. By giving our agents emotional models that change their willingness to follow the narrative, a degree of unpredictability is added to each run-through of the show, giving the impression that the agents are indeed characters capable of free will.

REFERENCES

[1] Junghyun Ahn, Stéphane Gobron, David Garcia, Quentin Silvestre, Daniel Thalmann, and Ronan Boulic, 'An NVC emotional model for conversational virtual humans in a 3d chatting environment', in *Articulated Motion and Deformable Objects*, 47–57, Springer, (2012).

[2] Rafael H Bordini, Jomi Fred Hübner, and Michael Wooldridge, *Programming multi-agent systems in AgentSpeak using Jason*, volume 8, John Wiley & Sons, 2007.

[3] Owen Cliffe, Marina De Vos, and Julian Padget, 'Specifying and reasoning about multiple institutions', in *Coordination, Organizations, Institutions, and Norms in Agent Systems II*, 67–85, Springer, (2007).

[4] Martin Gebser, Benjamin Kaufmann, Roland Kaminski, Max Ostrowski, Torsten Schaub, and Marius Schneider, 'Potassco: The potsdam answer set solving collection', *Ai Communications*, **24**(2), 107–124, (2011).

[5] Robert Kowalski and Marek Sergot, 'A logic-based calculus of events', in *Foundations of knowledge base management*, 23–55, Springer, (1989).

[6] Jeehang Lee, Vincent Baines, and Julian Padget, 'Decoupling cognitive agents and virtual environments', in *Cognitive Agents for Virtual Environments*, eds., Frank Dignum, Cyril Brom, Koen Hindriks, Martin Beer, and Deborah Richards, volume 7764 of *Lecture Notes in Computer Science*, 17–36, Springer Berlin Heidelberg, (2013).
[7] Pablo Noriega, *Agent mediated auctions: the fishmarket metaphor*, Citeseer, 1999.
[8] Vladimir Propp, 'Morphology of the folktale. 1928', *Trans. Svatava Pirkova-Jakobson. 2nd ed. Austin: U of Texas P*, (1968).
[9] Raymond Reiter, 'The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression', *Artificial intelligence and mathematical theory of computation: papers in honor of John McCarthy*, **27**, 359–380, (1991).
[10] James A Russell, 'A circumplex model of affect.', *Journal of personality and social psychology*, **39**(6), 1161, (1980).

Search and Recall for RTS Tactical Scenarios

Jason Traish, James Tulip and Wayne Moore [1]

Abstract. The success of a Real-Time Strategy agent is heavily dependent on its ability to respond well to a large number of diverse tactical situations. We present a novel method of tactical decision making called Search and Recall (S&R) which is a hybrid of Search and Case Based Reasoning (CBR) methods. S&R allows an agent to learn and retain strategies discovered over the agent's history of play, and to adapt quickly in novel circumstances.

The sense of memory that S&R provides an RTS AI agent allows it to improve its performance over time as better responses are discovered. S&R demonstrates an minimum win rate of 92% in standard scenarios evaluated in this paper.

S&R decouples search from the main game loop which allows arbitrary computational complexity and execution time for search simulations. Meanwhile in-game decision making is based on CBR and remains fast and simple.

This paper presents an S&R model which extends the ability of an RTS AI agent to deal with complex tactical situations. These situations include special unit abilities, fog of war, path finding, collision detection and terrain analysis.

1 Introduction

Real-time strategy (RTS) games are a popular genre of commercial games that require substantial practice, skill and experience to master. In order to conquer an opponent, a player must manage a number of in-game systems with precision, using a large number of possible commands. In-game systems include research, economics, exploration, managing an army, and executing a strategy with the potential of defeating the opponent's strategy. On top of all this complexity a player is expected to complete all these tasks in a real-time environment of uncertainty.

The number of in-game systems and possible commands illustrate the complexity of the RTS genre and form the basis of it's appeal to players and researchers alike. Developing an RTS agent poses many challenges that are not present in traditional strategy board game environments such as GO and Chess. In particular, the large number of units and possible commands, the uncertain environment, the effect of terrain, and the real-time execution constraints are unique to the computer based RTS genre.

It is very difficult to write scripted agents that vary their responses in different situations. This results in easily exploited AI agents which fail to give experienced players an enjoyable challenge.

Case Based Reasoning is one approach that has been used to create adaptive RTS AI agents [1, 2, 8, 9]. Search based methods have also became a point of interest to the RTS research community [3, 4, 12]. However, both approaches have intrinsic limitations.

[1] Charles Sturt University, Mining Lab, Australia, email: {jtraish & jtulip} @csu.edu.au & wmoore@lisp.com.au

1.1 Case Based Reasoning and Search in RTS AI

Case based reasoning (CBR) methods have been used successfully to create adaptive RTS agents. In general, such methods store plans with an associated game state and use this data to reason about future encounters.

Aha et al. [1] demonstrated a CBR agent capable of identifying and adapting to a randomly selected opponent which demonstrated good results. Their agent relied on the availability of a set of pre-generated responses, each capable of winning against an opponent from a given position.

McGinty et al. [8] improved CBR approaches by changing the structuring and case retrieval approach, leading to significantly better results. Their agent demonstrated a high win rate in experiments with imperfect information. Other CBR methods have focused on the use of recorded human player interactions to make decisions [2, 9].

However, while CBR has been successful in creating adaptive RTS agents, they face a number of challenges. Responses derived from human players can be of inconsistent quality due to the diversity of human player skills and the nature of human play. Standard CBR approaches are also ill equipped to make decisions if there is no similar recorded context.

As a result, search based methods, and in particular Monte Carlo simulations have gained the interest of the RTS AI research community [3, 4, 5, 6, 12]. Search based methods enable an agent to adapt in real-time to whatever circumstances it is currently facing, assuming the simulator can correctly predict the outcome of a given response action. Significant research on adaptive agents using search based techniques has been performed in the context Chess and GO [7] and the application of such techniques to RTS games is an attractive prospect.

However, complexities such as path finding and collision detection are required for an agent to appropriately handle commercial game type tactical situations. Such situations include moving units in a environment affected by terrain, or engaging armies of many varied unit types, some with special abilities.

The complexity inherent in commercial RTS games places huge computational demands on the simulations required to perform a search for a tactical solution. For this reason most of the published search simulation approaches are very simple relative to the demands of fully realised commercial game agents and ignore issues such as terrain, path finding, and collisions between units.

The problem is that simulations conducted within the game loop are heavily constrained to execute in an extremely limited amount of time, due to the demands of other aspects of the game loop such as animation and rendering.

In the rest of this paper we present a hybrid search/CBR approach called Search and Recall (S&R) which enables simulations capable of dealing with commercial grade RTS game complexity, while offering CBR level in-game performance. We demonstrate these capabil-

ities in the context of Starcraft Broodwar; a commercial RTS which has become a popular RTS AI research platform.

The main contribution of this work is to demonstrate the utility of responses generated using Search simulations as recorded responses in a CBR-like database. The technique was inspired by case base reasoning literature that focused on constructing databases using player responses [10]. We also demonstrate an approach for making computationally intensive search simulations feasible in the context of a real-time game.

2 Search and Recall - Overview

Search and Recall (S&R) is a novel method of tactical decision making which is a hybrid of Search and Case Based Reasoning (CBR) methods. It allows an agent to learn and retain strategies discovered over the agent's history of play, and to adapt quickly in novel circumstances.

Similarly to CBR methods, S&R uses a database of previously discovered successful responses associated with a collection of identified game states. S&R agents use these responses to quickly identify a solution without extensive simulation within the game loop. However, unlike other CBR methods, S&R does not populate it's response database with a static set of game states identified from previously played games. Rather, it populates the database dynamically with the results of search simulations conducted in response to actual game states encountered during play.

By combining the adaptive learning of MCS with the memory of CBR, S&R allows an agent to improve the quality of its responses over the course of multiple games.

In essence, we decouple the search tasks from the game loop by allowing them to execute asynchronously and in parallel with the game loop. Searches are pushed into concurrent threads, allowing them to take as long as necessary without delaying game rendering. The agent makes its decisions based on its current database of solutions, and the search tasks update that database asynchronously with the results of new simulations based on possible responses to the current game state. As many searches can be carried out as are appropriate to the CPU resources available to the game.

Search time is limited only by the length of a game or an arbitrary stopping condition, and is substantially longer than the 5ms generally allocated for an agent's decision making process within the standard game loop. The downside is that the longer it takes to evaluate potential decisions the more likely it is that the response will come too late to be useful in the current situation. However, the next time a similar situation is encountered, the simulation results will be available in the CBR database (response library) ready for near instant access.

Search results are used to update a CBR like database as they become available, and the AI task within the game loop is reduced to selecting the appropriate response as in a conventional CBR system.

We apply this architecture in the context of the commercial game Starcraft Broodwar. Starcraft is an immensely popular and sophisticated RTS game, famous for its balanced asymmetric game play and status as a professional spectator sport in Korea. Starcraft Broodwar is a version of Starcraft for which an external programming interface has been developed called the Brood War API (BWAPI). The availability of BWAPI has made Broodwar an attractive platform for RTS AI research.

3 Search and Recall - Agent Components

The S&R agent is composed of a recall-playback component (RPC), a search component (SC), and a response library (RL). This basic architecture is illustrated in Figure 1. The RPC component acts as coordinator for the agent and interacts with the BWAPI interface. As soon as a Broodwar game begins the S&R agent starts the recall-playback component and initialises the search component with a number of threads.

3.1 Recall/Playback Component (RPC)

The RPC matches the current game state against the game states currently recorded in the response library. Game states in the database are identified by a simplified descriptor containing only the number and types of unit present.

The RPC then retrieves the response associated with the current game state from the response library. The response associated with a game state is always the most favourable response generated by the search simulations carried out in the search component. If no matching game state is found, the RPC assigns random behaviours to the agent's units. If a response was loaded earlier from a previous game state then those previous behaviours are not changed.

The RPC has a simulator similar to those being used for searching. It uses this to simulate a single time step using the unit actions specified in the response. This step is carried out in order to map from the actions specified in the response to a set of Broodwar commands that must be issued through the BWAPI interface. The raw actions that the units must perform are recorded (e.g. move[x,y], attack[unitId]) and forwarded to the BWAPI.

Although games states are identified in the RL only by the number and type of units present, actual game state is defined with considerably more information on unit positions, current unit states, what projectiles have been created, which units are damaged, and which weapons have entered their cool down periods. All of this information is captured from the BWAPI and sent through to the search component (SC) in addition to the number and type of units present in the scenario. The RPC buffers these changes in actual game state for the SC, updating the information used by that component as a basis for simulation only after 200 simulations have completed. This allows a sufficient number of searches associated with a particular game state to complete to be useful in subsequent games.

The execution of the RPC is constrained to take less than 5ms per frame since it executes as a part of the main game loop. This constraint is easily achieved since the simulator used to calculate the BWAPI commands simulates only a single time step.

3.2 Search Component (SC)

The search component is represented in Figure 1 as the Concurrent Search Simulators (CSS). It consists of a number of search threads which repeatedly run simulations for the combat scenario utilizing the current actual game state, a simulator engine, and a set of actions assigned to each unit in the scenario.

At the beginning of a simulation, each search thread is given the current identifying game state (unit numbers and types) as well as information describing the actual game state (terrain, unit positions, unit health, current unit action states, etc).

We randomly assign behaviours to each unit for each simulation so we can evaluate the effect of utilising different tactics on the outcome of a battle . If simulations complete quickly, many different

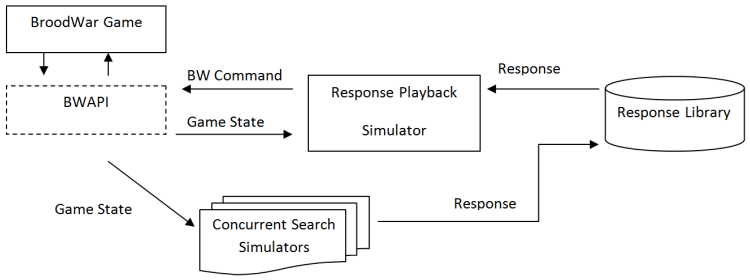


Figure 1. Search and recall agent process

possible outcomes can be calculated and used to update the solution available to the RPC before the time allotted to its execution within the game loop (5ms) expires. However, if it takes longer to simulate an outcome than the time Broodwar allows, then the result will not be available to the RPC during the current game loop. This results in the game agent taking longer to respond to a game state in real time, although simulation results do become available to the RPC over the next few game loop cycles as simulations complete.

When a simulation completes, the quality of the response is calculated as the total health percentage of the remaining allied units at the end of simulation. A quality of 0 is given for prediction in which all allied units are killed. This formula favours victories with lower casualties, and ranks all losses equally.

As in [6], the simulator is a mathematical model of the combat mechanics implemented in Broodwar, that allows simulations to be run without any frame rate derived speed limitations. As such, it is not an exact model of the combat mechanics implemented in Broodwar.

3.2.1 Simulators

Asynchronous execution of search allows the complexity and execution cost of the simulation engine used to be increased arbitrarily. In this work we explore the effect of increasing the complexity of the simulation engine used by evaluating the performance of two different simulators. These are:

1) Basic Simulator: This simulator handles unit health, shields, healing, attacking, and movement without collision or path finding. It can complete up to 2000 combat simulations per second per thread.
2) Complex Simulator: This simulator handles unit health, shields, healing, and attacking. However, the movement function detects collisions and finds paths around obstacles such as terrain and other units. Influence maps from [11] have also been integrated to support a 'kiting' behavior which has been added to the list of available behaviours. This simulator can complete only up to 200 combat simulations per thread per second.
   Kiting is a highly successful behaviour that fast moving ranged units can use against slower units. Kiting is the act of attacking an enemy unit and then moving away while reloading.

3.2.2 Response Divergence

A response grows stale the longer it is in effect. This is due to differences in mechanics between the Broodwar game and the simulator that even a very sophisticated model will find challenging to eliminate, in particular because there are random elements built into the Broodwar game engine. We call the differences between the simulated outcome and what actually happens in Broodwar as divergence. Divergence represents the cumulative error between the game states of the simulation and Broodwar as time passes.

Different game systems suffer differing amounts of divergence. While systems like health regeneration and attack damage are straight forward, other components such as attack cool downs are randomisied slightly, introducing small changes in combat outcomes. The precise mechanics of other systems such as path finding are unknown and this also increases the divergence of simulations from actual game encounters. Furthermore, an opponent model is not necessarily a precise model of the Broodwar AI, and this also leads to a large amount of divergence. Finally, the actual precise game state used to drive the search simulation that generated the response recorded in the database may differ from the precise current game state. If differences in precise unit location and health affect the outcome of the battle, divergence will occur.

3.2.3 Opponent Models

In order to combat the effects of divergence, solutions that generalise well are sought. The simulation outcome is heavily dependent on the strategy used by the opponent, so we attempt to find generalized solutions by taking the minimum of the solution quality score over a small set of opponent models. This favours the selection of robust strategies that are successful against a variety of opponent models for the response library (RL). In the current work this set of opponent models contains only 2 strategies; one using an 'Attack Weakest strategy, and the other using an 'Attack Closest' strategy.

3.2.4 Unit Behaviours and Grouping

A behaviour describes what action the unit should take in any given circumstance. A behaviour consists of a series of actions which a unit executes in sequence, moving on to the next action when the previous action is complete or appropriate conditions are met. For each behaviour we identify a primary action, and a secondary action which is applied if multiple targets are identified for the primary action. For example, if 'Attack Weakest' is the primary action, and all enemy units have the same health, then the secondary action 'Attack Closest', is applied. Behaviours are described in Table 1.

In order to allow the S&R agent some flexibility in terms of choosing and targeting particular units or types of enemy units, we provide the agent with the ability to separate the enemy into groups.

When setting up a simulation, not only are a random set of behaviours assigned to the agent's units, but the enemy is divided into 4 random groups. Actions are then made specific to groups. For example, the generic "Attack Closest" behaviour becomes "Attack Closest in Group 1". Grouping allows the agent to create plans that can focus fire individual or groups of units. This greatly increases the degree of freedom with which the agent can respond to situations.

3.3 Response Library Component (RLC)

The S&R agent receives its recall ability from the use of the response library. The response library is responsible for the storage and communication of the best recorded responses from the search simulations. The database is updated asynchronously by the SC, and queried from within the game loop by the RPC. It acts as a constantly growing and improving database of best seen responses to recorded tactical situations.

3.3.1 Game State and Response Descriptors

Preliminary testing identified that actual game state needed to be generalized for successful game state matching to occur. Furthermore, only a small number of game state attributes were required for the agent to adapt competently. Hence, the attributes used to identify game state within the RLC include only the number and type of each unit involved in the current scenario. Adding more detailed game state descriptors such as those describing unit health or position causes an explosion of possible states, this drastically shortens the time that a game remains in a particular state, and makes it difficult to match the current game state with a state recorded in the response library.

Describing game state by only the number and type of units involved results in relatively stable states that recur sufficiently frequently to make matching effective, and balances the frequency of response adaption. This approach effectively forces the chosen response to change only to when units are removed from or added to the game.

In addition to the game state information that is used as a key in the response library, each entry in the response database records the behaviour assigned to each unit, and the groupings assigned to the enemy units.

Response behaviours do not correspond with BWAPI commands: they need to be mapped into BWAPI commands by the simulator associated with the RPC.

4 Experimental Setup

The following experiments contain four tactical scenarios that an agent cannot resolve with a singular response. These are illustrated in Figure 2 and listed below:

A) 3 Zealots vs 3 Vultures (Attack Closest agent): This scenario pits 3 fast ranged units (Vultures) controlled by the agent against 3 slow close attack units (Zealots). This scenario favours the kiting strategy as it is extremely difficult to solve without it.
B) 6 Fast Zerglings vs 2 Dragoons (Attack Closest agent): This scenario pits 2 strong ranged units (Dragoons) controlled by the agent against 6 fast close attack units (Fast Zerglings). Once again a kiting solution is favoured, but far more precision is required to make this work.
C) 3 Zealots and 3 Dragoons vs 3 Zealots and 3 Dragoons (Default AI): This is a symmetrical scenario pitting ranged (Dragoons) and close attack (Zealots) units against each other. Precise control over unit attacks which enemy unit as well as unit placement is required to be successful.
D) 8 Dragoons vs 8 Dragoons (Default AI): Once again this is a symmetrical scenario that pits equal numbers of ranged units against each other. Control of attack strategy is important in this scenario, but unit placement is less important than in Scenario C.

The experimental setup is based on work by [5] although the experimental setups for scenarios A and B differ from Churchill's implementation. Due to problems encountered with the BroodWar AI's default behaviour it was replaced with a scripted agent designed to constantly attack the closest unit.

Each scenario is run against a particular configuration of the S&R agent for a total of 200 games at an acceleration of 5ms per frame. This is necessary since due to stochastic variation between games, the outcome of an actual game is not completely deterministic. The scores recorded in Table 2 are defined by the following function to the nearest percentage.

$$Score = (wins + draws/2)/200$$

Our experiment compares several different configurations of the SR agent. The performance of the basic and the complex simulator engine are compared in two modes: in pure search mode (ie without access to any stored responses), and in combined search and recall mode (with access to stored responses). This tests whether there is any advantage in retaining results from earlier simulations. For comparison purposes, the performance of two scripted agents was also evaluated: one based on an 'Attack Closest' strategy, and another which favours Kiting. Each configuration or agent is tested on the four scenarios listed above.

For the S&R agents, each configuration is initialised with a new empty response library at the beginning of the evaluations for all scenarios. All recorded responses are generated by simulations run during the actual games.

All S&R experiments utilise 4 threads within the SC for running simulations. Each search was limited to 2000 time steps although this number of steps was never reached. The results of the experiments are shown in Table 2.

5 Results and Discussion

The results of the experiments for the scripted agents show clearly that to do well in all four scenarios requires adaptive agent behaviour. The 'Attack Closest' scripted agent performs poorly in scenarios A and B, but is successful in scenarios C and D while the reverse is the case for the 'Kiting' scripted agent.

Results for the simple simulator, which does not have a kiting behaviour available are similar to the 'Attack Closest' scripted agent. This illustrates the importance of the simulator model containing a set of behaviours sufficient to cover what is required in a scenario.

On the other hand, results in scenarios A and B for the complex simulator show that agent clearly discovered and utilized the appropriate kiting behaviour. Results in Scenario A are stronger than in Scenario B, likely because the large speed difference between Vultures and Zerglings makes a wide range of successful kiting solutions relatively easy to find. In Scenario B, if the Dragoons performed a suboptimal action for even a small period they would lose to the larger numbers of Zerglings.

Table 1. Behaviour Descriptions

Behaviour	Primary Function	Secondary Function	Condition
G1, G2, G3 and G4	Attack unit of least health in group X	Attack closest unit in group X	No units in group X
Attack Closest	Attack closest unit	Attack unit of least health	N/A
Attack Wounded	Attack unit of least health	Attack closest unit	N/A
Kite	Attack unit of least health in range when ready to fire	Move away from all enemies and terrain	N/A

Table 2. Experiment 1 Results. S&R: Search and Recall. IM: Influence Map.

Setup	Churchill Search	Search	S&R	Search (IM)	S&R (IM)	Attack Closest	Kiter
A	0.81	0	0	0.96	1.00	0	1.00
B	0.65	0	0	0.65	0.92	0	1.00
C	0.95	0.95	0.80	0.76	0.94	0.77	0.26
D	0.96	1.00	1.00	1.00	1.00	0.97	0.14

The results for the complex simulator with recall enabled are better than for search alone, indicating that the recall capability provides a considerable advantage. The advantage conferred by the recall ability is much greater in Scenario B than in Scenario A. This suggests that the advantage of accumulating knowledge in the response database is greatest when solutions are relatively exact, and the exploration of the solution space is relatively slow.

Results for the simple simulator are equivalent or better than the complex simulator for scenarios C and D. This indicates that the range of behaviours available to the simple simulator are sufficient in these scenarios, and that the complexities introduced for the complex simulator have little impact in these scenarios. This result is not terribly surprising since the influence map affects only the kiting behaviour which is not necessary in these scenarios, and the close ranged combat and lack of terrain features in these scenarios reduces the impact of the path finding capability of the complex simulator. Given these considerations, it may be that the much greater number of simulations that the simple simulator can perform (2000 vs 200 per second) allows it to find better solutions than the complex simulator.

Results for scenario C yield are the most varied. The winning solutions for this scenario required more complex behaviours than in the other scenarios. Scenario C is similar in some respects to Scenario B with its rigorous success requirements.

Results for the complex simulator in Scenario C show a large difference between search only and combined search and recall. Once again it appears that that the recall capability becomes a significant advantage when solutions are hard to find and the exploration of solution space is slow.

Results degrade when recall is enabled for the simple simulator. It is likely that this is an example of the effects of divergence. The simulator has discovered an action set that is effective in simulation, but that does not translate well into the actual game. This indicates the importance of the simulator's combat model being a close match to the actual game's.

Results for Scenario D are both extremely strong and uniform across both the simple and complex simulators, both with and without recall enabled. This is probably a result of the scenario being relatively easy to solve, as indicated by the strong result also generated by the 'Attack Closest' scripted agent.

Over all scenarios, the strongest performance is shown by the complex simulator with recall enabled. This configuration of the S&R agent adapts strongly to all scenarios, even though its performance without recall enabled is relatively weak. The result is important, since it indicates that the build up of experience over many game cycles becomes greatly beneficial when solutions are hard to find, and simulation rates are slow. This is exactly the situation faced when attempting to apply accurate simulation models to complex commercial grade RTS AI problems.

Note that for all the search based configurations, results between zero and one are in some ways a measure of divergence, since the simulations return what they estimate as a winning solution or a loss. Solutions that win sometimes reflect differences between what the simulators calculate and what actually happens in Broodwar. This tends to impact weaker solutions to a greater extent, resulting in lower scores where search is less effective. Given this interpretation of each scenario score, it is an important result that the scores for the complex simulator with recall enabled are consistently high across all scenarios. This reflects relatively little divergence between what the complex simulator predicts and what happens in Broodwar, given a sufficient accumulation of simulations, and the capacity to retain the results.

Another important result is that the benefits of recall are delivered to the agent relatively quickly. There is a marked improvement for the complex simulator with recall enabled in the difficult scenarios even though the scenario is evolving in real time. This indicates that the advantage of receiving high quality solutions outweighs the disadvantage of them taking more than a game cycle to calculate.

In comparison with Churchill's results, the complex simulator with recall enabled dominates by a large margin in all but Scenario C, where it is only marginally weaker. Given the divergence interpretation of the evaluation scores, the results suggest that the complex simulator is a much closer approximation of the Broodwar combat mechanics, and that the predictions made by the complex simulator are much more accurate. The 'complex simulator with recall' approach is an approach worth pursuing.

6 Conclusions and Future Work

Overall the results of this preliminary study can be summed up as: high quality responses are worth remembering, when solutions are hard to find, the exploration rate of the solution space is low, and when the fidelity of the simulations is high.

The results strongly indicate that retention of results from search simulations is worthwhile, and that Search and Recall is a useful approach. This eliminates the need for a huge and uneven quality

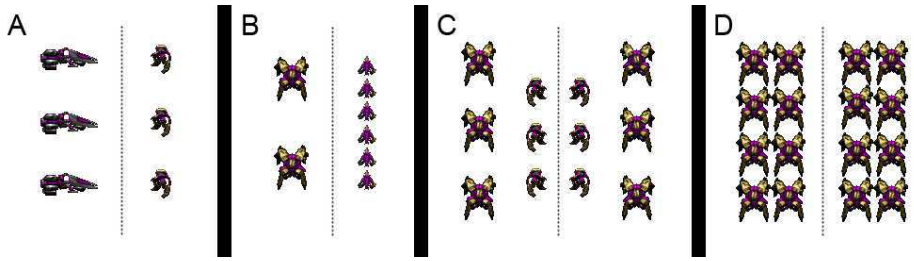


Figure 2. Experimental Setup

database of pre-played games on which to base CBR, and allows the situations a game AI can respond intelligently to grow over time. At the same time it guarantees fast decision making within the game loop.

An important implication of the proposed architecture is that because simulations are decoupled from the game loop, they become amenable to parallel, distributed, or offline processing. The exact actual game states sent to the SC, could instead be sent out over the network, or logged for later processing. Regardless of whether results arrive in time to advantage the S&R agent in the current game, the fact that the results are generated improves the response database over time, even when the game is not being played. Another implication is that simulation results from many separate instances of a game can be shared between games, allowing games to cooperate in improving the AI for all games.

A final implication is that simulations are not restricted to the CPU capacity of an ordinary gaming PC. Simulations could be conducted on server farms or supercomputers in the cloud, and the results used to update a global database available to all instances of a game.

Because the constraints on execution times and hence simulations complexity have been eased, future work could extend simulation models to scenarios of greater complexity such as working with terrain and larger unit encounters. It would also be interesting to explore the feasibility and utility of more detailed game state descriptors, and the associated much larger response databases required.

Once response databases become larger and more populated, game progression paths through state space and discovering general patterns of game progression could prove interesting. The sensitivity of results to the range of available behaviours also indicates that further work into more complex behaviour sets is also warranted.

S&R removes computational execution time restrictions on search but retains the ability of search based agents to adapt to new situations. The S&R agent model allows simulators used in searches to use much more complex models to deal with complex tactical situations. Simulators can include path finding, unit and terrain collision avoidance, and specialized behaviours. These complex simulators greatly improve the fidelity of the results produced, which reduces the divergence between predicted outcomes and those produced by the game. This makes the S&R method potentially useful in applying search techniques to commercial grade levels of combat scenario complexity.

REFERENCES

[1] David W. Aha, Matthew Molineaux, and Marc Ponsen, 'Learning to win: Case-based plan selection in a real-time strategy game', in *Case-Based Reasoning Research and Development*, volume 3620 of *Lecture Notes in Computer Science*, 5–20, Springer Berlin / Heidelberg, (2005).

[2] Klaus-Dieter Althoff, Ralph Bergmann, Mirjam Minor, Alexandre Hanft, Neha Sugandh, Santiago Ontan, and Ashwin Ram, 'Real-time plan adaptation for case-based planning in real-time strategy games', in *Advances in Case-Based Reasoning*, volume 5239 of *Lecture Notes in Computer Science*, 533–547, Springer Berlin / Heidelberg, (2008).

[3] Radha-Krishna Balla and Alan Fern, 'Uct for tactical assault planning in real-time strategy games', pp. 40–45. Morgan Kaufmann Publishers Inc., (2009).

[4] Michael Chung, Michael Buro, and Jonathan Schaeffer, 'Monte carlo planning in rts games', in *IEEE Symposium on Computational Intelligence And Games (Cig)*, (2005).

[5] David Churchill and Michael Buro, 'Incorporating search algorithms into rts game agents', in *Eighth Artificial Intelligence and Interactive Digital Entertainment Conference*, (2012).

[6] David Churchill, Abdallah Saffidine, and Michael Buro, 'Fast heuristic search for rts game combat scenarios', *AIIDE*, (2012).

[7] Leandro Soriano Marcolino and Hitoshi Matsubara, 'Multi-agent monte carlo go', pp. 21–28. International Foundation for Autonomous Agents and Multiagent Systems, (2011).

[8] Lorraine McGinty, David Wilson, Ben Weber, and Michael Mateas, 'Conceptual neighborhoods for retrieval in case-based reasoning', in *Case-Based Reasoning Research and Development*, volume 5650 of *Lecture Notes in Computer Science*, 343–357, Springer Berlin / Heidelberg, (2009).

[9] Manish Mehta and Ashwin Ram, 'Runtime behavior adaptation for real-time interactive games', *IEEE Transactions on Computational Intelligence and Ai in Games*, **1**(3), 187–199, (2009).

[10] Santiago Ontan, Kinshuk Mishra, Neha Sugandh, and Ashwin Ram. Case-based planning and execution for real-time strategy games, 2007.

[11] Alberto Uriarte and Santiago Ontan, 'Kiting in rts games using influence maps'. AIIDE, (2012).

[12] Wang Zhe, Kien Quang Nguyen, Ruck Thawonmas, and Frank Rinaldo, 'Using monte-carlo planning for micro-management in starcraft', in *GAMEON Asia*, pp. 33–35, Japan, (2012).

Follow-up on Automatic Story Clustering for Interactive Narrative Authoring

Michal Bída, Martin Černý and Cyril Brom[1]

Abstract. One of the challenges in designing storytelling systems is the evaluation of resulting narratives. As the story space is usually extremely large even for very short stories, it is often unfeasible to evaluate every story generated in the system by hand. To help the system designers to maintain control over the generated stories a general method for semi-automatic evaluation of narrative systems based on clustering of similar stories has been proposed. In this paper we report on further progress in this endeavor. We added new distance metrics and evaluated them on the same domain with additional data. We have also successfully applied the method to a very different domain. Further, we made first steps towards automatic story space exploration with a random user.

1 INTRODUCTION

Developing interactive storytelling (IS) systems is a challenging task involving multi-disciplinary knowledge, yet a number of IS systems was developed in the past, such as Façade [1], ORIENT [2] or FearNot! [3]. Bída et al. [4] notes that the evaluation of complex IS systems is a demanding process often requiring extensive effort. To mitigate this, the authors propose a computer assisted method of story evaluation based on clustering the stories into clusters according to their similarity. The general idea is that by meaningful clustering of the stories into groups the human designer will not be required to evaluate all the stories, but only few from each cluster and thus save development time. Authors also reported on the performance of the method on two domains - SimDate3D (SD) Level One and SD Level Two [5]. The first results indicated that the main metric could scale better than the other metrics on the complex domain of SD Level Two.

In this paper we report on further progress in a similar endeavor. Firstly, we have added two new features for the clustering algorithm in the SD domain - a) automatic extraction of sub-scenes from the recorded story and b) condensed tension difference curve based on the sub-scenes. We have managed to reproduce previous results on an extended domain of SD Level Two getting good performance using some of the new features. Secondly, we have implemented a random user that tries to explore the story space of SD Level Two by playing differently than an input set of previous stories hence exploring parts of story space not seen in the input set of stories. We show the performance of the metrics in distinguishing between stories generated by the random user and the original set of stories. Thirdly, we have applied the method on stories generated by the MOSS system [6] in order to investigate the performance of the method on a different domain.

Aside from the work mentioned, little has been done on story clustering. Weyhrauch [7] implemented several evaluation functions specific for his emergent narrative system. Ontañón and Zhu [8] proposed an analogy-based story generation system, where they evaluated the quality of resulting stories by measuring their similarity to "source" stories (input human-made stories). Compared to the approach in this paper, they were solving a problem of generation of the stories rather than the analysis of the stories.

This paper is organized as follows: First, we will describe the story domains we used in the experiments, then we will discuss updates of the method for narrative analysis and afterwards we present results of the new experiments. We will conclude the paper with discussion and future work.

Figure 1. SimDate3D Level Two screenshot showing Thomas and Nataly in the park with emoticons above their heads having a conversation about music.

2 DOMAINS

The experiments detailed in this paper have been conducted on IS system SD Level Two detailed in [5] and MOSS system [6].

SD game (Figure 1) is a 3D dating game taking place in a virtual city, with three protagonists: Thomas, Barbara and Nataly. The characters communicate through comic-like bubbles with emoticons indicating the general topic of the conversation

[1] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.
Email: {michal.bida,cerny.m}@gmail.com, brom@ksvi.mff.cuni.cz

(see Figure 1). The user partially controls one of the characters actions (typically Thomas). The users' goal is to gain the highest score by achieving certain kind of things, e.g. Thomas kissing one of the girls. The game features four possible endings.

The MOSS system [6] developed by M. Sarlej generates short stories with morals (e.g. greed, retribution, etc) in three domains (animals, family and fairytale). Each moral has its own emotional pattern that is used to generate stories with moral of a particular category. Internally the system uses Prolog abstraction to generate the stories, which is then translated to human readable text with Perl scripts. We worked directly with the internal prolog representation of the stories, which we parsed and analyzed with the system.

3 METHOD

Here, we will briefly overview the method we evaluated (which is given in detail in [4]). The main idea is to cluster the resulting narratives of a given IS system into groups of similar stories. The human designer then needs to see only several stories from each group to gain sufficient understanding of all the stories the cluster contains, saving development time. The clustering is done with the k-means algorithm. In the previous work, the clustering was based on two general features of stories: a) story action sequence and b) story tension (dramatic) curve.

The story action sequence is created by taking the sequence of actions done by all the characters in the story. Each of the actions available in the domain is assigned a letter and the sequence of these letters forms the *action string*. This way, standard string distance metrics (*Levenshtein*, *Jaro-Winkler* and *Jaccard* distances) are applicable to measure similarity between *action strings* representing different stories. In previous work, Jaccard distance has been shown to be of little use for story clustering in SD domains and is therefore tested here only for MOSS stories.

The *tension curve* is extracted from emotions experienced by the story protagonists. In SD this is straightforward as the characters are equipped with emotion model. The tension in SD is computed as follows: Every 250 ms we make a snapshot of all characters' emotions. Then we take the sum of these emotions where every positive emotion is counted with a minus sign and every negative emotion is counted with a plus sign. The resulting number encodes the tension value at the moment. The *tension curve* is then simply the piecewise linear function defined by these values.

In the MOSS system the emotions are also defined explicitly as a part of the generated stories. We again take the sum of positive and negative emotions at each time point of the story and the resulting value is the tension value at the specific time point of the story.

We propose two new features for clustering the SD stories: *sub-scene sequence string* and *condensed tension difference curve*. A sub-scene is a time span in the story where a) the set of characters that are in the proximity of the main protagonist do not change and b) the location of the main protagonist does not change. Let us suppose that Thomas (the main protagonist) is with Barbara (character) at the restaurant (place) – this is one sub-scene. After 5 minutes, Nataly arrives and joins them. At this moment, the old sub-scene ends and a new one begins. The new sub-scene features Thomas, Barbara and Nataly at the restaurant. Sub-scenes are extracted automatically from the story logs. The time span of sub-scenes varies from 5 seconds (enforced lower limit) to the whole duration of the story.

To measure distance between sub-scene sequences we assign strings to sub-scenes in the following way: one letter represents a location of the story (e.g. P for park) and the consecutive letters represent characters in the sub-scene (e.g. T for Thomas; one letter per each character present). For example, the "TBR" string represents a sub-scene where Thomas is with Barbara at the restaurant. The sub-scene sequence string is simply a concatenation of the individual strings. We then apply string distance algorithms as is the case with action strings.

Condensed tension difference curve is extracted from sub-scenes. We look at the tension value at the beginning and at the end of the sub-scene. The difference between these two values represents the tension difference for respective sub-scene. The condensed tension difference curve is defined as a sequence of all of these differences.

We have not implemented sub-scenes for MOSS stories, because the MOSS stories are already relatively short and composed of at most two sub-scenes. To check whether the clustering really captures non-trivial properties of the stories, we also tested difference in story length as distance metric for the MOSS domain.

All pairwise distances between stories have been computed, normalized and standardized prior to clustering.

3.1 Story space exploration with a random user

IS systems are often interactive, requiring a human user in the loop. Exploring the story space of such systems may be problematic as one needs many users and many story runs to get a reasonable coverage of the story space. For semi-automatic analysis the designer would benefit from an algorithm that would be able to explore parts of the story space automatically. For SD we have implemented a random user that is able to play the game alone. In addition, the random user tries to steer away from a given set of stories. Hence exploring parts of story space not covered in the given set of stories revealing previously unseen parts of the story space to the designer. This is achieved as follows: The random user (controlling Thomas) extracts the sub-scene sequences from the given set of stories and then tries to achieve a different sub-scene sequence in the story he is playing in. E.g., if the random users detects that most of the given stories started with characters at the restaurant, he will try to change location in the story by inviting the characters for example to the cinema and so forth for the second and the n-th sub-scene in the sequence. The random user has simple domain-specific knowledge that limits the actions he considers only to those contextually appropriate (e.g. he does no try to become intimate with a girl at the restaurant).

3.2 Evaluating clustering quality

As there is no generally accepted method for evaluating the quality of a clustering independent of the application, we use ad hoc method suitable for our scenario. Intuitively, a clustering is good, if stories in the same cluster have many features in common. Let us have a feature function $f: S \rightarrow V$, where S is the set of all possible stories and V is a finite set representing possible values of a feature the designer might be interested in.

For a cluster $X \subset S$ we define *precision with respect to f* as the proportional size of its largest subset sharing the same value of the feature:

$$precision(X, f) = \frac{\max\{|M| : M \subset X, \forall m, n \in M : f(m) = f(n)\}}{|X|}$$

In other words, precision of 0.62 means 62% of stories in the cluster produce the same value for *f*. The precision of the whole clustering is simply the average of per-cluster precisions. A system that clusters stories can be considered useful, if it provides high precision across multiple domains and multiple features.

In the experiments, we tested three features: the ending of the story (Experiment 1), the type of user (random vs. human) that generated the story (Experiment 2) and the MOSS moral of the story (Experiment 3).

As k-means depends on random initialization we ran each analysis 100 times to get robust results. In further text, we always report the average precision of these 100 clustering runs. To provide a simple baseline to the measurements, we also tried assigning stories to clusters at random. Once again an average of 100 random assignments is measured.

To provide a more robust evaluation of the methodology, it would be best to measure precision with respect to similarity of stories as perceived by humans. This however poses multiple methodological issues. In our view, a biggest obstacle to human evaluation is finding a useful dataset. Since humans cannot effectively cluster more than a handful of stories, the dataset needs to be small, which is usually unsuitable for machine clustering as the algorithm can easily pickup artifacts in the data. We left this as a future work.

4 EXPERIMENT 1

In Experiment 1 we analyzed an extended dataset of 70 human play sessions of SD Level Two using additional features – sub-scenes sequence string distance and condensed tension difference curve based on sub-scenes. Precision is measured with respect to the ending of the story. A graph of the results is presented in Figure 2.

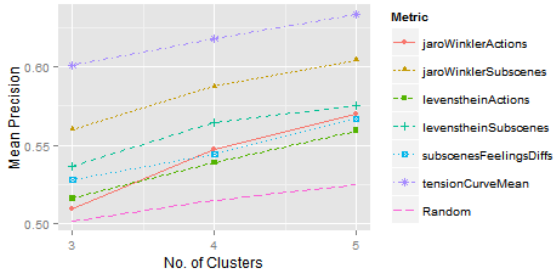


Figure 2. SD Level Two clustering results. Cluster precision weighted averages can be seen for three, four and five clusters (this is chosen arbitrarily based on that there are four possible endings). The results are averaged over 100 clustering runs with different initial cluster positions. The precision is calculated with respect to story ending.

As in previous work [4] we see that the tension curve outperforms other approaches in mean precision (0.6 for three clusters to 0.63 for five clusters). The interesting observation is that the sub-scene string sequence (metrics marked as "Subscenes" on Figures 2, 3, 4) outperform action strings (metrics marked as "Actions" on Figures 2, 3, 4) on this dataset. This indicates that sub-scene sequence is a meaningful feature in SD domain, relevant to story ending. Also note that Jaro-Winkler distance on sub-scenes (average 0.58) slightly outperforms Levenshtein (average 0.56). This is somewhat unexpected as Jaro-Winkler distance is usually a sub-par choice for clustering as it does not satisfy the triangle inequality. However this distance gives more weight to differences between first four characters of the string. The good performance of Jaro-Winkler on sub-scene sequences may then be explained by a large impact of the beginning of the story on its ending. Assigning higher weight to story start and/or story end might be an interesting extension of the approach as it would reflect the way stories are perceived by humans.

The compressed tension difference curve (metrics marked as "subscenesFeelingDiffs" on Figures 2, 3) scored on par with action strings distance metrics (average 0.55), but did not match the uncompressed original tension curve.

All metrics scored significantly better than the random cluster assignment. However compared to previous results [4] the addition of more stories resulted in lower precision for all previously measured metrics (tension curve and action strings). This might be partly caused by the larger size of the dataset, but it indicates that the metrics need to be made more robust.

Examples of the stories from this dataset and their clustering can be found in the appendix.

5 EXPERIMENT 2

In Experiment 2 we analyze a dataset containing 41 original human play sessions (as analyzed in [4]) and 66 randomly selected play sessions gathered from the random user. Precision is measured with respect to the type of the user that generated the story. A graph of the results is presented in Figure 3.

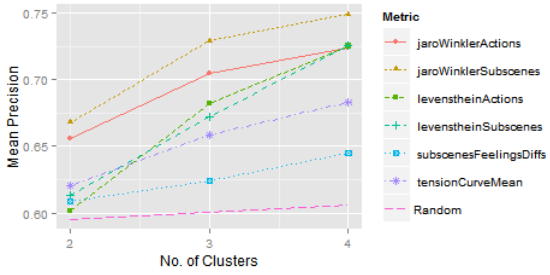


Figure 3. Experiment 2 clustering results. Figure shows the average precision of clustering with respect to the users that created the stories as a function of number of clusters.

The best metric for distinguishing between human and random user is Jaro-Winkler distance on sub-scenes (with precision 0.67 on two and 0.72 on four clusters). This can be explained again by the feature of the algorithm putting more weight on the first characters of the string. The random user tried to achieve different sub-scene sequence than the human users. Even though the story always begins the same (the first sub-scene is always the same), the random user immediately tried to change the sub-scene, so the second one differed from the average done by human users. This was picked up by Jaro-Winkler resulting in better performance of the algorithm.

The tension curve performed worse on this task (average 0.65). This is understandable as different sub-scene sequences in the story may produce similar tension curves. However this also indicates that the problem of similarity of the stories is multi-layered and to grasp this properly a combination of features is likely to be required.

6 EXPERIMENT 3

In Experiment 3, we ran the method on stories generated by the MOSS system. We have analyzed 3000 stories from fairytale domain of MOSS with recklessness, retribution and reward morals (1000 from each). Half of the stories comprised of two dramatic actions, and the other half comprised of four dramatic actions. In both cases, the resulting stories contained about 30 atomic actions. The precision was measured with respect to the moral of the story. A graph of the results is presented in Figure 4.

We can see that the precision of clustering is very high for almost all clustering metrics. For MOSS stories of length four, tension curve achieved precision of 0.99 on three clusters. The sum of normalized story length and Levenshtein on action strings was the second best scoring 0.93 on three clusters. On MOSS stories with length two, these two metrics performed a bit worse. The best was Levenshtein on action strings which averaged on 0.94 and the tension curve with 0.88 precision on average. The story length metric was outperformed by almost all other metrics and it also did not bring significant improvements to the Levenshtein distance indicating that the MOSS generating process did not produce artifacts in story length. Similarly to previous results on the SD domain, Jaccard distance did not perform well.

This overall good performance is caused by the fact that stories in MOSS are generated through templates that use emotional patterns. Stories in one domain exhibit the same or very similar emotional patterns resulting in similar tension curves. This is picked by the tension curve metric really well. The comparable performance of string metrics on action strings is likely caused by the presence of emotional actions in the action strings. The overall slightly worse performance on stories with dramatic length two is probably caused by the fact that less dramatic actions in the story offer less space to distinguish the stories from each other (however the performance was still remarkably good).

Examples of the stories from this dataset and their clustering can be found in the appendix.

7 CONCLUSIONS AND FUTURE WORK

We have presented new data for a methodology for semi-automatic evaluation of interactive storytelling systems based on clustering of similar stories. We have reproduced and refined previous results in the area.

New results showed that the method can be transferred successfully to other domain. However we need to take this with a grain of salt as the MOSS story generator abstraction was very favorable to the method as it uses emotional patterns to define categories of the stories.

Next, we have added new feature of stories, sub-scene sequence, that was used in the implementation of random user designed to explore unvisited parts of the story space of SimDate3D domain and we have shown the performance of the method on distinguishing random user from the human users. Some of the metrics scored worse than expected indicating that to grasp story similarity properly a combination of features will be required.

The semi-automatic exploration of the story space with a random user proved useful and will be further investigated in future work.

We have also shown the performance of the method on an extended dataset from SimDate3D Level Two. Although we

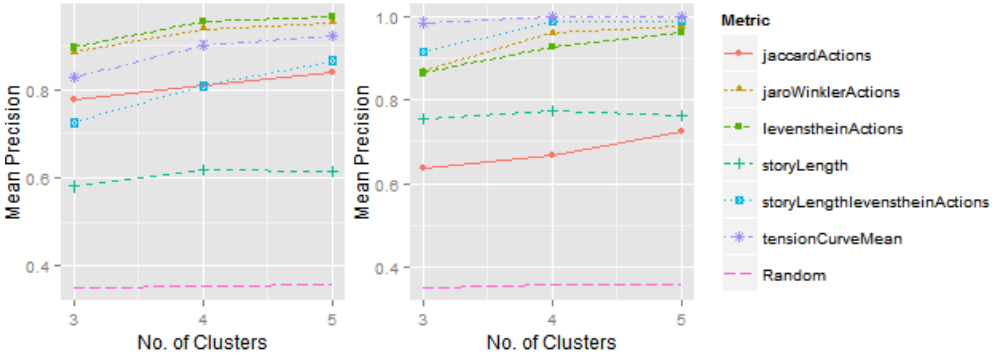


Figure 4. Experiment 3 MOSS domain clustering results. On the left there are precisions of clustering for three, four and five clusters when distinguishing between stories of the dramatic length two with particular moral. On the right there is the same for stories with the dramatic length four. All results were averaged over 100 clustering runs.

have reproduced the performance ordering of the metrics, the overall results were worse than in previous paper. The reason may be that the metrics do not accurately represent story similarity and pick a large amount of noise. A detailed analysis of stories in the same clusters could shed more light onto this and it is planned as future work, including comparison with story similarity as perceived by humans.

In line with conclusions from previous work, the tension curve provided best overall results across domains and feature functions, but as it did not work very well in Experiment 2 it cannot be considered universal and better metrics are needed. A combination of tension curve and one of the string distances might prove useful.

Other future work includes experiments with combination of distance metrics for the clustering algorithm and further enhancements and additional experiments with the random user. Finally, it would be beneficial to experimentally determine, how humans would cluster some of the stories.

ACKNOWLEDGEMENTS

This research is partially supported by SVV project number 260 224 and by student grant GA UK No. 559813/2013/A-INF/MFF.

REFERENCES

[1] M. Mateas, and A. Stern. Façade: An Experiment in Building a Fully-Realized Interactive Drama. In: *Game Developer's Conference: Game Design Track.* (2003).

[2] R. Aylett, M. Kriegel, and M. Lim: ORIENT: Interactive Agents for Stage-Based Role-Play. In: *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, vol. 2, pp. 1371–1372 (2009).

[3] R. Aylett, M. Vala, P. Sequeira and A. Paiva: FearNot! – An Emergent Narrative Approach to Virtual Dramas for Anti-Bullying Education. In: *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling*, LNCS Vol. 4871, Springer, pp. 199-202, (2007).

[4] M. Bída, M. Černý, C. Brom: Towards Automatic Story Clustering for Interactive Narrative Authoring. In: *Interactive Storytelling.* LNCS, Vol. 8230, Springer, pp. 95–106. (2013)

[5] M. Bída, M. Černý and C. Brom: SimDate3D – Level Two. In: *Proceedings of ICIDS 2013*. LNCS, Vol. 8230, Springer, pp. 128-131. (2013)

[6] M. Sarlej, M. Ryan.: Generating Stories with Morals. In: *Interactive Storytelling.* LNCS, vol. 8230, Springer, pp. 217-222. (2013)

[7] P. Weyhrauch: Guiding Interactive Drama. PhD. Thesis, Carnegie Mellon University, Pittsburgh (1997).

[8] S. Ontañón, and J. Zhu: On the Role of Domain Knowledge in Analogy-Based Story Generation. In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 1717–1722, (2011).

APPENDIX – EXAMPLE STORIES

Here, we present several examples of stories from SimDate3D and MOSS domains and show examples of the clustering of stories using the tension curve metric. In both cases we provide simple handcrafted natural language representations of the actions in the story.

A. SimDate3D Domain

Story 1: Thomas went with Barbara to the cinema. After the movie, he was rude to her. They have parted ways. Thomas went to Nataly's home to pick her up. They went out for a walk, but they did not speak much. Thomas insulted Nataly. They met Barbara. An argument started and both girls left Thomas.

Story 2: After the movie, he was rude to her. They have parted ways. Thomas went to Nataly's home to pick her up. Thomas was rude to Nataly. They went out for a walk and Thomas was rude to Nataly. They met Barbara. An argument started and both girls left Thomas.

Story 3: Thomas spent a long time with Barbara in the cinema, then he was very rude on her. Nataly was in the restaurant alone. Then Thomas and Barbara got very angry on each other, but continued talking. Nataly noticed them on their way from restaurant and she run towards them. An argument started and Thomas ended up with Nataly.

Stories 1, 2 and 3 get clustered together in most cases. Stories 1 and 2 are extremely similar and end the same, while story 3 is an example of a story that is relatively similar to the other two, but does not end the same.

Story 4: Thomas and Barbara were on a way to cinema. Thomas asked Barbara to kiss him and to cuddle, she refused. Then they've run into Nataly, argument started and Thomas ended up with Barbara.

Story 5: Thomas and Barbara were going to the cinema. Thomas was making jokes on the way. Before they've get to the cinema they've run into Nataly, argument started and Thomas ended up with Barbara.

Stories 4 and 5 on the other hand are also very similar and end the same but were almost never clustered together.

B. MOSS Domain

Story 1: A wizard gets hungry. He picks up a rose. A troll kidnaps a princess. The troll also kidnaps a dwarf. A knight rescues the princess from the troll. (Generated as an example for recklessness)

Story 2: A wizard gets hungry. He picks up a rose. A troll kidnaps a princess. The troll also kidnaps a dwarf. A dragon gives a treasure to the dwarf. (Generated as an example for recklessness)

Story 3: A dwarf kills a princess. A troll kidnaps the dwarf. A dragon tries to kidnap a unicorn, but fails. Fairy gives magical dust to the dragon. Dragon gives the dust back to the fairy. (Generated as an example for retribution)

All those stories are from the same cluster. While it is clearly visible, how stories 1 and 2 are extremely similar, story 3 seems very different.

aMUSE: Translating Text to Point and Click Games

Martin Černý [1] and **Marie-Francine Moens** [2]

Abstract. In this demo we will show aMUSE — a system for automatically translating text, in particular children stories, to simple 2D point and click games. aMUSE consists of a pipeline of state-of-the-art natural language processing tools to analyse syntax, extract actions and their arguments and resolve pronouns and indirect mentions of entities in the story. Analysed text serves as data the game mechanics operate on, while the story is represented graphically by images the system downloads from the Internet. The system can also merge multiple stories from a similar domain into a branching narrative. Users will be able to both play games created by aMUSE and create games from their own texts using the aMUSE editor.

1 INTRODUCTION

Video games are a powerful media for telling stories and for transferring experiences and feelings in a more general sense. Games are different to most other art forms in that they require active collaboration on the receiver's part. Thus adapting a story to the video game genre requires more than visualisation of the story events on screen: The game mechanics must also be designed to support the story or actively convey parts of the experience.

Recent research has shown that both game design and adaptation of text to game can be, to some extent, performed automatically. Most of the work so far either a) focuses on the game mechanics and does not consider the story of the game, or b) uses a large amount of domain-specific knowledge.

In this demo we will show aMUSE — a system that can automatically translate stories given in natural language to simple games without using any domain-specific knowledge. As our focus is on the story, we have chosen to generate games in the 2D point and click adventure genre. Games in this genre are inherently story-driven and consist of the player clicking on various objects to trigger interactions. If the correct interaction is found, the story progresses further. We have chosen this genre as it allows for a very direct mapping between the story and the game mechanics.

2 RELATED WORK

A system called Angelina can fully automatically design simple 2D and even 3D games [3, 2]. Game-o-matic [9] uses common-sense knowledge databases to generate 2D arcade games involving given topics. Our work is orthogonal to these efforts as it translates a story written in a natural language to a predefined game mechanic instead of generating the mechanics.

In the context of adapting a text to an interactive experience, De Mulder et al. [4] discuss transforming patient guidelines into educational 3D experiences. The authors use a large domain-specific knowledge base to provide common-sense grounding to the fragmentary information present in the text.

Some progress has been made on generating 3D scenes from text to be later used in a whole interactive experience [5]. However, the system is not fully automatic, as it relies on crowdsourced domain-specific knowledge to correctly position the entities in the scene and does not produce playable experiences yet.

3 THE SYSTEM

The aMUSE system consists of four parts: editor, translator, server and frontend. The editor is a graphical application that lets the user enter stories, group stories to form projects and control the execution of the translator. The translator is responsible for finding an interactive representation of the story which is passed to the frontend. For fast startup of the translator and due to some technical aspects of the technologies used, some of the tasks performed by the translator are carried on a dedicated server. The frontend is a simple game engine written in Flash that visualises the game provided by the translator.

To translate a story, the translator first passes it to the server. The most important part of server-side processing is *semantic role labelling* (SRL) using the Lund pipeline[3]. SRL builds upon syntactic features of the sentence to discover *semantic frames*. A frame represents a concept in the sentence (the *root*) and annotated arguments of the concept (the *roles*). We use frame definitions given in PropBank[4].

For example the sentences "The city was taken by the Romans" and "The Romans took the city" have different syntax, but both contain the frame $take.01(taker : Romans, thingTaken : city)$. The numbered suffix to the frame root distinguish between various meanings of the same word: e. g., "I cannot take it anymore" would resolve to $take.02(tolerator : I, thingTolerated : it)$. The Lund SRL was trained on news texts, so we used transfer learning [8] to adapt it to handle stories better.

The last crucial part of server-side processing is coreference resolution using Stanford CoreNLP [6]. Coreference resolution links all mentions of the same entity (pronouns, in particular) throughout the whole story. The annotated text is then returned to the translator.

The translator uses the semantic frames to find possible interactions for each sentence of the story. In our case, interaction is an agent-action-target triplet, where either agent or target may be omitted (but not both). All frames with roots that are verbs are candidates for interactions. Simple hard-coded heuristics are used to choose the agent and the target among the frame's roles.

Now, every story is represented as a linear multigraph with sentences as nodes and possible interactions as edges from the previous sentence to the sentence that defined the interaction. Optionally, the translator can merge multiple stories to form a non-linear story

[1] Charles University in Prague, Czech Republic, email: cerny.m@gmail.com
[2] KU Leuven, Belgium, email: sien.moens@cs.kuleuven.be

[3] http://nlp.cs.lth.se/
[4] http://verbs.colorado.edu/propbank/

graph. To achieve this we check all pairs of sentences A, B. If they are from different stories, but have similar frames then for each interaction (X, A) we add (X, B) to the graph and vice versa, i.e., at these nodes the game can switch to a different story, depending on the interaction chosen by the player. This approach was inspired by the story generation process described in [7].

The translator then lists all the entities present in the story and schedules at which point in the story they should appear. As coreference resolution is not flawless, we make the simplifying assumption that two entities with the same name are the same and merge the respective entity mentions. The translator then requests images for the entities from the server which uses Spritely [1] for this task.

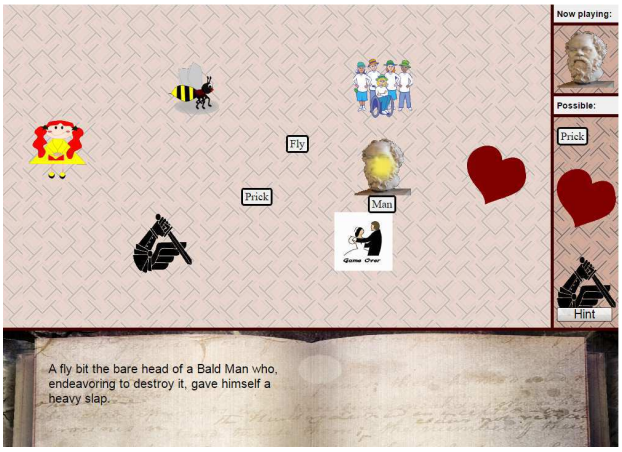


Figure 1. Screenshot of the aMUSE frontend.

The frontend then uses the story graph as the basic structure to guide gameplay. It keeps the current node in the graph and when the user performs an interaction corresponding to any of the outgoing edges, the story progresses to the edge's target, i.e., every action of the user corresponds to progressing the story one sentence further.

Originally, we intended that the user will represent the protagonist of the story and perform only the interactions where he is the agent. In this case, the other interactions would be performed by the system automatically as a kind of a cutscene. This however led to a large number of non-interactive nodes, so we decided to alter the game design a little: the user is no longer a character in the story; he represents a disembodied entity, whose single goal is to make the story happen. To do this, the user can take control of any active entity and act (click on objects) on its behalf. The resulting interactions are very abstract and it is almost impossible to decipher the story from the interactions themselves. To allow the player to follow the story, the original text of the sentence is shown in a stylized book. The screenshot of the frontend is given in Figure 1.

So far, we have not been able to finish our work on extracting spatial relationships between the entities from text, so the entities only float around the screen without any structure.

4 CONCLUSION

Our system is capable to automatically translate stories written in natural language into a specific type of playable experiences. While many of the interactions that the system produces make sense, it also produces absurd options, mostly due to imperfections in natural language processing (NLP). To some extent, this can be enjoyable from the user perspective, but there is definitely room for improvement.

The system works reasonably well on short stories targeted at very small children, as the vocabulary and syntactical structure is simple. However, the main reason that short stories work better than longer ones is that the gameplay is very limited and it is not fun to click through a longer story. Although longer stories also degrade accuracy of coreference resolution. Semantic and syntactic complexity of the text is currently the most limiting factor for our tool. We tested the system on Aesop's fables, where the resulting gameplay was still more often relevant to the story than not. However, when run on fairy tales collected by Andrew Lang, which have long and complex sentences and archaic language style, only a minority of the resulting interactions were reasonable. Further issues arise from incorrect association of words with images.

Our system can serve as a demonstration of the power (and remaining deficiencies) of the contemporary NLP technology. We believe that NLP is at the level where it can improve games and gaming experience. While we are aware of game-related research using syntactic analysis of texts, we are not aware of usage of SRL in this context, although there are high possible benefits.

Examples of games created by the system can be played online[5] and the system itself is fully open-source.

ACKNOWLEDGEMENTS

This research is partially supported by the EU FP7-296703 project MUSE, student grant GA UK No. 559813/2013/A-INF/MFF and by SVV project number 260 224.

REFERENCES

[1] M. Cook. Spritely — autogenerating sprites from the web. `http://tinyurl.com/spritelypost`, (2013). Last checked: 2015-01-16.
[2] M. Cook and S. Colton, 'Ludus ex machina: Building a 3D game designer that competes alongside humans', in *Proceedings of the Fifth International Conference on Computational Creativity*, pp. 54–62, (2014).
[3] M. Cook, S. Colton, A. Raad, and J. Gow, 'Mechanic miner: Reflection-driven game mechanic discovery and level design', in *16th European Conference on Applications of Evolutionary Computation*, volume LNCS 7835, pp. 284–293. Springer, (2013).
[4] W. De Mulder, Q. Ngoc Thi Do, P. Van den Broek, and M.-F. Moens, 'Machine understanding for interactive storytelling', in *Proceedings of KICSS 2013: 8th International Conference on Knowledge, Information and Creativity Support Systems*, pp. 73–80, (2013).
[5] R. Hodhod, M. Huet, and M. Riedl, 'Toward generating 3D games with the help of commonsense knowledge and the crowd', in *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*, pp. 21–27, (2014).
[6] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, 'The Stanford CoreNLP natural language processing toolkit', in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, (2014).
[7] N. McIntyre and M. Lapata, 'Plot induction and evolutionary search for story generation', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1562–1572. Association for Computational Linguistics, (2010).
[8] Q. Ngoc Thi Do, S. Bethard, and M.-F. Moens, 'Text mining for open domain semi-supervised semantic role labeling', in *Proceedings of the First International Workshop on Interactions between Data Mining and Natural Language Processing*, pp. 33–48, (2014).
[9] M. Treanor, B. Blackford, M. Mateas, and I. Bogost, 'Game-o-matic: Generating videogames that represent ideas', in *Proceedings of the The Third Workshop on Procedural Content Generation in Games*, p. 11, (2012).

[5] `http://tinyurl.com/amuseExamples`

Data Collection with Screen Capture

Jason Traish, James Tulip and Wayne Moore[1]

Abstract. Game traces are an important aspect of analysing how players interact with computer games and developing case based reasoning agents for such games. We present a computer vision based approach using screen capture for extracting such game traces. The system uses image templates of to identify and log changes in game state. The advantage of the system is that it only captures events which actually occur in a game and is robust in the face of multiple redundant commands and command cancellation.

This paper demonstrates the use of such a vision based system to gather build orders from Starcraft 2 and compares the results generated with those produced by a system based on analysing log files of user actions. Our results show that the vision based system is capable of not only automatically retrieving data via screen capture, but does so more accurately and reliably than a system relying completely on recorded user interactions.

Screen capture also allows access to data not otherwise available from an application. We show how screen capture can be used to retrieve data from the DotA 2 picking phase in real time. This data can be used to support meta-game activity, and guide in-game player behaviours.

1 Introduction

Game traces allow researchers to follow the evolution of game state as a game is played. Retrieving game traces is necessary to further understand decisions made by players in different game states, and to support the development of AI agents.

Data for board games such as GO [4] and Chess [7] are obtained from a sequential list of user interactions with the game. In GO and chess the user interactions with the game are very limited and the effect of any player action in the game is deterministic. For example, in GO it is known that when a player places a stone such that an opponent's stones are surrounded, then the opponent's stones will be eliminated. However, in commercial RTS games such as Starcraft 2 [1], the set of user interactions for is often far larger than in a classic board game, and the effect of player actions on game state is uncertain. A user can move a camera, move units, construct buildings, train units, buy upgrades and much more. Some of these commands (eg camera movement) have no effect on game state, and for others, (eg unit movement) the effect is indirect. Furthermore, the actual internal game state is inaccessible.

In both GO and Starcraft 2 a game is recorded as a sequence of user interactions. However, while in GO this sequence corresponds directly to changes in game state, in Starcraft 2, multiple redundant commands may be issued in a short space of time, many may never have effect, or they may be cancelled before they are enacted. The only way to tell what actually happens is to play the user interactions back through the game environment.

The other problem that occurs, particularly in RTS games, is that the rules determining how player actions affect the game environment can change due to developers tuning and rebalancing game play. This makes it unfeasible to recreate game state based on user interactions because their effects are constantly changing. In the case of Starcraft 2, while we can access the list of raw commands given by the user via game replay files, access to this list does not allow recreation of game state unless it is used to replay the game using the actual game environment. Once again, the solution is to directly monitor game state changes rather than user inputs.

Lack of access to internal game state makes it difficult to develop AI agents, and much game AI agent research is based on the use of appropriately instrumented simulators. Unfortunately, many of these are highly simplified versions of the original game. Samothrakis [8] suggests that a screen capture approach would solve this problem. Screen capture also offers a standardized way to provide AI agents with input. This is necessary to meaningfully compare the performance of AI agents. Screen capture also allows retrieval of game state from closed source commercial games, and so enables testing of agents using the original game rather than a simulator.

This paper demonstrates the use of a screen capture system to analyse Starcraft 2 build orders. Build orders describe a player's sequence of creating units, buildings, and upgrades to reach a specific strategic goal. The efficiency of a player's build order can significantly impact their chance of winning, and there is considerable research in build order optimization [6]. We demonstrate the extraction of build orders using screen capture and compare the results generated with those produced by Sc2Gears [3]. Sc2Gears calculates build orders based on a log of user interactions. We also demonstrate the real-time use of screen capture to monitor hero selection in the online game Defense of the Ancients (DotA 2) [2]. DotA 2 is a multi player online battle arena (MOBA) game where players select heroes with various characteristics and do battle in teams. Hero selection and team combinations have a large impact on team success, and there is much interest in predicting game outcomes based on hero combinations [9].

2 Screen Capture

The screen capture system models a human observer tracking and recording changes in a game. It identifies areas of the screen which display information relevant to game state and then monitors changes in those areas, interpreting them in terms of game state.

Initially, the area of the game window that the relevant information will appear is specified. Then all patterns showing the information to be recognised in that area of interest are recorded as a set of templates. All templates have the same dimensions to simplify and

[1] Charles Sturt University, Mining Lab, Australia, email: {jtraish & jtulip} @csu.edu.au & wmoore@lisp.com.au

speed up matching. After all templates have been loaded into the system, PCA [5] is used to compress each set of templates down to 30 descriptors per template. This enables a reduction in the number of comparisons for each template and facilitates real-time analysis of captured images. The application is then started with the replay for game trace retrieval. The windows contents are captured via the Windows API and stored in an OpenCV image as often as the refresh rate allows. The image is then decomposed into the identified areas of interest such as the game timer, player's production icons, progress bars, and resource supply. Game specific heuristics are then used to extract information from the screen using template matching and to monitor game state events. The processed game state information is then stored as a game trace.

The screen capture system can be summarised as taking the following steps:

1. Load pre-labeled templates.
2. Decompose templates into basic descriptors.
3. Open the application's associated replay file.
4. Capture the game window using the Windows API.
5. Store and decompose the windows contents into areas of interest.
6. Match templates against areas of interest using a multi-threaded framework.
7. Process the results and store the resulting game trace.
8. Repeat from step 3 to analyse further games.

3 Starcraft 2 Build Orders

Figure 1 displays the replay interface in Starcraft 2 that was used to retrieve game traces. Before starting the process, the system must be aware of where to look for which templates. The templates are stored in sets, one each for the production icons of each player selectable race, and an extra one for other GUI elements. This reduces the number of comparisons necessary as a player can only produce items for their chosen race.

To retrieve the build order we now identify what is displayed using our library of PCA refined descriptors. The top left hand corner of Figure 1 shows seven units/buildings in production. Each item of production shows an identifying image, a number showing how many units are being produced simultaneously, and a green progress bar reflecting the completion percentage of that item. Each different production icon indicates that a build queue is active within that area of interest. In this case we would say that 4 build queues are active for player 1 and 3 are active for player 2. The icon positions are then posted to different worker threads which compare the captured image with an assigned template set. Figures 2 and 3 show the matching templates used to identify production icons collected from the scene shown in Figure 1. Each template is labelled with the name of the production icon.

After identifying the production icon, the game trace heuristic then finds the number on each template as shown in Figure 4.Numbers are identified using a relatively naive yet accurate method. Because numerals are imprinted against a production icon's image they contain a small amount of noise. The noise is reduced by only accounting for pixels that are very similar to white. Then the filtered image is compared against a set of number templates where the closest is selected as the matching number. Following this process identifies the digit shown Figure 4 as matching the template shown in Figure 5.

The completion percentage shown in the production icon is then determined (Figure 6). For this we simply perform a threshold check for predefined pixel values along the length of the progress bar, returning when an empty pixel is found.

Figure 1. Sample screen capture

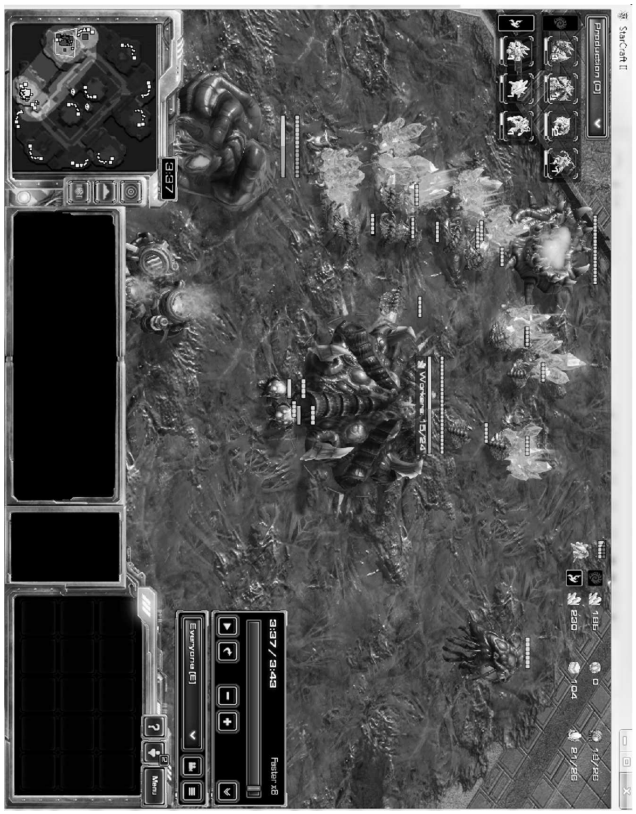


Figure 2. Player 1 - Matching production templates

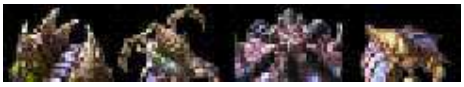


Figure 3. Player 2 - Matching production templates

Figure 4. Digit with noise

Figure 5. Pre-labeled Digit (Matching Template)

Figure 6. Progress bar

Since each template comparison is independent, all template comparisons can be run in separate threads. Once all threads have finished analysing each real-time acquired production icon image the information is used to update each players build order along with the game state, and the game time at which the image was retrieved.

As each player's build order is updated, it is possible that a previously recorded production item is cancelled. If a production item is not listed but was less than 97% complete when last identified then that item is assumed to have been cancelled and is removed from the recorded build order. Within a game of Starcraft 2, this can occur at any point in time when a user selects a production item and cancels it. A cancellation is also noted if the number of items listed within the production icon drops while the current completion percentage is under 97%. This leads to the flaw in the current game trace heuristic that if a production item is almost complete then any number of production items of the same type can be cancelled and they will be falsely recorded as completed. In practice, this rarely occurs.

When a new production item type appears or the production count increases then the game trace heuristic appends that item to the build order. If the number of items in production recorded by a production icon number remains the same for longer than the time to create that item, then another production item of that type is appended to the build order. This deals with the case of when a series of probes/workers are queued. Since they are created one at a time, a constant production count of 1 appears over an extended period. Thus, keeping track of how long it is from when a production icon first appears we can determine when an item repeats production. The exception is when production is halted or paused which can be detected when the progress bar is halted.

After a game has completed, each players build order is recorded to file and the next game is opened and the process repeated. The replay interface is controlled by sending Windows API keyboard messages to Starcraft 2 to display the production icons and accelerate the play back. The replay playback is accelerated to the maximum of eight times the normal playback rate.

4 Comparison with User Interaction Logs

An experiment was conducted to evaluate how the screen capture system performs in capturing a build order in comparison with the established tool Sc2Gears [3]. Sc2Gears applies the user interaction approach to analyse build orders. Both systems were tested using a set of 100 public Starcraft single player versus single player ladder games. Comparisons were made only on the first 10 minutes of game play so that replays of diverse lengths would not affect the results significantly. Each of the 100 games was also processed by a human to generate a ground truth set of build orders. The accuracy for the automated systems was calculated as the number of matching build order steps compared with the human verified sequence.

Table 1 shows that the screen capture technique was able to significantly reduce the number of errors in calculated build orders compared with an analysis based on raw user interactions. The screen capture system still generated a small number of errors in cases where actions were cancelled on the last frame (thus appearing to have actually been completed). Table 2 shows an example of an opening build order extracted using screen capture compared with one using Sc2Gears from the same game. The extracted information is significantly different. Sc2Gears incorrectly identifies the creation of three probes and an additional pylon. In this case, the player requests production of an additional Probe without the necessary resources, a situation that can only be determined by running the game replay. The extra pylon identified by Sc2Gears was the result of the player ordering construction of a pylon and then moments later changing the location of its construction. These errors highlight the issues encountered when using user interaction methods to extract game traces.

Table 1. Error Rates

Error	Screen Capture	Sc2Gears
Mean	0.39%	30.71%
StdDev	0.96%	27.75%

Table 2. Example Game Trace

Sc2Gears	Screen Capture
1. Probe	1. Probe
2. Probe	2. Probe
3. Probe	3. Probe
4. Probe	4. Pylon
5. Probe	
6. Probe	
7. Probe	
8. Probe	

5 Hero Selection in Defence of the Ancients 2

The screen capture technique was also applied to Defence of the Ancients 2 (DotA 2) to test its real-time capabilities of the screen capture framework, and its capacity to generalise beyond Starcraft 2. This section re-enforces the application of the screen capture frame work in retrieving data from a 2D display interface. The data from the DotA 2 interface is retrieved without error and thus no comparison against other methods is given, instead a potential use of the retrieved data is given. The experiment with DotA 2 shows flexibility and versatility of the screen capture approach.

DotA 2 is a multi player online battle (MOBA) game that involves 2 teams of 5 players. Each player must pick a hero, and after a hero is selected and locked in it can not be picked by any other player. Players can select a hero they intend to pick before locking it in, and this is referred to as shadow picking. A shadow pick will only display to the allied team, and is important in influencing the heroes other members of the team will select.

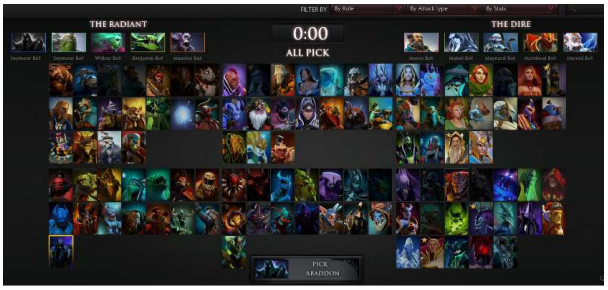
Heroes fall into general categories based on their abilities and how they interact with other heroes within the game. The picking process leads to a diverse set of combinations that can be formed between the 2 teams. However, some of these combinations are weaker than others due to the interaction of hero's strengths and weaknesses. Each hero has synergies with certain allied heroes and/or are able to exploit weaknesses in particular enemy heroes. Thus, it is an interesting problem to see how players adapt their choice of hero during the 1 minute picking phase. It is also interesting to see how these picks can be used to predict the winning team and what rate of success they might have.

In DotA 2, there is much interest in real-time capture of game actions since such a capability offers the potential to support real-time

guidance on hero selection. It also provides information useful to calculating the likelihood of final outcomes. Screen capture potentially can achieve this while user interaction logs are available only after a game has ended.

Figure 7 shows a standard DotA 2 'all pick' mode selection screen. It can be seen that all players have locked in their hero choices except for the player shown on the upper left . This player's portrait is rendered in grey scale to show that it the depicted hero has only been shadow picked. During the picking phase we use the screen capture

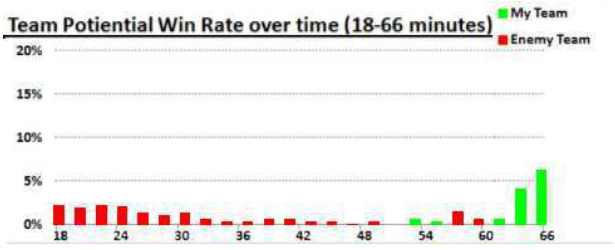
Figure 7. DotA 2 Hero Selection Screen



framework to identify which heroes have been locked in or shadow picked. This data is then analysed using a statistical algorithm based on hundreds of thousands of games of DotA 2.

The current program then displays the win rate for any point in time during the game as shown in Figure 8. This graph can be used loosely to identify when one team is stronger than another and can be used as an indicator for players to become more aggressive within the favoured time zones. It can also be used by lower skilled players to help better identify hero picks that complement their team, and to see what effect their pick would have on the progress of the game. Figure 8 shows that the enemy team has a small advantage that decreases over time until around the 60 minute mark, at which point My Team increases substantially in strength.

Figure 8. DotA 2 Predicted Game Balance



6 Discussion and Further Work

The Starcraft 2 experiment shows that the screen capture approach can help generate more accurate build orders than conventional systems based on logs of player actions. Its application to analysing hero selection in DotA 2 shows that the principles can be applied generally to any game, and for any analytical purpose, using different sets of image templates and different analytical heuristics. The technique can be applied to almost any application where a streaming 2D display record is available. Furthermore, no access to game code or proprietary APIs is required. This opens up data collection and analysis for previously inaccessible games and other applications. The high performance provided by the simplified PCA based image descriptors and parallel template matching allows the development of real-time in-game decision support systems, once again without access to game code or proprietary APIs. The screen capture system takes advantage of using the game display to retrieve actual game events while user interaction logging methods can result in noisy data that can detrimentally affect further analysis.

However, currently screen capture has only been applied to applications where the state is represented with scale and rotation invariant 2D images. There would be considerable challenges in applying the technique to applications that display their state in 3D.

The technique could also be extended to live game data retrieval, such as a Starcraft 2 commentator agent. An agent could be set up to watch two players play a competitive game, giving viewers predictions and feedback in a similar way to how real commentators perform.

The screen capture system could also be used to track in game auction house item prices. The retrieval of the changing value of game items could allow systems to graph, analyse and predict market trends in online worlds.

It could also be used for non game applications such as watching a user's screen and determining the time spent interacting with different windows. This could help system analysts trace work flow and productivity in given applications without access to the source code.

While analytic techniques relying on replays to retrieve game data have to wait until a game has been played and recorded before analysis can be applied, a screen capture system can be used to analyse live games, allowing interested parties to use the data in prediction systems or other applications.

7 Conclusion

Screen capture data retrieval offers great advantages to researchers and applications looking to gather data from complex environments with 2D displays. The system is flexible and more accurate than user interaction logs for such applications.

REFERENCES

[1] 'Blizzard entertainment'. http://www.blizzard.com.
[2] 'Valve corporation'. www.dota2.com.
[3] 'Sc2gears'. https://sites.google.com/site/sc2gears, (2012).
[4] T. Bossomaier, J. Traish, F. Gobet, and P. C. R. Lane, 'Neuro-cognitive model of move location in the game of go', in *(IJCNN), The 2012 International Joint Conference on Neural Networks*, pp. 1–7.
[5] Ian T Jolliffe, *Principal component analysis*, volume 487, Springer-Verlag New York, 1986.
[6] Matthias Kuchem, Mike Preuss, and Günter Rudolph, 'Multi-objective assessment of pre-optimized build orders exemplified for starcraft 2', in *Computational Intelligence in Games (CIG), 2013 IEEE Conference on*, pp. 1–8. IEEE, (2013).
[7] Peter CR Lane and Fernand Gobet, *Using chunks to categorise chess positions*, 93–106, Springer, 2012.
[8] S. Samothrakis, D. Robles, and S. Lucas, 'Fast approximate max-n monte carlo tree search for ms pac-man', *Computational Intelligence and AI in Games, IEEE Transactions on*, **3**(2), 142–154, (2011).
[9] Pu Yang, Brent Harrison, and David L Roberts, 'Identifying patterns in combat that are predictive of success in moba games', *Proceedings of Foundations of Digital Games*, (2014).

Cognitive Navigation in PRESTO

Paolo Calanca and **Paolo Busetta**[1]

Abstract. The PRESTO project has developed an AI infrastructure and an agent framework called DICE for the creation of game-independent, modular NPC behaviours based on a BDI (Belief-Desire-Intention) approach enriched with cognitive extensions for human simulation. Behavioural models can be combined via end-user development tools to form the behavioural profiles of NPCs in a game. Furthermore, PRESTO is producing a set of behavioural models targeted at its pilot project's needs or expected to be of common use. This paper focuses on a fundamental building block: navigation of (human and non-human) characters, implemented as the interplay between a set of behavioural models encapsulating higher-level decision making concerning e.g. speed control, activation of gates, replanning when faced with the impossibility to going forward and lower-level modules for path planning, steering and obstacle avoidance that focus on performance and simpler perception-driven choices. These lower-level modules are embedded into the PRESTO infrastructure and contain a few novel algorithms. The higher level navigation behavioural models in DICE can encapsulate very different physical and emotional profiles; they deal with short-term memory and background knowledge concerning spatial knowledge and impose constraints on path planning based on physical as well as cognitive considerations (e.g. risks or threats). DICE provides the coordination between body-controlling behavioural models (for navigation as well as posture, facial expressions, actioning) and decision-making models representing e.g. the standard operating procedures of professional roles, the cognitive appraisal of events and perceptions, the modality of reaction to unplanned events occurring during a game.

1 INTRODUCTION

PRESTO (Plausible Representation of Emergency Scenarios for Training Operations) [2] aims at adding semantics to a virtual environment and modularising the artificial intelligence controlling the behaviours of NPCs. Its main goal is to support a productive end-user development environment directed to trainers building scenarios for serious games (in particular to simulate emergency situations such as road and industrial accidents, fires and so on) and in general to game masters wanting to customize and enrich the human player's experience. The framework for behavioural modeling in PRESTO, called DICE, was inspired by a BDI (Belief-Desire-Intention) [1, 9] multi-agent system with cognitive extensions, CoJACK [10, 6]. PRESTO offers powerful end-user development tools for defining the parts played by virtual actors (as end user-written behaviours) and the overall session script of a game. PRESTO supports a specific virtual reality, XVR from E-Semble, a well known tool in use for Emergency Management and Training (EMT) in a number of schools and organisations around the world, as well as Unity 3d and, at least in principle, is agnostic with respect to the game engine in use.

The rest of this introduction briefly explains the motivations behind PRESTO with an example and gives an overview of the system. The following sections are dedicated to its navigation subsystem, first discussing lower-level facilities for path planning and steering and then introducing a higher-level layer that takes into account cognitive aspects including memory and appraisal of the perceptions according to the semantics of the environment and the NPC's own psychological profile.

Directing NPCs as virtual actors in a virtual stage. Serious games have the potential to dramatically improve the quality of training in a number of fields where the trainee has to face complex and potentially life-threatening situations. In particular, open-world 3D simulations (also called ''sandbox" or ''free-roaming" games) have been used for quite a long time by the military, with a few products reaching a significant market success, and are becoming common in civilian emergency training because they allow the rapid construction of scenarios for the rehearsal of safety procedures. The main limitation of current technology concerns NPCs, whose behaviour may be quite sophisticated when performing predefined tasks but is often unaffected by context; further, a professional programmer is required for the implementation of any procedure that cannot be described with the simple selection of a few waypoints and the choice of a few actions, let alone introducing variants due to psychological factors. These issues lead to repetitive and hardly credible scenarios and to the slow and costly development of new ones when many NPCs are involved.

As an example, consider a fire breaking in a hospital ward during daytime with patients with different impairments, visitors of various ages and professionals with different roles, experiences and training. In this scenario, which is taken from the pilot project of PRESTO, most characters are NPCs while the human players, i.e. the trainees, are either health professionals that could be in charge for a ward at the time of an accident or emergency staff called to help. A training session would require two apparently conflicting abilities from NPCs. From the one hand, they should act autonomously according to a variety of parameters concerning e.g. their physical and psychological state, their current position, their capabilities; e.g. visitors may act rationally and follow well-marked escape routes or flee panicking to the closest exits, nurses at the start of their shift are fully responsive and careful while at the end of the shift fatigue may lead to errors, and so on. On the other hand, in order to make training effective and engaging, the trainer supervising a simulation session should be able to temporarily suspend it (e.g. to give feedback to the trainees), change the course of events or affect the way certain characters behave (e.g. to introduce more drama or rehearse different procedures), as well as introducing or removing characters in following runs of the same scenario. Hardcoding all possibilities, assuming that this is

[1] Delta Informatica Spa, Trento, Italy, email: name.surname@deltainformatica.eu

supported by the game in use, is a laborious task to say the least.

The objective of PRESTO is to allow NPCs to act as "virtual actors" because they are able to "interpret" a part written at a higher level of abstraction than with common scripting languages, with additional modalities (that may correspond to, e.g., levels of skills or psychological profiles) that can be selected at the beginning but changed during a game as a result of the application of rules or by explicit user choice. The game's master (i.e. the trainer) is empowered to become a "director" able to "brief" virtual actors, that is, to define the parts the artificial characters have to play by means of a language aimed to non-programmers that composes more fundamental even if potentially very complex behaviours into game-specific sequences. Key enablers are end-user development tools [7] and the ability to mix and match behavioural components taken off-the-shelf from a market place (similar in principle to asset stores in popular gaming platforms such as Unity).

Semantics and NPC programming in PRESTO. PRESTO provides facilities for the semantization of the game environment in order to support decision-making based on game- and scenario-independent properties. Most importantly, ontologies are used for the classification of objects and locations and for annotating them with properties and states (called "qualities") that allow abstract reasoning, while navigation areas can be annotated with various properties [5]; some of these aspects are discussed in Sec. 3.

DICE (Fig. 1) supports multi-goal modeling of NPC behaviours, where navigation, body postures and facial expressions, manipulation of objects and decision-making concerning tactical and long-term objectives are controlled by concurrent threads (implemented, in BDI speak, as intention trees achieving independent hierarchies of goals and subgoals). Furthermore, decision-making in DICE happens at two levels, controlled by independent "planned" and "reaction" intention trees. A decision-making behaviour started in reaction to an event pre-empts and blocks the execution of a planned behaviour until it is fully completed, at which point the planned behaviour is resumed. This allows, for instance, to have short-term reactions to perceptions (such as hearing a noise) that partially change the NPC state (e.g. by pointing the head towards the source of the noise) while not affecting navigation or longer-term procedures if not required. All behaviours in the body-controlling intention trees and in decision-making can be overriden by new behaviours at any time, e.g. as new perceptions are processed, as part of a decision-making routine, as a user choice from a GUI, as a command from a PRESTO session-controlling script; at any time, no more than one behaviour for each intention tree is active.

Changes in behaviours due to emotions, fatigue or other non-rational factors can be dealt within DICE in various ways, of which the most novel (and dramatic) is by defining behavioral rules that select alternative models according to the current cognitive state of the NPC. These rules can be defined directly by the end user, who is enabled to change the behavioural profiles of her characters according to the evolution of the game or even in real-time by explicit choice and from the session-level script. As in CoJACK [10], cognitive states are represented in DICE by moderators (i.e. numeric values modeling specific factors such as fear and fatigue levels) and a set of cognitive parameters computed from those moderators (modeling e.g. reactivity and accuracy), even if greatly simplified with respect to the original. Any behavioural model, including navigation, can use moderators and cognitive parameters to tune its own internal parameters, e.g. to decide the speed of execution of action or memory fading. Changes to moderators are normally performed by behavioural models for cognition according to appraisal rules (concerning e.g. the perception of threatening things) and time; however, it is possible to force the value of moderators at any time from any behavioural model (e.g. because of the realization of a dangerous situation) or from the session-controlling script, thus allowing the trainer to fully control the overall behaviour of an NPC during a game.

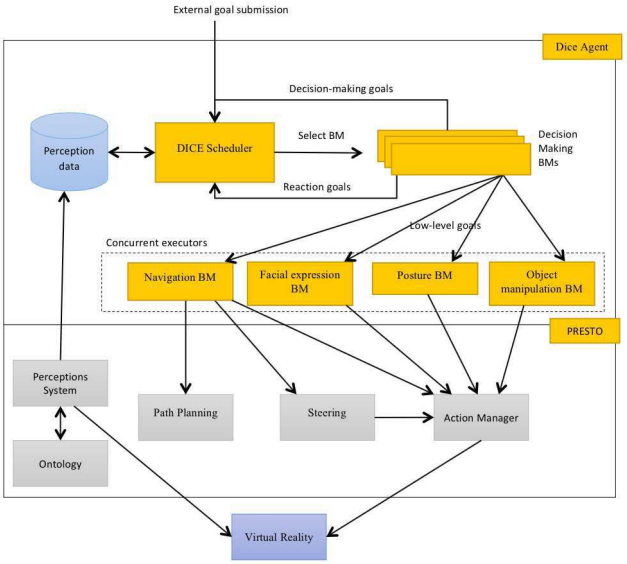


Figure 1. Simplified DICE architecture with navigation highlighted (BM: Behavioural Model)

One of the implications of the DICE approach on navigation is that, at any time, the travel direction (decided by a behaviour) can be changed and may be resumed later (e.g. when a reaction is completed). The APIs make programming this concurrent machinery a straightforward business, while the end-user development tool for behaviour modeling (called the DICE Parts Editor) provides an extremely powerful yet intuitive way to write scripts that affect one or more intention trees at each step [8].

As mentioned earlier, PRESTO has a facility to edit and control session-level scripts inspired by interactive books. A session script is composed by a set of scenes connected as a graph. At each scene, goals can be given to NPCs, their internal state changed (including emotions) and objects manipulated. The trainer starts a script at the beginning of a training session and advances it by manually navigating the graph of scenes or letting PRESTO choose the next one e.g. when certain events happen or when a timer expires. This allows a large, potentially unlimited number of different sessions to unfold from a single script with no need to reprogram the NPCs once equipped with all required behavioural models. In the hospital ward example presented earlier, the initial scene would command visitors, patiens and nurses to accomplish their routine goals; the script may continue with alternative scenes such as "fire breaking in a patient room" or "fire breaking in a surgical facility", each with different people involved, and then with sequences that may lead e.g. to smoke filling the area and visitors fleeing or an orderly managed situation with the intervention of fire fighters, chosen according to the decisions of the trainer and the events occurring during a session.

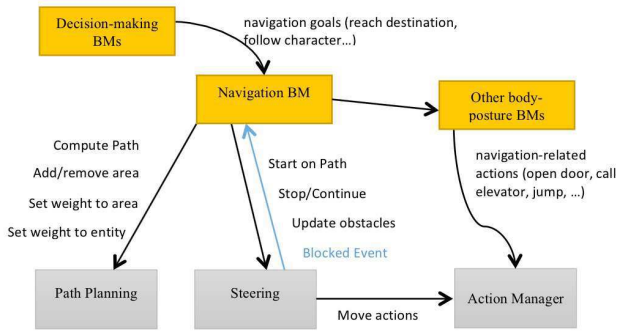


Figure 2. Navigation subsystem architecture

2 NAVIGATION ARCHITECTURE

The overall architecture of navigation within a DICE agent, shown in Fig. 2, closely resembles a standard model [3], with a path planning module, a steering module looking after actual body movements and simple obstacle avoidance, and the navigation behavioural model calling the path planner and the steering modules according to the goals provided by decision-making (e.g., of reaching a destination, of following another character, and so on).

The path planner uses a navigation graph which is instantiated for each agent and modified by the navigation behavioral model to reflect memory, navigation decisions and specific capabilities. From this graph, the path planner can compute one ore more paths to the desired destination and the behaviour can choose which one to follow based on any attached information. Once a path is chosen, the steering module is invoked by the behaviour to move along it. State information on the steering activity for a specific path, including an explanation in case of unsuccessful conclusion (e.g., facing a gate, impassable obstacles, aborted by another steering request typically generated by a reaction), is used by the behaviour to track progress and possibly perform actions to resume navigation. Analogously, the state of a goal given to the navigation behaviour is reported on a tracking object that allows higher-level decision-making behaviours to know when the goal has been satisfied or the reason for failure, including abort caused e.g. by a reaction submitting a different navigation goal.

The flow of perceptions goes to steering as well as to all behavioural models to update their own internal state. As a consequence, the navigation goal being currently pursued may be changed because e.g. of a reaction or the decision to take a different course of actions.

3 MESHES, AREAS AND SEMANTICS OF THE ENVIRONMENT

Configuration information affecting navigation is distributed in three main data structures, two of which concern meshes and are directly used by the navigation modules while the third is related to semantics for the decision-making layer.

Navigation meshes and navigation areas. PRESTO uses navigation meshes (that is, sets of adjacent convex polygons that share edges and cover a walkable / drivable / otherwise navigable surface) [11] to compute safe and efficient paths through the environment, avoiding walls, obstacles and precipices. Navigation meshes can be automatically built from the environment geometry and from parameters including the navigating object's radius, height and max acceptable steepness, so it is possible to generate meshes specialized per character type (including non-humans, e.g. vehicles).

Semantics data on the navigation meshes, such as the terrain type and traffic constraints (permitted directions, reserved paths, ...), can be added with a tool that allows the creation and annotation of navigation areas by selecting polygons of a mesh. Furthermore, as discussed below, behavioural models manipulate areas rather than polygons of a mesh.

Locations of Interest and navigation-affecting entities. PRESTO allows the end-user to classify and annotate locations of interests and objects within the environment with semantic information taken from an ontology. This is composed of a domain-independent core and one or more domain-specific extensions [5] and determines which behavioural models can be used in a specific game; for instance, the current PRESTO pilot project contains a hospital ontology that is used by models of nurses and doctors while a generic safety ontology is used by fire fighters. A small part of the semantic annotations is directly managed by the navigation subsystem as discussed later, most importantly the property of being a "gate", i.e. anything that has a state of openness that can be manipulated by a character. Being a gate is not automatically related to the classification of the object (e.g., a door is not a gate if it is permanently closed) and may even change dynamically. Anything else that may affect what the character does during its movements is handled by other behavioural models and especially by decision-making models. This separation of concerns relies on the possibility offered by DICE to stop and change navigation goals at any time, possibly as reactions that simply delay rather than abort the procedure being executed by a character.

4 LOWER-LEVEL NAVIGATION FACILITIES

Higher-level behavioural models and lower-level facilities share a navigation graph, manipulated by behaviours and used by the path planner, and status information on the current steering activity. A set of APIs allow behaviours to affect the navigation graph, invoke the path planner and trigger steering.

Navigation graph and path planning. The Path Planning module uses a navigation mesh to build a polygon adjacency graph, which in turn is used as navigation graph shared with the behavioural models. While navigation meshes are generated off-line and shared by all agents, a navigation graph is specific for each agent since it is based on the background knowledge of the agent, its capabilities, its memory and its decisions. For instance, the configuration of the background knowledge of an agent specifies which mesh to use and how much of it is known at the beginning of a game; furthermore, behavioural models can add or remove navigation areas (converted in polygons by the Path Planning API).

Edges in the navigation graph carry a weight, by default representing the euclidean distance between the centroids of the two polygons correspondent to two graph nodes. These weights can be manipulated by behavioural models to convey preferences to the path planner; this is done by specifying the weight for an entire area, which is like altering the area's distance from the remaining navigable areas.

The path planner computes the shortest path from a source point to a destination point by using the weigths and applying the well known A^* algorithm.

Steering and obstacle avoidance. The steering module moves the NPC controlled by the agent along a path computed by the path plan-

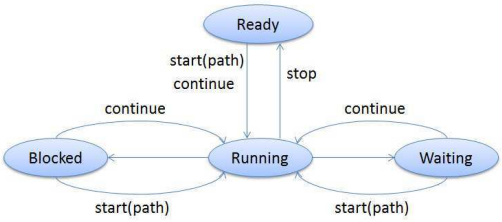


Figure 3. Steering FSM

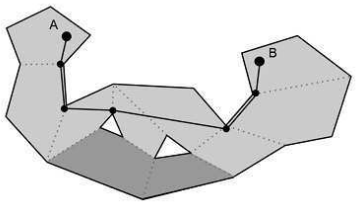


Figure 4. Trajectory generated by the Funnel algorithm

ner. To this end, it computes and updates a trajectory that avoids obstacles and moves the NPC along the points of the trajectory. While the path is computed from the start point to the destination point, the trajectory is computed locally, that is, from the current NPC position up to a maximum distance. The trajectory is frequently updated so that it continuously adapts to changing conditions. The trajectory is computed inside a "global path", i.e. the sequence of polygons computed by the path planner forming a tunnel in the selected navigation mesh. Only obstacles inside this tunnel, perceived by the agent and close to the current position of the NPC are considered by steering, which considers also their semantic properties; most importantly, objects classified as gates and in a "closed" state are not avoided. When the agent perceives that an obstacle has moved then the trajectory is immediately re-calculated.

Steering is a Finite State Machine, illustred in Fig. 3. The agent (that is, its navigation behavioural model) can query its state and send inputs that will cause state transitions; in particular, the behaviour can start steering on a selected path, stop it and later resume it on the current path or re-start it on a different one.

While Running, steering moves the NPC by calling PRESTO's "MOVE" action, which in turns controls the body's animation concerning legs or other moving parts (e.g. wheels), translate the NPC in space at the desired speed and adjust the NPC position on the ground. MOVE modifies the speed according to its initial value, providing any required acceleration; a complementary STOP action decelerates the NPC.

The Blocked state is entered when steering fails in computing a trajectory because the path is obstructed by too many obstacles. As discussed below, it is left to the behaviour to take a decision, e.g. waiting and later resuming or temporary removing the obstructed polygon from the agent's navigation graph and recomputing the path.

The Waiting state is entered when the NPC cannot go further because it is in front of a closed gate. Steering moves the NPC to an appropriate distance before entering Waiting. At this stage, the behaviour has to take an action depending on the gate's type, for example a door must be opened or an elevator must be called. Once the action has been performed, steering can be resumed. Note that the behaviour may decide to abort steering and change path because, for instance, the opening action fails for some reason not under navigation's control (e.g., the goal of opening a door cannot be achieved because a key is required and not owned by the NPC).

Steering trajectory computation. The trajectory is first computed ignoring ostacles, using the Funnel algorithm [4]. This algorithm is also known as "string-pulling" because the trajectory being generated is like a string pulled from the two extremes (Fig. 4).

The generated trajectory is modified to avoid obstacles, represented with simple geometries, like circles and rectangles, enlarged by the agent radius; an example of the algorithm is in Fig. 5. As first

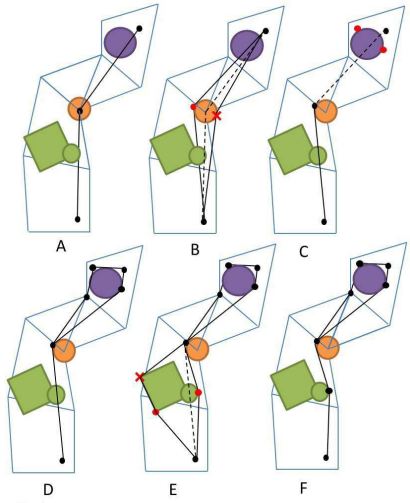


Figure 5. Obstacle avoidance algorithm, A: the output of the Funnel algorithm. B: the trajectory point is inside the orange obstacle, the right side is rejected. C: the new segment intersects the violet obstacle. D: the trajectory is recomputed to attach the two sides, but the first segment intersects the green obstacle cluster. E: the left side is rejected because a point is out of the path. F: the two final trajectories.

step, obstacles that intersect each other are clustered; each isolated obstacle forms a cluster by itself. Then each cluster is checked for intersections with the trajectory segments. If a segment intersects the cluster, the segment is discarded and two poly-lines are computed from its starting point to its ending point, passing to the right side and to the left side of the cluster. If no poly-lyne is within the path, the steering state is set to Blocked and the algorithm is stopped, eventually invoking the higher-level behavioural model. If exactly one of the computed poly-lines is inside the path, then the intersecting segment is substituted with that one. If both the poly-lines are inside the path, then the trajectory is duplicated. At this stage the checking process is repeated recursively on the resulting trajectories to handle further intersections with other clusters. The final output of the algorithm, if successful (i.e. if the Blocked state is never reached), is one or more trajectories; one is eventually chosen at random, to prevent the oscillations that typically arise when NPCs facing each other use the same deterministic steering algorithm.

5 HIGHER-LEVEL NAVIGATION BEHAVIOURAL MODELS

Navigation control in DICE is split in two types of behavioural models. One type, identified as "navigation BM" in Fig. 1 and 2, satisfies the navigation goals submitted by decision-making behaviours (e.g., of reaching a destination); slightly different navigation models are provided that depend on the main physical features of the NPC, e.g. of being a human rather than a vehicle, and consequently on the NPC's ability to move and affect the environment. As mentioned above, the navigation BM runs in its own intention tree (thread of execution) concurrently with decision-making and other body-controlling behaviours. The navigation BM calls path planning and controls steering, acting according to the latter's indication in particular when entering the Blocked or Waiting states. A number of different decisions can be taken according to the model and to the semantics of gates or obstructing objects, which may in turn cause goals to be submitted to other body-parts behaviours (e.g. opening a door, calling a lift, and so on).

A second type of behavioural model, referred to as "navigation capabilities" and included as a decision-making module in DICE, looks after some of the cognitive aspects of navigation. In particular, the navigation capability of an NPC decides which mesh to use on creation, then changes the default speed, default animations and so on according to the current sub-rational state of the agent (i.e. its moderators and cognitive parameters). Thus, PRESTO can provide capabilities specialized e.g. for quiet or excited people, for permanent or temporary physical impairments, for different types of vehicles, and so on. Navigation capabilities may access the cognitive state to tune their parameters (e.g. speed or animations); furthermore, behavioural rules may be defined to switch navigation capabilities entirely during a game depending on the NPC's moderators. For instance, a high level of fear may select a model whose default speed is running and movement animations jerky, while a high level of fatigue may select a model doing exactly the opposite. Furthermore, the navigation capabilities satisfy goals concerning path selection, such as "stay out of sight of entity E" or "don't go thru location L" (which may have been classified as dangerous by a decision-making model according to the appraisal rules of the agent), by taking note of what to avoid and manipulating the navigation graph accordingly, based on current knowledge and the flow of perceptions.

Behavioural models in DICE have their own configuration parameters, called "background knowledge". As mentioned above, the background knowledge of the navigation capability of an agent determines how much the agent knows *a priori* about the environment – it can be everything or being limited to a few areas; the navigation graph is created accordingly. The flow of perceptions arriving from the PRESTO infrastructure includes also the visible navigation polygons of the various meshes; this data is used by the navigation capability to update the navigation graph. The cognitive model of DICE, not discussed here, looks after short-term memory management, which includes calling the navigation capability to purge the navigation graph; that is, the agent literally forgets about where to navigate according to timing and frequency of perceptions from the environment. Out of scope of the navigation subsystem, and not discussed here, is a "search" behaviour, which is a set of decision-making procedures that can be started when a navigation goal fails with an "unknown path" error.

In the hospital fire scenario presented in the introduction, the navigation capability of a patient on a wheel chair would use a different mesh than the one selected for a visitor with normal walking capabilities, e.g. to avoid steps and stairs. The patient's background knowledge would include the navigation areas of the entire ward (since she has been there for a while) while the visitor's knowledge would be initially empty and populated while she moves in the ward; a decision-making procedure of the visitor that invokes a goal such as "go to patient room nr. 3" would initially fail because, indeed, no path can be computed and a search behaviour would need to be invoked allowing the progressive discovery of the navigation areas of the selected mesh. If, at any time during the game, a fire alarm starts ringing, its perception on both visitor and patient would trigger a (decision-making) reaction that is handled different according to the currently active behavioural models, which in turn may depend on cognitive states such as fear. The perception of smoke and fire would submit goals such as "don't go thru that area" handled by the navigation capability as mentioned above. A rationally-behaving NPC that knows the position of a location ontologically classified as "fire exit" would navigate to the latter, with a speed and a modality that depend on the currently active navigation capability (excited / not excited, walking / pushing the wheel chair); an NPC that doesn't know about fire exits or that it's too fearful to act rationally would run to the closest exit.

Queuing and other coordinated behaviour. Steering looks after obstacle avoidance and thus somehow takes care of certain crowding behaviours. However, proper coordination is a matter for decision making at least partially outside of the scope of navigation. Work is in progress on game-theoretical descriptions of queuing and access to shared resources that allow the definition of policies at a very abstract (meta-) level. This exploits the support in DICE for introspection, semantic tagging of goals and plans, dynamic assignment and aborting of goals and intentions as well as the ability to dynamically manipulate semantic tags of any entities (including NPCs) offered by PRESTO. The specification of policies is expected to substantially reduce the coding required by models and allows the reuse of the same coordination patterns in many different situations, e.g. for queuing to pass through a gate (which will be part of the navigation BMs) as well as for queuing at the entrance of an office or at the cashier in a supermarket (which are decision-making behaviours not related to navigation goals).

6 CONCLUSIONS AND FUTURE WORKS

At the time of writing, testing and performance evaluation are still in progress. Initial results show that the navigation meshes are surprisingly small even in very large and complex indoor and outdoor environments; in turn, this makes the maintenance of per-agent navigation graphs and path planning computationally well affordable. Other work in progress concerns coordinated behaviour, as discussed above.

While the navigation algorithms described in this paper contain a few novelties, we believe that the most interesting part of the PRESTO approach is the coordination among navigation behaviour, other concurrent body-controlling intentions and the two-level decision making, all affected by cognitive elements such as short term memory management and emotions. When combined with its semantic facilities and end-user development tools for the creation of NPC behavioural profiles, PRESTO represents an interesting improvement to the state-of-the-art of game platforms, especially for serious game development.

ACKNOWLEDGEMENTS

We thanks all other members of Delta Informatica's technical team (Matteo Pedrotti, Mauro Fruet and Michele Lunelli). PRESTO has been funded by the Autonomous Province of Trento (PAT), Italy.

REFERENCES

[1] Michael E. Bratman, *Intention, Plans, and Practical Reason*, Harvard University Press, November 1987.

[2] Paolo Busetta, Chiara Ghidini, Matteo Pedrotti, Antonella De Angeli, and Zeno Menestrina, 'Briefing virtual actors: a first report on the presto project', in *Proceedings of the AI and Games Symposium at AISB 2014*, ed., Daniela Romano, (April 2014).

[3] Alex J. Champandard, *An Overview of Navigation Systems*, volume 2 of *AI Game Wisdom*, 131–139, Charles River Media, Massachusset, 2004.

[4] Xiao Cui and Hao Shi, 'An overview of pathfinding in navigation mesh', *IJCSNS International Journal of Computer Science and Network Security*, **12**, 48–51, (December 2012).

[5] Mauro Dragoni, Chiara Ghidini, Paolo Busetta, Mauro Fruet, and Matteo Pedrotti, 'Using ontologies for modeling virtual reality scenarios', in *to appear in Proceedings of ESWC 2015*.

[6] Rick Evertsz, Matteo Pedrotti, Paolo Busetta, Hasan Acar, and Frank Ritter, 'Populating VBS2 with Realistic Virtual Actors', in *Conference on Behavior Representation in Modeling & Simulation (BRIMS)*, Sundance Resort, Utah, (March 30 – April 2 2009).

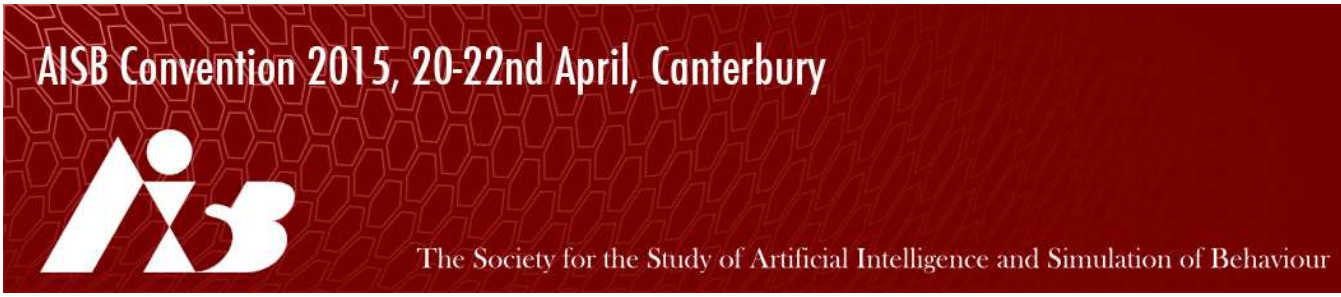
[7] Henry Lieberman, Fabio Paternò, Markus Klann, and Volker Wulf, 'End-User Development: An Emerging Paradigm', *End User Development*, **9**, 1–8, (2006).

[8] Zeno Menestrina, Antonella De Angeli, and Paolo Busetta, 'APE: end user development for emergency management training', in *6th International Conference on Games and Virtual Worlds for Serious Applications, VS-GAMES 2014, Valletta, Malta, September 9-12, 2014*, pp. 1–4. IEEE, (2014).

[9] Anand S. Rao and Michael P. Georgeff, 'Bdi agents: From theory to practice', in *IN PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON MULTI-AGENT SYSTEMS (ICMAS-95*, pp. 312–319, (1995).

[10] Frank E. Ritter, Jennifer L. Bittner, Sue E. Kase, Rick Evertsz, Matteo Pedrotti, and Paolo Busetta, 'CoJACK: A high-level cognitive architecture with demonstrations of moderators, variability, and implications for situation awareness', *Biologically Inspired Cognitive Architectures*, **1**, 2–13, (July 2012).

[11] Paul Tozour and Ion Storm Austin, *Building a Near-Optimal Navigation Mesh*, 171–185, AI Game Wisdom, Charles River Media, Massachusset, 2002.



AISB Convention 2015

University of Kent, Canterbury

Proceedings of the AISB 2015 Symposium on Updating the Anti-representation Debate: Behavior-oriented Perspectives

Edited by Martin Flament Fultot

ISBN: 978-1-908187-48-2

Introduction to the Convention

The AISB Convention 2015—the latest in a series of events that have been happening since 1964—was held at the University of Kent, Canterbury, UK in April 2015. Over 120 delegates attended and enjoyed three days of interesting talks and discussions covering a wide range of topics across artificial intelligence and the simulation of behaviour. This proceedings volume contains the papers from the Symposium entitled *Updating the Anti-representation Debate: Behavior-oriented Perspectives*, one of eight symposia held as part of the conference. Many thanks to the convention organisers, the AISB committee, convention delegates, and the many Kent staff and students whose hard work went into making this event a success.

—Colin Johnson, Convention Chair

Introduction to the Symposium

For more than 30 years now, two camps have been arm-wrestling about the place of representations in cognitive science. Radicals claiming that no representations whatsoever are deployed in biological systems in their everyday lives on the one side, radicals claiming that strict formal computation is all there is to cognition on the other side, and many other thinkers taking intermediate positions.

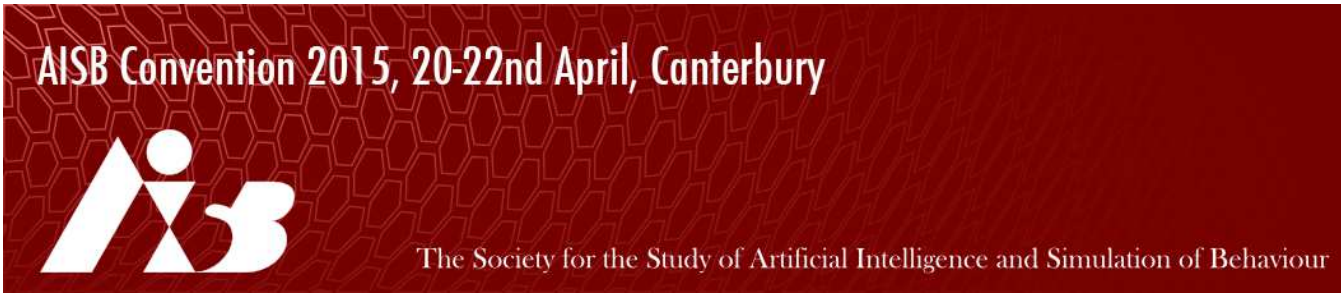
However, “representation” battles are seldom fought around the notion itself. Different answers to the question “how would the cognitive system benefit from deploying representations?” are provided according to different cognitive aspects or phenomena under observation. Thus, it has been claimed that while some phenomena could do well with non-representational explanations, others are more “representation hungry”. The problem with this ecumenic solution is that it owes an explanation of how such different cognitive solutions as those implied by representational and nonrepresentational mechanisms can cohabit harmoniously in, say, the brain, and be implemented by the same physical resources.

As an answer to this problem, one tradition among non-representationalists has been to focus on so-called “low level cognition” and to try to incrementally build increasingly complex models of cognition without having to appeal to representational mechanisms. As it turns out, of course, “low-level” does not mean “simple”. This is especially true in the case of adaptive behavior, since it presents the particular characteristic of having to cope with a changing environment, in real time, through the coordination of bodily activity, and in a flexible and meaningful way. Interestingly, this degree of complexity rather than systematically driving research towards representational solutions has proven a very fertile ground for the research of deeply interesting non-representational mechanisms that are far from trivial, contrary to what the notion of “low-level” would imply. Moreover, there is reason to think the non-representational mechanisms of behavior could underlie the phenomenon of cognition itself. Adaptive behavior would not be merely an end-product of cognitive processes (“output”), but cognition could be grounded--if not consist--in the very ongoing shaping of those sensorimotor patterns. However, there is not a consensus on the interpretation of these mechanisms as non-representational, thus the question always requires considerable philosophical efforts to make sense of the relation between behavior, cognition and representation.

This symposium aims to update the current state of research in behavioral phenomena and to make progress in the elucidation of the role of representations in the face of recent developments. The talks will address the following topics : the consequences of considering behavioral systems as agent-body-environment systems ; the relationship between representational states and properties of dynamical flows describing the evolution of behavior ; navigation through action-orienting recognition instead of place recognition ; how to design controllers for tensegrity-based locomoting bodies ; an interpretation of off-line cognition from a movement coordination perspective ; contrasting the notion of ‘agent’ with the notion of ‘organism’ as cognitive systems and how each notion places a different emphasis on representation ; dynamic behavior based origins of life ; mechanisms for muscle coordination and the possibilities of their being constituent of the rest of cognition ; defense of the use of representations to understand motor control.

—Martin Flament Fultot, Symposium Organiser

Contents



AISB Convention 2015

University of Kent, Canterbury

Proceedings of the AISB 2015 Symposium on Computational Creativity

Edited by Mohammad Majid al-Rifaie, Jeremy Gow and Stephen McGregor

ISBN: 978-1-908187-50-5

Introduction to the Convention

The AISB Convention 2015—the latest in a series of events that have been happening since 1964—was held at the University of Kent, Canterbury, UK in April 2015. Over 120 delegates attended and enjoyed three days of interesting talks and discussions covering a wide range of topics across artificial intelligence and the simulation of behaviour. This proceedings volume contains the papers from the *Symposium on Computational Creativity*, one of eight symposia held as part of the conference. Many thanks to the convention organisers, the AISB committee, convention delegates, and the many Kent staff and students whose hard work went into making this event a success.

—Colin Johnson, Convention Chair

Introduction to the Symposium

Over the last few decades, computational creativity has attracted an increasing number of researchers from both arts and science backgrounds. Philosophers, cognitive psychologists, computer scientists and artists have all contributed to and enriched the literature.

Many argue a machine is creative if it simulates or replicates human creativity (e.g. evaluation of AI systems via a Turing-style test), while others have conceived of computational creativity as an inherently different discipline, where computer generated (art)work should not be judged on the same terms, i.e. as being necessarily producible by a human artist, or having similar attributes, etc.

This symposium aimed at bringing together researchers to discuss recent technical and philosophical developments in the field, and the impact of this research on the future of our relationship with computers and the way we perceive them: at the individual level where we interact with the machines, the social level where we interact with each other via computers, or even with machines interacting with each other.

This 2nd International Symposium on Computational Creativity (CC2015) featured a number of presentations covering a range of topics in the evolving field of Computational Creativity. Issues addressed will include practical work in the area, theoretical approaches to creativity, and philosophical questions raised on the potential of non-human “creative” agents.

Topics of interest for this symposium included, but were not limited to: novel systems and theories in computational creativity, in any domain (e.g. drawing and painting, music, story telling, poetry, games, etc); the evaluation of computational creative systems, processes and artefacts; theory of computational aesthetics; representational issues in creativity, including visual and perceptual representations; social aspects of computational creativity, and intellectual property issues; creative autonomy and constraint; computational appreciation of artefacts, including human artworks.

We would like to thank all the members of the Programme Committee for the generous support and excellent work in evaluating the submissions.

—Mohammad Majid al-Rifaie, Jeremy Gow

Organising Committee

- Mohammad Majid al-Rifaie (Goldsmiths, University of London, UK)
- Jeremy Gow, (Goldsmiths, University of London, UK)
- Stephen McGregor (Queen Mary, University of London, UK) – Publicity

Programme Committee

- Mark Bishop (Goldsmiths, University of London, UK)
- Simon Colton (Goldsmiths, University of London, UK)
- Mark d’Inverno (Goldsmiths, University of London, UK)
- Pablo Gervás (Universidad Complutense de Madrid, Spain)
- Bipin Indurkhya (AGH University of Science and Technology, Kraków, Poland)
- Mohammad Ali Javaheri Javid (Goldsmiths, University of London, UK)
- Anna Jordanous (Kent University, UK)
- Francois Pachet (SONY Computer Science Laboratory Paris, France)
- Alison Pease (University of Dundee, UK)
- Georgi Stojanov (American University of Paris, France)
- Dan Ventura (Brigham Young University, USA)
- Geraint Wiggins (Queen Mary, University of London, UK)

Contents

Tightening the Constraints on Form and Content for an Existing Computer Poet

Pablo Gervás [1]

Abstract. Existing systems for the automated generation of poetry often attempt to simplify the task by taking advantage of free-form poetry - to avoid the need to achieve rigorous poetic form - and poetic licence - to avoid the need of conveying a specific message at a semantic level. This is acceptable as an initial step, but once acceptable solutions have been found for the simplified version of the problem, progress can be made towards higher goals by enriching the initial problem statement. The present paper describes an attempt where an existing computer poet, originally developed to produce poems in a given form but with no specific constraints on their content, is put to the task of producing a set of poems with tighter restrictions on both form and content. Alternative generation methods are devised to overcome the difficulties, and the various insights arising from these new methods and the impact they have on the set of resulting poems are discussed in terms of their potential contribution to better poetry generation systems.

1 Introduction

Computer generation of poetry is a flourishing area of research in the context of computational creativity. In the last few years there has been a significant increase in the number of approaches to the task, and an extension to work in languages previously untried. By its nature, the task of generating a poem, when addressed by either a computer or a human, has to satisfy constraints at two very different levels. One level concerns the sequence in which the words appear in the poem. For a draft to be acceptable there has to be some way in which the words in it appear to link to one another, to make sense as a linguistic message. This constraint is applicable to the whole poem but essentially it operates at a local level, based on how each word can be seen to follow on from the previous one. A different level concerns certain macro-structural features that may be desirable in a poem, such as being distributed over a number of lines of specific lengths in terms of syllables, or having rhyming words occur at the end of particular lines. This corresponds to the poem satisfying some form of poetic stanza.

The problem of poetry generation is in fact rather more complex, because these two levels of constraints are just formulations of the overall specification at the extremes of a continuum. In truth, the way in which the sequence of words builds up is also expected to satisfy constraints on form - usually based on the relative positions of stressed syllables within a line, sometimes expressed in terms of feet - and there must also be some sense to be made between the different parts of the poem at a linguistic level. This is why human quality poetry is a tall order that few computer programs can tackle to the satisfaction of their critics. However, two higher level characteristics of poetry can be exploited to simplify the problem from an engineering point of view. First, poetry can also exist in free form, where constraints on line length, stress patterns or rhyme may be waived in favour of a more expressive poem at a semantic level. Second, the concept of poetic licence allows poets to sometimes violate linguistic expectations in favour of a more pleasing poem in terms of form. Traditionally, these two characteristics are applied in opposition to one another: if free-form is chosen for a poem, it is usually so that its linguistic expression does not have to be forced in any way to express the poet's meaning; if poetic licence is applied, it is usually to fit the poet's meaning into a particular poetic form where conventional phrasings might not work. Computer generated poetry often operates at the confluence of these two approaches relying on one to avoid the need to achieve rigorous poetic form and on the other to avoid the need of conveying a specific message at a semantic level. As the full problem is so complex, it is acceptable to apply a certain degree of simplification so that progress can be made in spite of the difficulties. However, the original goal must be kept in mind, so that once acceptable solutions have been found for the simplified version of the problem, progress can be made towards it by enriching the initial problem statement.

The present paper describes such an attempt. An existing computer poet, originally developed to take advantage of the characteristics of poetry described above, is set to the task of producing a set of poems with tighter restrictions on both form and content. The approach previously followed to poetry generation is shown to have limitations when the task is rephrased in this way. These limitations are analysed in terms of the current theoretical descriptions of computational creativity, and alternative generation methods are explored.

2 Previous Work

The work presented in this paper brings together some of the existing theoretical accounts of computational creativity and a number of efforts for computer generation of poetry. Both of these separate topics are reviewed in the present section.

2.1 Computational Creativity

Much of the work done on computational creativity over the past few years has been informed by Margaret Boden's seminal work describing creativity in terms of search over a conceptual space [3]. Boden formulated the search of ideas in terms of search over a conceptual space. Such a conceptual space would be defined by a set of constructive rules. The strategies for traversing this conceptual space in search of ideas would also be encoded as a set of rules. This view of

[1] Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, email: pgervas@ucm.es

computational creativity was taken a step further in [25] by specifying formally the different elements involved (the universe of possible concepts, the rules that define a particular subset of that universe as a conceptual space, the rules for traversing that conceptual space, and a function for evaluating points in the conceptual space reached by these means).

In his pioneering work on the evaluation the creativity of computer programs, Ritchie [20] outlined a set of empirical criteria to measure the creativity of the program in terms of its output. Ritchie's criteria are defined in terms of two observable properties of the results produced by the program: novelty (to what extent is the produced item dissimilar to existing examples of that genre) and quality (to what extent is the produced item a high-quality example of that genre). He also put forward the concept of *inspiring set*, the set of (usually highly valued) artefacts that the programmer is guided by when designing a creative program. Ritchie's criteria are phrased in terms of: what proportion of the results rates well according to each rating scheme, ratios between various subsets of the result (defined in terms of their ratings), and whether the elements in these sets were already present or not in the inspiring set.

This idea of the inspiring set was taken a step further in [10], where the issue of how systems might take their prior output into account when evaluating the novelty of subsequent artifacts. This lead to the introduction of the concept of a *dynamic inspiring set*, one where system outputs are progressively updated into the inspiring set so they can inform later generative processes.

Colton and Wiggins [6] introduced the term *curation coefficient* to identify the proportion of system results that an impartial observer of system output would be happy to present to third parties. When estimated for a system addressing creative tasks it provides a reasonable measure of how much of the merit of presented system output can be attributed to the system itself and how much to the person actually selecting which particular outputs to present.

2.2 Computer Generated Poetry

Computer generation of poetry has traditionally addressed the constraints outlined in section 1 in terms of two different strategies: one is to reuse large fragments of text already formatted into poem-like structures of lines [7], and the other is to generate a stream of text by some procedure that ensures word-to-word continuity and then establish a distribution of the resulting text into lines by some additional procedure.

The reuse of text fragments already distributed into poetic lines was poineered by [17, 16] and it has more recently been used by [23, 12, 5, 24, 21, 4, 19]. In all these cases, either lines or larger poem fragments from exiting poems are subjected to modifications - usually replacement of some of the words with new ones - to produce new poems. A refinement on this method the selected fragment is stripped down to a skeleton consisting only of the POS tags of each line, and words corresponding to the desired content are used to fill this skeleton in. This procedure is followed in [8, 1, 22].

Alternative procedures rely on building a stream of text from scratch, and resort to various techniques to ensure the continuity of the textual sequence. One early approach was to rely on linguistic grammars to drive the construction. This was the approach followed in [13, 14], where TAG grammars were employed. A more popular alternative is the use of n-grams to model the probability of certain words following on from others. This corresponds to reusing fragments of the corpus of size n, and combining them into larger fragments based on the probability of the resulting sequence. This is the main approach for ensuring text coherence used in [2, 11, 9, 7]. All these different computer poets rely on various additional methods for establishing constraints on the resulting poem drafts.

To ensure that resulting poems satisfy constraints on poem structure in terms of lines, systems that build a stream of text from scratch rely on either building each line separately [7] or applying a separate procedure for distributing the resulting text into poetic lines [11, 9].

2.3 The WASP System

The development described in this paper was carried out over an existing version of the WASP system [11, 9].

Combining ngram modelling and evolutionary approaches, the WASP poetry generator had been built using an evolutionary approach to model a poet's ability to iterate over a draft applying successive modifications in search of a best fit, and the ability to measure metric forms. It operates as a set of families of automatic experts: one family of content generators or *babblers* – which generate a flow of text that is taken as a starting point by the poets –, one family of *poets* – which try to convert flows of text into poems in given strophic forms –, one family of *judges* – which evaluate different aspects that are considered important –, and one family of *revisers* – which apply modifications to the drafts they receive, each one oriented to correct a type of problem, or to modify the draft in a specific way. These families work in a coordinated manner like a cooperative society of readers/critics/editors/writers. All together they generate a population of drafts over which they all operate, modifying it and pruning it in an evolutionary manner over a number of generations of drafts, until a final version, the best valued effort of the lot, is chosen. In this version, the overall style of the resulting poems is strongly determined by the accumulated sources used to train the content generators, which are mostly n-gram based. Several versions have been developed, covering poetry generation from different inspirational sources as different sets of training corpora are used: from a collection of classic Spanish poems [11] and a collection of news paper articles mined from the online edition of a Spanish daily newspaper [9]. Readers interested in a full description are referred to the relevant papers. However, two specific aspects of this implementation are relevant for the present paper. First, the various judges assign scores on specific parameters – on poem length, on verse length, on rhyme, on stress patterns of each line, on similarity to the sources, fitness against particular strophic forms... – and an overall score for each draft is obtained by combining all individual scores received by the draft. A specific judge is in charge of penalising instances of excessive similarity with the sources, which then get pushed down in the ranking and tend not to emerge as final solutions. Second, poets operate mainly by deciding on the introduction of line breaks over the text they receive as input.

3 Can a Computer Poet Undertake a Commission for a Set of Themed Poems?

The work reported in this paper arose in response to a request received by the author to provide a set of poems generated by the WASP poetry system to be included in a book chapter about computational creativity. The request explicitly indicated that these poems should never have been published anywhere else, to avoid possible problems with copyright. Additionally, the author decided that the poems should aim to achieve a certain thematic unity, somehow relating to the circumstances in which they were commissioned. Finally, the author wanted to include data on the curation coefficient

applicable, and to maximise its value to highlight the relative merit of the system itself in the achievement.

These conditions posed a challenge to the existing implementation of the WASP system. First, because the system as it stood had no means for driving the resulting poems towards particular themes. Second, because the procedures already in place for ensuring originality were inefficient. Third, because prior versions of the system had relied on low values of the curation coefficient: only a very small subset of actual system output was worthy of presentation to a wider audience.

The final set of poems was achieved by a recombination of some of the existing modules with new modules specifically designed for the ocasion, and by a new procedure for generating poems that abandoned the original generate and test approach underlying the evolutionary version of the system for a more informed generative approach that applied backtracking in search of solutions that better fulfilled the driving constraints.

3.1 Developing Text Babblers for the Themed Commission

As the book for which the poems were commissioned was to be published in Mexico, it was decided that the poems should have a Mexican theme. As the babbler modules rely on an ngram model of language to produce sequences of text that are word to word coherent, the overall style of the resulting poems is strongly determined by the accumulated sources used to train the content generators. For this initiative, a corpus of training texts was constructed by combining an anthology of poems by Mexican poets compiled from the Internet, and a set of news articles mined from the web pages of an online Mexican daily newspaper.

Earlier attempts to generate based on the simpler model trained only over the set of news items resulted in a candidate texts that were very difficult to adjust to any given poetic form. This related to the fact that the sequences of words contemplated in the ngram model resulting from news items only did not include enough combinations with a potential for poetic form. When the training set was expanded with an additional set of poetic texts, the resulting set of candidate texts showed a greater potential for composition into poetic forms.

This observation corroborates the intuition that the set of training texts used to train the ngram model imposes a certain overall style on the texts that can be produced. But it also raises the question of whether the desired poetic form is obtained at the price of replicating fragments of the poems being used as part of the inspiring set. This issue is addressed below.

3.2 Limitations of the Original Evolutionary Approach

The original WASP evolutionary system was designed to produce an initial large population of drafts - based on its ngram-based babbler modules -, to compose these into poem drafts by inserting line breaks at appropriate places - relying on its poet modules -, and to select as output a quality subset from those candidate drafts by applying the fitness functions implemented in its judge modules. This procedure was effective because it allowed the system to zoom in towards the regions of the overall conceptual space - as defined by the ngram model of language being used - that held potentially valuable text fragments from the point of view of poetic form - as defined by the fitness functions. This procedure was reasonable when the only constraint on the result was that it satisfy a certain poetic form. Specific poet modules and fitness functions would be designed for the particular poetic form, say, for a *cuarteto*, and the system would explore all the possible poems of this form arising from the given ngram model. This approach had two disadvantages for the present initiative: one related to form and one related to theme.

The existing solution was devised to drive the system towards poems of a particular type. When giving priority to theme, a certain flexibility in form could be introduced, allowing for poems with different poetic forms as long as they were consistent with the theme. To achieve this in terms of an evolutionary approach required the development of a confusing set of composition modules - capable of generating drafts in several poetic forms - and complex fitness functions - allowing for different fitness according to which particular poetic form was being considered. This lead to the consideration of alternative implementations.

The existing solution also had no obvious way of constraining results to particular themes. The word content of the results is constrained by the ngram model used, but an ngram model small enough to ensure that particular themes are present in the result would be too small to allow sufficient word recombinations to achieve valuable poetic forms. Additional elements could be added to the fitness function to rule out candidate drafts diverging from the desired themes, but this solution clashed with the decision above to consider alternative implementations.

3.3 Redeploying WASP Modules with a Different Purpose

A first attempt was carried out to simply redeploy the existing WASP modules - babblers, poets and judges - with the new purpose in mind. Under the new circumstances, judgements on candidate drafts could become more radical: if drafts were not related to the desired theme, they could be ruled outright. This had another consequence on the overall design: the reviser modules, which allowed exploration of the conceptual space by replacing certain words with others at random were seen to have little positive effect. Given the accumulated set of constraints on the results, random changes had a high probability of reducing fitness rather than improving it.

A formative evaluation was carried out over the existing prototype, configured so that a very large population of drafts was built, composed into a number of possible poetic forms, and evaluated using judges that combined fitness functions for theme, the various poetic forms considered, and originality. The revision modules were switched off for this test.

Fitness functions for theme relied on a set of input words to characterise the desired theme, penalising the drafts that did not include any of them, and reinforcing the drafts that did.

Fitness functions for poetic forms were already available as judge modules, and a simple combination of judges for different poetic forms was employed.

The fitness function for originality was addressed by developing a specific judge module that held the complete set of texts in the training set as a master file. Every line appearing in a candidate draft was searched for in the master file, and the candidate draft was rejected if the particular sequence of words in any of its lines appeared as a continuous unit anywere in the master file. This ensured that only lines that combined elements from different parts of the training set in innovative ways were considered by the system.

This approach generated a very large set of results but with very low average quality. This might have been acceptable if the set of results was mined for valuable drafts, but this would imply a very

low curation coefficient for the final set.

3.4 Revising the Constructive Procedure to Match the New Circumstances

It was clear from the experiment described above that at least two improvements were required to fulfill the goals we had set out to fulfil. One was to improve the fitness functions overall so that only results of a higher quality survived the evaluation stage. Another was to somehow improve the construction procedure itself so that better quality results were produced. The evolutionary paradigm of the original approach required mostly random procedures for generation and revision, with quality to be achieved by means of evolutionary operators combined with selection in terms of the fitness function. But this approach clashes with the fact that the conceptual space that we want to explore is constrained to the set of texts that can be derived from the ngram model under consideration. For the evolutionary operators to guarantee that mutation and cross over produce results that are still within the desired conceptual space, they would have to be restricted to operations that take into account the ngram model during mutation and/or cross over.

The option of refining the revisers by enriching them with knowledge so that the changes they introduced were more informed was seen as impractical, and it was preferred to overhaul completely the generation procedure so as to take advantage of the available information to only generate valuable results in the first place.

The revised version of the construction procedure expanded the initial solution for babblers, which was based on extending a candidate sequence of words with further words that have a non-zero probability of appearing after the last word of the sequence, according to the ngram model. In both versions, at each choice point, the system is faced with a number of possible continuations. In the earlier version, this choice was taken randomly. In the new version, the choice is made taking into account additional criteria, covering the following issues: relation to theme, plausibility of sentence ending, control over repetition of sentences already generated, and restriction to overall length of sentences.

The first criterion to consider involves the initial constraints on theme, giving preference to options related to the desired theme.

The second criterion is designed to rule out cases where a draft is ended at a point where the word sequence under consideration does not allow the ending of the sentence.

The third criterion aims to avoid having the system repeat itself. A model of short term memory for sentences has been added, so that continuations of sentence drafts that replicate sentences constructed recently are avoided.

The final criterion ensures that text candidates are restricted to single sentences, and the overall length is restricted by introducing a check on the accumulated length of the word sequence that starts giving priority to continuations that close off the sentence after a given threshold length has been achieved.

The set of judges is revised so that drafts in any one of the following situations are ruled out directly:

- candidate drafts with line lengths beyond 14
- candidates drafts that have lines of different lengths

Additional judges have been developed that reinforce drafts were a certain pattern of rhyme can be spotted:

The procedure for composing candidate texts into valid poetic forms is revised in the following way. For any given candidate text the poetic composition module:

- finds the set of line lengths that have a potential to give an exact break down of the total number of syllables in the text
- composes a number of candidate draft poems based on the input text, each one distributing the text into lines of the corresponding length as worked out above
- evaluates the resulting set of poem drafts
- returns only those that are positively evaluate in terms of the judges for metric form

The described adaptations result in an exploratory software that takes a long time to run - as it explores exhaustively the portions of the conceptual space established by the given ngram model that include words from the desired theme - and produces a much smaller set of candidate drafts. These candidate drafts are of high quality in terms of poetic form - they correspond to stanzas of lines of the same length in syllables - but are surprisingly short in length - they very rarely exceed two lines. This restriction on length is a result of the interplay between the configuration that limits texts to single sentences and the restriction that the system start trying to close sentence as soon as a minimally valid length has been reached. In spite of the fact that judges have been included to prioritise poem drafts that exhibit rhyme, the set of results very rarely does.

This set of results is not in itself a convincing set of poems with which to satisfy the received commission. But it constitutes a treasure trove of valuable material generated by the system: it is by construction innovative - in terms of p-creativity as described by Boden, given that the originality judges check each line against the master file built from the training corpus and rule out any replications - and it is remarkable in its poetic form - as guaranteed by the remaining judges. It is a small set, but large enough to allow a further step of recombination of these poem snippets with one another.

The construction procedure was therefore extended with a further stage that considered these poem drafts as possible ingredients to combine into larger poems. The heuristics considered to drive this recombination process were as follows:

- the set of poem snippets was classified into groups according to the length of their lines in syllables
- poem snippets of the same length of line were further grouped together into sets related by shared rhymes
- larger poem drafts were built by combining together the sets of snippets of the same line length that had shared rhymes

The initial set of small poem drafts was produced in 6 separate runs with the same configuration, designed to carry out 1000 attempts to build poem drafts fulfilling the constraints as described above. The data on number of valid poem drafts found in each of these runs is presented in Table 1. Runs 2, 3 and 6 had to be aborted without finishing for practical reasons unrelated to system operation.

Run #	Valid drafts found
1	149
2	46
3	106
4	150
5	8
6	10

Table 1. Rates of success in the runs for collecting an initial set of poem snippets.

The average rate of success over this limited set of data - excluding the data for aborted runs on the grounds that no record is available of

the number of attempts they had carried out before being stopped - is 13.5 %. Given the complexity of the conceptual space that is being searched, this rate is considered very acceptable.

The total number of snippets obtained in this way that was used as input for the procedure for composing larger poem drafts was 469.

The procedure for recombining the generated poem snippets into larger poem drafts produced 42 poem drafts, as described in Table 2. Overall these poems have used 18 different rhymes, irregularly spread over the set of resulting poems. The numbers provided for the complete set of poems do not correspond to the addition of the specific values for different line lengths because poem lengths and rhyme schemes are sometimes repeated for different line lengths.

Line lengths	Poems	Poem lengths	Rhyme schemes
6	1	1	1
7	9	6	7
8	6	3	6
9	10	7	10
10	14	8	14
12	2	1	-
All	42	11	31

Table 2. Description in terms of line lengths of the set of poem drafts obtained by recombination of snippets

The poems that resulted from this process were of different size, and for each particular poem size a rhyme schema results from the way in which snippets sharing rhyming lines have been combined. The analysis of the resulting set in terms of these emerging stanzas and rhyme schemes is presented in Tables 3 and 4.

Stanza size	Rhyme schemes
10	ABABCAABDA
14	BAABACADAEFBAA
	ACABABADAEAFCA
15	EACAFGBADADAHCA
20	AEACABABFAGAADADACAC
21	BACBAADAAAEAFAGHAAIJA

Table 4. Description of the longer stanzas in the set of poem drafts obtained by recombination of snippets

Of the 42 poems generated, 13 poems were deemed to be unusable as a result of problems in the generation process. The type of problems that were identified included issues of incorrect scanning of line lengths due to the appearance of punctuation signs not covered by the parsing procedures (2), undesirable repetition of subsets of lines (5), occurrence of unknown words (4), inclusion of unacceptable rude words (2).

The issue with incorrect scanning of line lengths has now been corrected.

Repetition of fragments of poems of more than one line is discouraged. The ones appearing in the result set have been tracked down to a small bug in the recombination process that should be easy to fix.

Some of the unknown words appear because the corpus of news items is mined directly from the web and the pre-processing procedures applied to clean up the html code sometimes miss non-words that end up in the training set. Improvements on the clean up procedure already under way should avoid this problem in the future.

Another source of problematic words is the use of foreign languages proper names, also frequent in news items. These words are acceptable in terms of their semantics contribution but their spelling confuses the metric analysis module of the system, which computes an incorrect number of syllables for them. This in its turn affects the composition processes that convert the resulting text into poetic form.

Rude words seem to have been used in some of the news items in the corpus, or possibly in some of the poems. But they are not considered desirable for the commissioned set of poems.

Of the remaining 29 poems, 7 were selected to be included in the book chapter that gave rise to the commission. This selection was based on general quality, but also on how well the selected poems fitted the desired theme. The 22 poems that were not selected show acceptable quality, and they were excluded from the selection for one of the following reasons:

- they shared some lines with the poems already selected
- their relation to the desired theme was not clear
- they included mentions of entities too specific to Mexican current news to be easily identified by a general public
- they included proper names of individuals featuring in the Mexican news
- they were overlong

Example results of the poems produced in this way are presented in Tables 6, 7,8,9,10 and11. These examples correspond to a second stage of selection out of the 22 poems that had not been chosen for inclusion in the set of poems commissioned for the book chapter.

The poems presented in Tables 5, 6, 7 and 8 correspond to four-line poems of different number of syllables per line (7, 7, 8 and 9 respectively), and showing different rhyme schemes (BACA, ABCA, ABAC). Together they illustrate the ability of the system to find the most metrically appropriate form for presenting a given text, using different lengths of line in syllables as required. They also illustrate the ability of the system to operate with different rhyme schemes to make the most of a given text.

Toda era una ave larga	Everything was a long bird
que cuando se conforman.	that when they conform.
Admitió que se tienen	He admitted that they have
registradas personas.	registered persons.

Table 5. Example of a poem of 4 lines of 7 syllables with rhyme scheme BACA, with an approximate English translation.

Muestra también. Esta	Shows as well. This
noche adonde yo soy.	night were I am.
Subraya que para ellos	Underlines that for them
ya no salgas. Estrellas.	come out no more. Stars.

Table 6. Example of a poem of 4 lines of 7 syllables with rhyme scheme ABCA, with an approximate English translation.

Aspecto que se encontraban	Aspect that they were finding
ejemplares. Nuevamente.	exemplars. Again.
Señalaron que no haya	They pointed out that there should not be
más daños como los niños.	more harm like the children.

Table 7. Example of a poem of 4 lines of 8 syllables with rhyme scheme ABAC, with an approximate English translation.

Stanza size	4	5	6	7	8	9
Rhyme schemes	ABBA	BACDA	BAAAAB	CBADABA	BACADAAE	ABABCBBDB
	ABAB	ABCDA	ABABAB	BACDAAE		ABCDEAFBA
	AAAA		ABCABC			ABACDEAFA
	BACA		ABABCA			BACADAEAF
	ABAC		ABABAC			ABACDAAEF
	ABCA		BAACAA			
			ABCADA			
			ABCAAD			
			ABACAD			

Table 3. Description in terms of the shorter stanzas in the set of poem drafts obtained by recombination of snippets

Zona militar. Qué delicia	Military zone. What a delight
delgada incomprensible. Amiga.	thin incomprehensible. Girl friend.
Agueda era luto pupilas	Agueda was mourning green
verdes. Sobrepasa. Guerrero.	pupils. Overshoots. Warrior.

Table 8. Example of a poem of 4 lines of 9 syllables with rhyme scheme ABAC, with an approximate English translation.

The poem presented in Table 8 is made of 4 *eneasílabos* of 9 syllable lines. Lines 2 and 3 share an asonant rhyme in *i-a*. The restriction on early closure of sentences has produced here a certain *staccato* feeling that is in line with the topic being addressed. Serendipity has led to a marked contrast between "military" and "delight", followed up with a surprisingly appropriate "incomprehensible". In spite of the choppy phrasing, as "girl friend" and "delight" agree in gender in Spanish, there is an implicit thread to the first two lines that is quite evocative. The third line mentions the female proper name "Agueda", rounding up this impression. This is again serendipitous. But it poses the question of whether similar criteria might not be used to derive selection heuristics so that future versions of the system can attempt to achieve similar effects. The final word "warrior" is ambiguous, and may originally have been intended as a reference to the Mexican state of Guerrero, but also links up with the military theme.

Juegan el largo recorrido	They play the long tour
desde su muerte ya no salgas.	from his death come out no more.
Séptimo. Cordero tranquilo	Seventh. Peaceful lamb
cordero que paces tu grama.	lamb that grazes its grass.
Silencios. Cordero tranquilo	Silences. Peaceful lamb
cordero que paces tu grama.	lamb that grazes its grass.

Table 9. Example of a poem of 6 lines of 9 syllables with rhyme scheme ABABAB, with an approximate English translation.

The poem presented in Table 9 is composed of 6 *eneasílabos* of 9 syllable lines. Lines 1, 3 and 5 rhyme together, and so do lines 2, 4 and 6. The rhyming is poor because it basically involves some the line endings being repeated twice. However, this arises from a parallelism trope - same linguistic structure used repeatedly with slight variations of content - and this makes the repeated rhyme somewhat more acceptable. The repetition is serendipitous and arises from the fact that particular sequences of words that match well a given poetic form tend to be reused to fill in certain stanzas ("Cordero tranquilo // cordero que paces tu grama. "), relying on different fragments of similar length to cover the initial first few syllables ("Séptimo.", Silencios."). Remember the constituent snippets were originally built separately, and they are only combined by application of the described composition heuristics. The apparent rhetorical effect is a consequence of the interaction between the limitation in the constructive procedure for poem snippets and the composition heuristics. Having noticed this interaction, we hope to include it as a system feature in future releases. In this particular case, the sequence in which the different fragments appear also achieves a significant effect, with the neighbouring mention of "death" and "lamb" evoking a certain hint of Christian symbolism. The effect of the early closing policy for sentences is also apparent in this poem.

Engalanados por los derechos	Garlanded by the rights
del niño indígena. Apago soles.	of the indigenous child. I switch off suns.
Concluido el objetivo que exista	Having achieved the goal that it exists
todo el mes para que ya sin nombre.	the whole month so that now nameless.
Dichosa puerta que nos transforman.	Happy gate that they transform for us.
Solidaridad vocación. Hombres.	Solidarity vocation. Men.
Acción nacional tiene un enorme	National action has an enormous
pez que se ilumina. Guatemala.	fish that lights up. Guatemala.

Table 10. Example of a poem of 8 lines of 10 syllables with rhyme scheme BACADAAE, with an approximate English translation.

The poem presented in Table 10 is composed of 8 *decasílabos* of 10 syllable lines. Lines 2, 4, 6 and 7 rhyme together. It presents interesting features that arise from the fact that sentences in the news items corpus are not generally well suited for partition over several valid metric lines, which lead to them being cut off abruptly at points where the closure makes syntactic sense. The texts in the poetic part of the corpus perform better in this sense, possibly as a result of being composed with metric form in mind. Drafts where the system alternates fragments from the two different parts of the corpus tend to achieve greater sentence lengths, as well as interesting contrasts between day to day pragmatic topics arising from news items and grander and more abstract topics obtained from the poems in the corpus. The Mexican theme is hinted at by the mention of the indigenous child.

The poem presented in Table 11 is included as an example of a longer poem. It has 15 *heptasílabos* or 7 syllable lines. Lines 2, 3, 5, 7, 9, 13, and 14 share asonant rhyme in *a-o* and lines 4 and 12 share asonant rhyme in *a-a*. This results in a rhyme scheme of the form CAABADAEAFGBAA. The Mexican theme is apparent in the mentions of citizens of two different Mexican states ("michoacanas", women from Michoacán; and "Queretanos", men from the town of Querétaro).

Tus ojos. Vinos tintos	Your eyes. Red wines
blancos rosados. Nardo.	whites rosés. Tube rose
Amo tus ríos claros.	I love your clear rivers.
Tal vez esta medalla.	Maybe this medal.
Antes que este hachazo	Before this axblow
nos sacude. Imagínate.	shakes us. Imagine.
Séptimo. Pinté el tallo	Seventh. I painted the stem
luego el cáliz después.	then the calyx afterwards.
Ganar. Solicitamos.	Winning. We request.
Sólo soy un prisionero.	I am only a prisoner.
Admitió que se estima	He admitted that it is estimated
que mil michoacanas	that a thousand Michoacans
acudan. Queretanos.	turn up. Queretans.
Valor. Acompañado	Valour. Accompanied
por Margarita Flores.	by Margarita Flores.

Table 11. Example of a poem of 15 lines of 9 syllables with rhyme scheme CAABADAEAFGBAA, with an approximate English translation.

4 Discussion

Over the complete run, 29 out of 42 poems were considered acceptable, and 13 of those have been submitted for publication in different media. The remaining 16 poems are less impressive but acceptable overall - they are not included here for lack of space -, although they do have the disadvantage of sharing some lines with the preferred poems. This, however, should not be considered as a demerit of the poems themselves. Instead, it should be thought of as an issue of incompatibility between possible system outputs in terms of originality. Once one particular line has been included in a poem submitted to the public, the system should refrain from including it in further output. This issue had already been described in [10], and more attention should be paid to it in poetry generators in the future.

These numbers lead to a curation coefficient for the described system run of around 69 %. It is important to note that higher curation coefficients are desirable. This is contrary to intuition, which suggests that high values of the curation coefficient imply a need for significant mediation between system output and publishable results. The contrary is in fact the case, as a high curation coefficient implies that a large percentage of system output can be passed to the public directly.

An interesting feature of the system described in this paper is that instead of establishing as configuration parameters values for features such as number of lines, number of syllables per line, or rhyme scheme to use, it relies on an exploratory procedure that allows the system to find optimal values for these features depending on the text that it has to convey. This leads to the variety of line and poem lengths, and the broad range of rhyme schemes that appear in the result set.

This variation in the range of rhyme schemes might be presented as an argument in favour of the perceived creativity of the system. The exploratory procedure in place relies on a fitness function that assigns higher value to poems that exhibit rhyming lines, but it does not prescribe any particular patterns for the rhymes. This results in output that satisfies rhyme schemes not traditionally used by human poets. This could be interpreted as a shortcoming, but it can also be considered as a creative feature.

The reliance on a corpus of training texts to produce candidate texts to compose into poetry introduces a number of dependencies between the particular training set chosen and the range of output text that can be generated. In the examples above this has been shown to lead to poems satisfying certain thematic constraints, not necessarily arising from explicit theme related constraints but simply as a result of having constrained the corpus to text somehow related to the theme. The issue of explicit constraining on theme needs to be explored further.

The influence of the training corpus has also been shown to affect the plasticity of the resulting texts when trying to compose them into poetic metrical forms. Certain styles of prose, such as that used in news items, are less conducive to composing into metrically acceptable forms than those custom-composed for such a form of expression. This should be taken into account when building training corpora for this type of system. On the other hand, the combination of corpus elements coming from different domains can lead to interesting contrasts that may result in a perception of originality in the final results.

One interesting point is the role of punctuation. As a result of the way the ngram models are constructed, most punctuation sign are stripped away from the texts before training. Question and exclamation marks are left in because they impact the syntax of the sentences they appear in. The output candidate texts are therefore generally devoid of punctuation. This introduces a degree of freedom that provides some leeway for human readers to find possible valid interpretation of the resulting poems. Readers should consider the possibility of revising the poems to consider whether simple punctuation, like the insertion of commas or semi-colons at certain points might improve them. It is after all, a task that editors of poetry sometimes do take out of the hands of their poets, even when they are human. In any case, having noticed the possible significance of this issue, the development of a system module to address such a refinement task is being considered as future work.

A final question to be considered is that of the originality of the output set in contrast to the inspiring set, here understood to correspond to the training corpus of texts. This question features prominently in Ritchie's set of criteria for evaluating the output of creative systems [20]. The system presented in this paper includes by construction a filter on candidate poem drafts that rejects them if they include a line that can be found as a continuous sequence anywhere within the training set. This should ensure that no line in any of the resulting poems correspond to lines in the poems in the training set, and it should also reduce significantly the chance that sentences in the training corpus are replicated verbatim.

Although many of the points outlined above deal with features that are specific to poetry, some of them can clearly be considered as valuable insight for computational creativity beyond poetry generation. First, the idea that creative systems should evolve towards versions where the role of a human observer curating a a subset of system outputs as valid for publication is reduced to a minimum. Second, the need to consider not only originality with respect to the inspiring set but also with respect to other elements in the result set. Third, the observation that, once the desired target is sufficiently specified, the introduction of randomness in the constructive procedure can have a negative impact. Fourth, that tightening the constraints on the desired target is likely to lead to increases in the time taken to produce results, and to decreases in the amount of results produced. However, the results obtained in this way are more likely to be of high quality. This point can be related to the stated view of Douglas Hofstadter that constraints are crucial to creativity.

5 Conclusions

The evolutionary solutions attempted in the past for poetry generation in the WASP system worked very well for the unconstrained exploration of broad conceptual spaces, where all parts of the space

from a thematic point of view where equally valid as solutions, and constraints could be specified only in terms of metrical form. When constraints on theme are taken into consideration, it pays to relax the constraints on form, so that the system may look for the optimal poetical form covering a given theme. This has lead to the development of an exploratory procedure that sets its own values at run time for features such as poem length, line length, and rhyme scheme.

The refinement of the procedure for generating sentences to certain types of candidate - sentences of acceptable length and that can be understood as acceptably closed - had the consequence of restricting the possible outputs of the initial poem composition procedure to very short poem drafts. To compensate, a second stage of poem draft recombination has been added that builds larger poems from the set of initial candidate drafts. This recombination procedure is based on line length and shared rhymes, which leads to a result set that emulates reasonably well the composition of poems in terms of stanzas shaped together by rhyme.

The ratio of acceptable system outputs over total system outputs is reported, and it is argued to result in a very positive value of the curation coefficient.

The analysis of system outputs has lead to the identification of a number of positive features that have been included by serendipity, but which hold a very high potential for inclusion in future releases of the described system as quality-enhancing improvements. To handle these features might require an elaboration of the construction procedure as an interaction between a number of cooperating experts, in the way described in [15].

ACKNOWLEDGEMENTS

This paper has been partially supported by the projects ConCreTe 611733 and PROSECCO 600653 funded by the European Commission, Framework Program 7, the ICT theme, and the Future and Emerging Technologies FET program.

REFERENCES

[1] Manex Agirrezabal, Bertol Arrieta, Mans Hulden, and Aitzol Astigarraga, 'Pos-tag based poetry generation with wordnet', in *Workshop on Natural Language Generation (ACL 2013)*, (2013).

[2] Gabriele Barbieri, Franois Pachet, Pierre Roy, and Mirko Degli Esposti, 'Markov constraints for generating lyrics with style.', In Raedt et al. [18], pp. 115–120.

[3] M. Boden, *Creative Mind: Myths and Mechanisms*, Weidenfeld & Nicholson, London, 1990.

[4] John Charnley, Simon Colton, and Maria Teresa Llano, 'The flowr framework: Automated flowchart construction, optimisation and alteration for creative systems', in *5th International Conference on Computational Creativity, ICCC 2014*, Ljubljana, Slovenia, (06/2014 2014).

[5] Simon Colton, Jacob Goodwin, and Tony Veale, 'Full-FACE poetry generation', in *Proceedings of the International Conference on Computational Creativity 2012*, pp. 95–102, (2012).

[6] Simon Colton and Geraint A. Wiggins, 'Computational creativity: The final frontier?', In Raedt et al. [18], pp. 21–26.

[7] Amitava Das and Björn Gambäck, 'Poetic machine: Computational creativity for automatic poetry generation in bengali', in *5th International Conference on Computational Creativity, ICCC 2014*, Ljubljana, Slovenia, (06/2014 2014).

[8] P. Gervás, 'WASP: Evaluation of different strategies for the automatic generation of spanish verse', in *Proceedings of the AISB-00 Symposium on Creative & Cultural Aspects of AI*, pp. 93–100, (2000).

[9] P. Gervás, 'Evolutionary elaboration of daily news as a poetic stanza', in *Proceedings of the IX Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados - MAEB 2013*, (2013).

[10] Pablo Gervás, 'Dynamic inspiring sets for sustained novelty in poetry generation', in *Second International Conference on Computational Creativity*, México City, México, (04/2011 2011).

[11] Pablo Gervás, 'Computational modelling of poetry generation', in *Proceedings of the AISB13 Symposium on Artificial Intelligence and Poetry*, (2013).

[12] Hugo Gonçalo Oliveira, 'PoeTryMe: a versatile platform for poetry generation', in *Proceedings of the ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence*, C3GI 2012, Montpellier, France, (August 2012).

[13] H. M. Manurung, 'Chart generation of rhythm-patterned text', in *Proc. of the First International Workshop on Literature in Cognition and Computers*, (1999).

[14] H. M. Manurung, *An evolutionary algorithm approach to poetry generation*, Ph.D. dissertation, University of Edimburgh, Edimburgh, UK, 2003.

[15] Joanna Misztal and Bipin Indurkhya, 'Poetry generation system with an emotional personality', in *5th International Conference on Computational Creativity, ICCC 2014*, Ljubljana, Slovenia, (06/2014 2014).

[16] Oulipo, *Atlas de littérature potentielle*, number vol. 1 in Collection Idées, Gallimard, 1981.

[17] R. Queneau, *100.000.000.000.000 de poèmes*, Gallimard Series, Schoenhof's Foreign Books, Incorporated, 1961.

[18] Luc De Raedt, Christian Bessière, Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz, and Peter J. F. Lucas, eds. *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*, volume 242 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2012.

[19] Fam Rashel and Ruli Manurung, 'Pemuisi: A constraint satisfaction-based generator of topical indonesian poetry', in *5th International Conference on Computational Creativity, ICCC 2014*, Ljubljana, Slovenia, (06/2014 2014).

[20] G. Ritchie, 'Some empirical criteria for attributing creativity to a computer program', *Minds & Machines*, **17**, 67–99, (2007).

[21] Jukka M. Toivanen, Oskar Gross, and Hannu Toivonen, 'The officer is taller than you, who race yourself! using document specific word associations in poetry generation', in *5th International Conference on Computational Creativity, ICCC 2014*, Ljubljana, Slovenia, (06/2014 2014).

[22] Jukka M. Toivanen, Matti Järvisalo, and Hannu Toivonen, 'Harnessing constraint programming for poetry composition', in *Proceedings of the International Conference on Computational Creativity 2013*, pp. 160–167, (2013).

[23] Jukka M. Toivanen, Hannu Toivonen, Alessandro Valitutti, and Oskar Gross, 'Corpus-based generation of content and form in poetry', in *Proceedings of the International Conference on Computational Creativity 2012*, pp. 175–179, (2012).

[24] Tony Veale, 'Less rhyme, more reason: Knowledge-based poetry generation with feeling, insight and wit', in *Proceedings of the International Conference on Computational Creativity 2013*, pp. 152–159, (2013).

[25] G. Wiggins, 'Searching for Computational Creativity', *New Generation Computing, Computational Paradigms and Computational Intelligence. Special Issue: Computational Creativity*, **24**(3), 209–222, (2006).

An Informational Model for Cellular Automata Aesthetic Measure

Mohammad Ali Javaheri Javid ✉ [1] and **Mohammad Majid al-Rifaie** [2] and **Robert Zimmer** [3]

Abstract. This paper addresses aesthetic problem in cellular automata, taking a quantitative approach for aesthetic evaluation. Although the Shannon's entropy is dominant in computational methods of aesthetics, it fails to discriminate accurately structurally different patterns in two-dimensions. We have adapted an informational measure to overcome the shortcomings of entropic measure by using information gain measure. This measure is customised to robustly quantify the complexity of multi-state cellular automata patterns. Experiments are set up with different initial configurations in a two-dimensional multi-state cellular whose corresponding structural measures at global level are analysed. Preliminary outcomes on the resulting automata are promising, as they suggest the possibility of predicting the structural characteristics, symmetry and orientation of cellular automata generated patterns.

1 INTRODUCTION

Cellular Automata (CA) initially invented by von Neumann in the late 1940s as material independent systems to investigate the possibility self-reproduction. His initial cellular automaton to study the possibility of self-reproduction was a two-dimensional (2D) cellular automaton with 29 states and 5-cell neighbourhood. A cellular automaton consists of a lattice of uniformly arranged finite state automata each of which taking input from the neighbouring automata; they in turn compute their next states by utilising a state transition function. A synchronous or asynchronous interactive application of state transition function (also known as a *rule*) over the states of automata (also referred to as *cells*) generates the global behaviour of a cellular automaton.

The formation of complex patterns from simple rules sometimes with high aesthetic quality has been contributed to the creation of many digital art works since the 1960s. The most notable works are "*Pixillation*", one of the early computer generated animations [32], the digital art works of Peter Struycken [31, 36], Paul Brown [5, 12] and evolutionary architecture of John Frazer [18]. Although classical one-dimensional CA with binary states can generate complex behaviours, experiments with 2D multi-state CA have shown that adding more states significantly increases the complexity of behaviour, therefore, generating very complex symmetrical patterns with high aesthetic qualities [21, 22]. These observations have led to the quest of developing a quantitative model to evaluate the aesthetic quality of multi-state CA patterns.

This work follows Birkhoff's tradition in studying mathematical bases of aesthetics, especially the association of aesthetic judgement with the degree of complexity of a stimulus. Shannon's information theory provided an objective measure of complexity. It led to emergence of various informational theories of aesthetics. However due to its nature, the entropic measure fails to take into account spacial characteristics of 2D patterns which is fundamental in addressing aesthetic problem for CA generated patterns.

[1] Goldsmiths, University of London, email: m.javaheri@gold.ac.uk
[2] Goldsmiths, University of London, email: m.majid@gold.ac.uk
[3] Goldsmiths, University of London, email: r.zimmer@gold.ac.uk

2 CELLULAR AUTOMATA ART

The property of CA that makes them particularly interesting to digital artists is their ability to produce interesting and logically deep patterns on the basis of very simply stated preconditions. Iterating the steps of a CA computation can produce fabulously rich output. The significance of CA approach in producing digital art was outlined by Wolfram in his classical studies on CA behaviours in [39]. Traditional scientific intuition, and early computer art, might lead one to assume that simple programs would always produce pictures too simple and rigid to be of artistic interest. But extrapolating from Wolfram's work on CA, "it becomes clear that even a program that may have extremely simple rules will often be able to generate pictures that have striking aesthetic qualities-sometimes reminiscent of nature, but often unlike anything ever seen before" [39, p.11].

Knowlton developed "*Explor*" system for generating 2D patterns, designs and pictures from explicitly provided 2D patterns, local operations and randomness. It aimed not only to provide the computer novice with graphic output; but also a vehicle for depicting results of simulations in natural (i.e. crystal growth) and hypothetical (e.g. cellular automata) situations, and for the production of a wide variety of designs [23]. Together with Schwartz and using *Explor*'s CA models, they generated "*Pixillation*", one of the early computer generated animations [32]. They contested in the *Eighth Annual Computer Art Contest* in 1970 with two entries, "*Tapestry I*" and "*Tapestry II*" (two frames from *Pixillation*). The "*Tapestry I*" won the first prize for "*new, creative use of the computer as an artist's tool*" as noted by selecting committee and covered the front page of *Computers & Automation* on Aug. 1970.

Meertens and Geurts also submitted an entry to the *Eighth Annual Computer Art Contest* with "*Crystalization*" as an experimental computer graphics generated by a asynchronous cellular automaton. Their entries were four drawings intended to generated patterns that combine regularity and irregularity in a natural way [20]. Peter Struycken, the Dutch contemporary digital artist has created many of his works "*Computer Structures*" (1969), "*Four Random Drawings for Lien and Ad*" (1972), "*Fields*" (1979-1980) with binary and multi-state CA [31, 36]. Paul Brown, the British contemporary digital artists also applied various CA rules in his static and kinematic computer arts. "*Neighbourhood Count*" (1991), "*Infinite Permutations V1*" (1993-94), "*Infinite Permutations V2*" (1994-95), "*Sand*

Lines" (1998), "*My Gasket*"(1998) "*Chromos*" (199-2000) [5, 12] are some of his CA generated works.

John F. Simon Jr created a series of art projects called "*Art Appliances*" using a CA based software and LCD panels to exhibit CA pattern formations. "*Every Icon*" (1996), "*ComplexCity*" (2000) and "*Automata Studies*" (2002) are examples of his art appliances. Driessen and Verstappen have produced "*Ima Traveler*" (1996) and "*Breed*"(1995-2007) digital arts in a three-dimensional CA space. Dorin's "*Meniscus*" [13] and McCormack's "*Eden*" [27] are further examples of interactive artworks built on the bases of CA rules. In addition, a combination of CA with other Alife techniques (e.g. evolutionary computing or L-systems) has been used to explore a set of rules generating patterns with aesthetic qualities [9, 34].

Fig. 1 shows some experimental patterns generated by the authors to demonstrate the generative capabilities of CA in creating appealing complex patterns from various initial configurations.

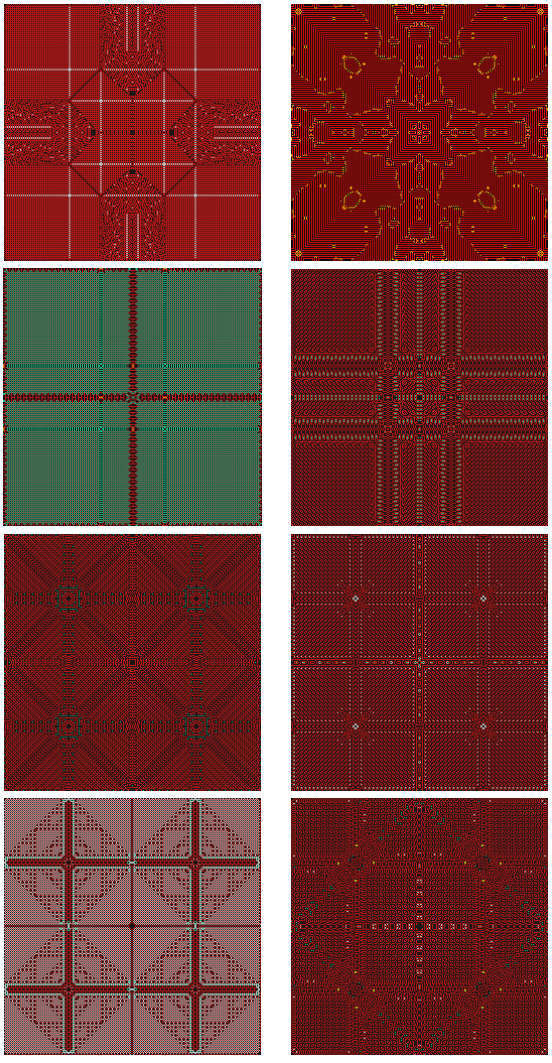


Figure 1. Sample 2D CA generated complex symmetrical patters

3 DEFINITION OF CELLULAR AUTOMATA

In this section, formal notions of 2D CA are explained and later referred to in the rest of the paper.

Definition 1: A cellular automaton is a regular tiling of a lattice with uniform deterministic finite state automata as a quadruple of $\mathcal{A} = \langle L, S, N, f \rangle$ such that:

1. L is an infinite regular *lattice* in $\mathbb{Z}$,
2. $S \subseteq \mathbb{N}^0$ is a finite set of integers as *states*,
3. $N \subseteq \mathbb{N}^+$ is a finite set of integers as *neighbourhood*,
4. $f : S^{|N|} \mapsto S$ is the *state transition function*.

The state transition function f maps from the set of neighbourhood states $S^{|N|}$ where $|N|$ is the cardinality of neighbourhood set, to the set of states $\{s_0, .., s_{n-1}\}$ synchronously in *discrete time* intervals of $t = \{0, 1, 2, 3, ..., n\}$ where t_0 is the *initial time* of a cellular automaton with *initial configuration*. A mapping that satisfies $f(s_0, ..., s_0) = s_0$ where $(s_0 \in S)$, is called a *quiescent state*.

In a 2D square lattice ($\mathbb{Z}^2$) if the opposite sides of the lattice (up and down with left and right) are connected, the resulting *finite* lattice forms a torus shape (Fig.2) which is referred as a lattice with *periodic boundary conditions*.

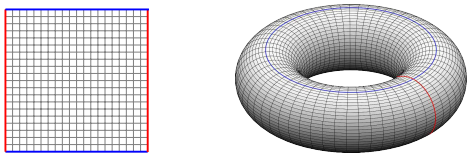


Figure 2. Connecting the opposite sides of a lattice forms a torus

The state of each cell at time $(t+1)$ is determined by the states of immediate surrounding neighbouring cells (nearest neighbourhood) at time (t) given a neighbourhood template. There are two commonly used neighbourhood templates considered for 2D CA. A five-cell mapping $f : S^5 \mapsto S$ known as *von Neumann neighbourhood* (Eq. 1) and a nine-cell mapping $f : S^9 \mapsto S$ known as *Moor neighbourhood* (Eq. 2).

$$s_{i,j}^{t+1} = f\begin{pmatrix} & s_{(i,j+1)}^{t} & \\ s_{(i-1,j)}^{t} & s_{(i,j)}^{t} & s_{(i+1,j)}^{t} \\ & s_{(i,j-1)}^{t} & \end{pmatrix} \tag{1}$$

$$s_{i,j}^{t+1} = f\begin{pmatrix} s_{(i-1,j+1)}^{t} & s_{(i,j+1)}^{t} & s_{(i+1,j+1)}^{t} \\ s_{(i-1,j)}^{t} & s_{(i,j)}^{t} & s_{(i+1,j)}^{t} \\ s_{(i-1,j-1)}^{t} & s_{(i,j-1)}^{t} & s_{(i+1,j-1)}^{t} \end{pmatrix} \tag{2}$$

Since the elements of the S are non-negative integers and discrete instances of time are considered, the resulting cellular automaton is a *discrete time-space* cellular automaton. These type of CA can be considered as *discrete dynamical systems*.

4 INFORMATIONAL AESTHETICS

The topic of determining aesthetics or aesthetic measures have been a heated debate for centuries. There is a great variety of computational approaches to aesthetics in visual and auditory forms including mathematical, communicative, structural, psychological and neuroscience. A thorough examination of these methodologies from different perspective has been provided in [19]. In this section, some informational aesthetic measures are presented. Our review is focused on informational theories of aesthetics as these are the ones that conform with this work directly.

Birkhoff suggested an early aesthetic measure by arguing that the measure of aesthetic (M) is in direct relation with the degree of *order* (O) and in reverse relation with the *complexity* (C) of an object [11]. Given that order and complexity are measurable parameters the aesthetic measure of (M) is:

$$M = \frac{O}{C} \tag{3}$$

Even though the validity of Birkhoff's approach to the relationship and definition of order and complexity has been challenged [38, 15, 16, 14], the notion of *complexity* and objective methods to quantify it remains a prominent parameter in aesthetic evaluation functions.

Shannon's introduction of *information theory* provided a mathematical model to measure the degree of uncertainty (entropy) associated with a random variable [33]. The entropy H of a discrete random variable X is a measure of the average amount of uncertainty associated with the value of X. So $H(X)$ as the entropy of X is:

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log_2 P(x) \tag{4}$$

The definition of entropy for X has a logarithm in the base of 2 so the unit of measure of entropy is in $bits$.

Moles [28], Bense [7, 6, 8] and Arnheim [2, 3, 4] were pioneers of the application of Shannon's entropy to quantify order and complexity in Birkhoff's formula by adapting statistical measure of information in aesthetic objects. Berlyne used informational approach in his psychological experiments to determine humans perceptual curiosity of visual figures [10]. Bense argued that aesthetic objects are "vehicles of aesthetical information" where statistical information can quantify the aesthetical information of objects [7]. For Bense order is a process of artistic selection of elements from a determined repertoire of elements. The aesthetic measure (M_B) is a the relative redundancy (R) of the reduction of uncertainty because of selecting elements from a repertoire $(H_{max} - H)$ to the absolute redundancy (H_{Max}).

$$M_B = \frac{R}{H_{max}} = \frac{H_{max} - H}{H_{max}} \tag{5}$$

where H quantifies entropy of the selection process from a determined repertoire of elements in $bits$ and H_{max} is maximum entropy of predefined repertoire of elements [8]. His informational aesthetics has three basic assumptions. (1) Objects are material carriers of aesthetic state, and such aesthetic states are independent of subjective observers. (2) A particular kind of information is conveyed by the aesthetic state of the object (or process) as *aesthetic information* and (3) objective measure of aesthetic objects is in relation with degree of order and complexity in an object [29].

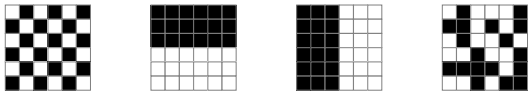
Herbert Franke put forward an *aesthetic perception* theory on the ground of *cybernetic aesthetics*. He made a distinction between the amount of information being stored and the rate of information flowing through a channel as *information flow* measured in $bits/sec$ [17]. His theory is based on psychological experiments which suggested that conscious working memory can not take more than 16 $bits/sec$ of visual information. Then he argued that artists should provide a flow of information of about 16 $bits/sec$ for works of art to be perceived as beautiful and harmonious.

Staudek in his multi criteria approach (informational and structural) as *exact aesthetics* to Birkhoff's measure applied information flow I' by defining it as a measure assessing principal information transmission qualities in time. He used 16 $bits/sec$ reference as channel capacity $C_r = 16\ bits/sec$ and a time reference of 8 seconds ($t_r = 8s$) to argue that artefacts with $I > 128\ bits$ will not fit into the conscious working memory for absorbing the whole aesthetic message [35].

Adapting Bense's informational aesthetics to different approaches of the concepts of order and complexity in an image in [30], three measures based on Kolmogorov complexity [25], Shannon entropy (for RGB channels) and Zurek's physical entropy [40] were introduced. Then the measures were are applied to analyse aesthetic values of several paintings (Mondrian, Pollock, and van Gogh). Machado and Cardoso [26] proposed a model based on Birkhoff's approach as the ratio of *image complexity* to *processing complexity* by arguing that images with high visual complexity, are processed easily so they have highest aesthetic value.

5 INFORMATION GAIN MODEL

Despite the domination of entropic measures to aesthetic evaluation functions, it has a major shortcoming in terms of reflecting structural characteristics of 2D patterns. Examples in Fig.3 illustrate this shortcoming by showing the calculations of entropy for 2D patterns with the same density but different structural regularities and complexities. Fig.3a is a uniformly distributed patterns (a highly ordered pattern), Fig.3b and Fig.3c are patterns with identical structures but in vertical and horizontal orientations. Fig.3d is randomly arranged pattern (a random pattern). As it is evident from the comparison of the patterns and their corresponding entropy value, all of the patterns have the same entropy value. This clearly demonstrates that Shannon's entropy fails to differentiate structural differences among these patterns. In the case of measuring complexity of CA generated patterns especially with multi-state structures, it would be problematic if only entropy used as a measure of complexity for the purpose of aesthetic evaluation.



(a) $H = 1$ (b) $H = 1$ (c) $H = 1$ (d) $H = 1$

Figure 3. The measure of entropy H for structurally different patterns with the same density of 50%

In order to overcome this problem we have adapted *information gain* model introduced as a method of characterising the complexity of dynamical systems [37]. It has been applied to describe quantitatively the complexity of geometric ornaments and patterns arising in random sequential adsorption of discs on a plane [1]. The informa-

tion gain G, also known as Kullback-Leibler divergence [24], measures the amount of information required to select a discrete random variable X with state j if prior information about variable X is known at the state of i.

$$G_{x_{ij}} = -\log P_{(x_i|x_j)} \tag{6}$$

where $P_{(x_i|x_j)}$ the conditional probability of the discrete random variable x at state i given its state j. Then from Eq. 6 *mean information gain* $\overline{G}$ would be the average information gain from possible states $(i|j)$:

$$\overline{G} = \sum_{i,j} P(i,j)G_{ij} = -\sum_{i,j} P_{i,j} \log P(i|j) \tag{7}$$

where $P_{(i,j)}$ is the joint probability of the variable x at state i and variable x at state j. Considering Eq. 7, we can define a structural complexity measure for a multi-state 2D cellular automaton as follows:

Definition 2: *A structural complexity measure is the mean information gain of a cell having a heterogeneous neighbouring cell in a multi-state two-dimensional cellular automaton pattern.*

$$\overline{G} = -\sum_{i,j} P_{(i,j)} \log_2 P_{(i|j)} \tag{8}$$

where $P_{(i,j)}$ is the joint probability of a cell having the i state (colour) and the neighbouring cell has the state (colour) j in a given neighbouring cell. And $P_{(i|j)}$ is the conditional probability of the state (colour) i given that its neighbouring cell has state (colour) j in one of four directions of up, low, left or right. The quantity $\overline{G}$ measures average information gain about other elements of the structure (e.g. the state of the neighbouring cell in one of the four directions), when some properties of the structure are known (e.g. the state of a cell). It can be noted that the combined probabilities of $P_{i,j}$ and $P_{i|j}$ describe spatial correlations in a pattern so that $\overline{G}$ can detect inherent correlations of patterns. Considering neighbourhood templates of a 2D CA in Eq.1 and Eq. 2, following variations of $\overline{G}$ can be defined where for each cell in i state given its neighbouring cell in j state in any of directions.

$$\overline{G}_u = -\sum_{i,j_{(x,y+1)}} P_{(i,j_{(x,y+1)})} \log_2 P_{(i|j_{(x,y+1)})} \tag{9}$$

$$\overline{G}_d = -\sum_{i,j_{(x,y-1)}} P_{(i,j_{(x,y-1)})} \log_2 P_{(i|j_{(x,y-1)})} \tag{10}$$

$$\overline{G}_l = -\sum_{i,j_{(x-1,y)}} P_{(i,j_{(x-1,y)})} \log_2 P_{(i|j_{(x-1,y)})} \tag{11}$$

$$\overline{G}_r = -\sum_{i,j_{(x+1,y)}} P_{(i,j_{(x+1,y)})} \log_2 P_{(i|j_{(x+1,y)})} \tag{12}$$

The measure is applied to calculate structural complexly of sample patterns in Fig 4 to demonstrates the ability of $\overline{G}$ in discriminating structurally different 2D patterns. The calculations have been performed for each elements of the patterns having a heterogamous colour in one of the four directions from two possible colours.

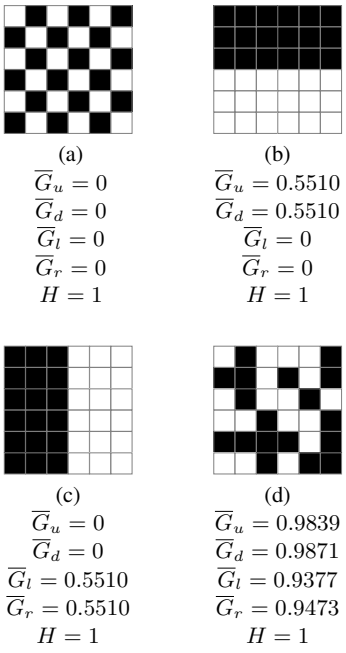


Figure 4. The comparison of entropy H and $\overline{G}$ for structurally different patterns but with the same density of 50%

6 EXPERIMENTS AND RESULTS

A set of experiments were designed to examine the effectiveness of $\overline{G}$ in discriminating structurally different patterns generated by multi-state 2D CA. The chosen experimental cellular automaton maps three states represented by *green*, *red* and *blue* colour cells. The quiescent state cells represented with *white* colours. The size of the lattice is set to 129×129 cells. Two set of experiments are conducted: (1) a single cell as initial configuration and (2) a randomly seeded initial configuration with 50% destiny of three states (*green*, *red*, *blue*). Both of the experiments are conducted for 300 successive time steps. The $\overline{G}$ for four directions along with their corresponding entropy H are measured in $bits$.

Fig. 7 and Fig. 8 illustrate the formation of 2D patterns for a sample of 12 time steps $\{0, 10, 20, 30, 40, 50, 60, 70, 80, 100, 200, 300\}$ starting from two different initial configurations and their corresponding $\overline{G}$ and H. Figs. 7 and 6 shows the plots of $\overline{G}$ and H for 300 time steps. The $\overline{G}$ measures in Fig. 7 which shows the formation of 2D patterns from a single cell are conforming in directional calculations; it means that each cell in the patterns has exactly the same amount of information regarding their neighbouring cell in one of the four directions. Therefore it indicates that the development of the patterns are symmetrical in the four directions. In other words, the cellular automaton with a single cell as its initial configuration has created 2D pattens with four-fold rotational symmetry. The measure in Fig. 8 starts with $\overline{G} \approx 1.7$ for the random initial configuration and with $H \approx 1.5$ (maximum entropy for a three-state patterns since $\log_2 3 = 1.5848$). The formation of patterns with local structures reduces the value of $\overline{G}$. The values of $\overline{G}$ are not conforming in any directional calculations which indicates the development of less ordered ("chaotic") patterns. From the comparison of H with $\overline{G}$ in the set of experiments, it is clear that it would be very unlikely to discriminate the structural differences of patterns with a single measure

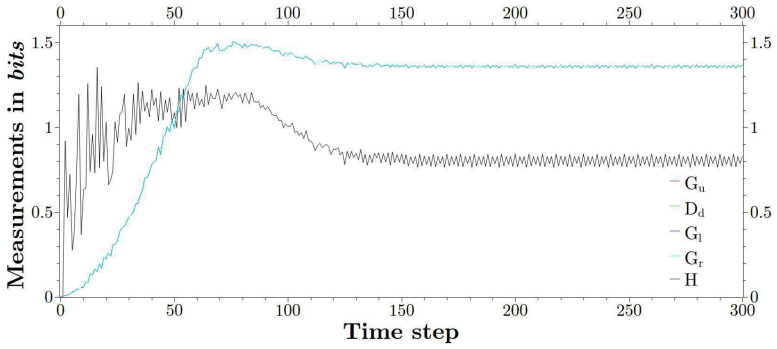


Figure 5. The plot of $\overline{G}$ and H for 300 time steps starting from single cell

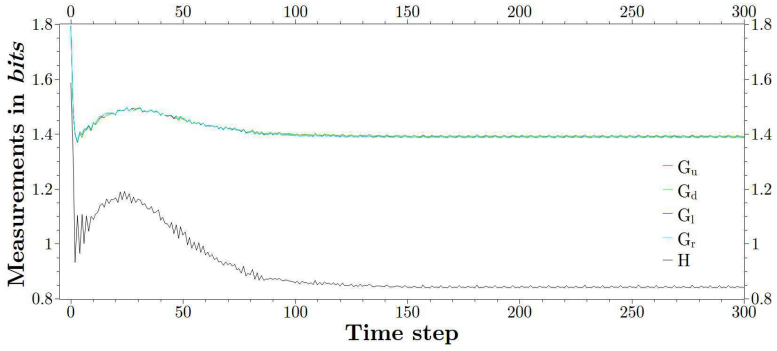


Figure 6. The plot of $\overline{G}$ and H for 300 time steps starting from random initial configuration

of H given the diversity of patterns that can be generated by various 2D CA state transition functions. Computing directional measures of $\overline{G}$ and comparing their values provides a more subtle measure of structural order and complexity of a 2D pattern. The conformity or non-conformity of $\overline{G}$ measure in up, down, left and right neighbouring cells clearly gives us not only an accurate measure of structural characteristics of 2D patterns but they also provide us with information about the orientation of the patterns as well.

7 CONCLUSION

Cellular automata (CA), which are fundamental to the study of self-replicating systems, are powerful tools in generating computer art. The multi-state 2D CA rule space is a vast set of possible rules which can generate interesting patterns with high aesthetic qualities. The application of CA in digital art has been reviewed; and the concepts of order and complexity from Shannon's information entropy perspective in the CA framework has been analysed concluding that existing informational aesthetic measures do not capture structural differences in 2D patterns. In order to address the shortcomings of informational approaches to computational aesthetics, a mean information gain model was adapted to measure both structural complexity and distinguish symmetrical orientation of 2D CA patterns. The measure takes into account conditional and joint probabilities of the information gain value that a cell offers, given a particular position of its neighbouring cells. The effectiveness of the measure is shown in a series of experiments for multi-state 2D patterns generated by a cellular automaton. The results of the experiments show that the mean information gain model is capable of distinguishing the structural complexity of 2D CA patterns as well as their symmetrical orientation. Having a model to evaluate the aesthetic qualities of CA generated patterns could potentially have a substantial contribution towards further automation of the evaluative component in the CA based computer generated art. This could also enable us to have an integrated process of generation-evaluation which is a subject of on going research.

8 ACKNOWLEDGEMENTS

We would like to offer our special thanks to anonymous reviewers for their constructive suggestions to improve the quality of this paper.

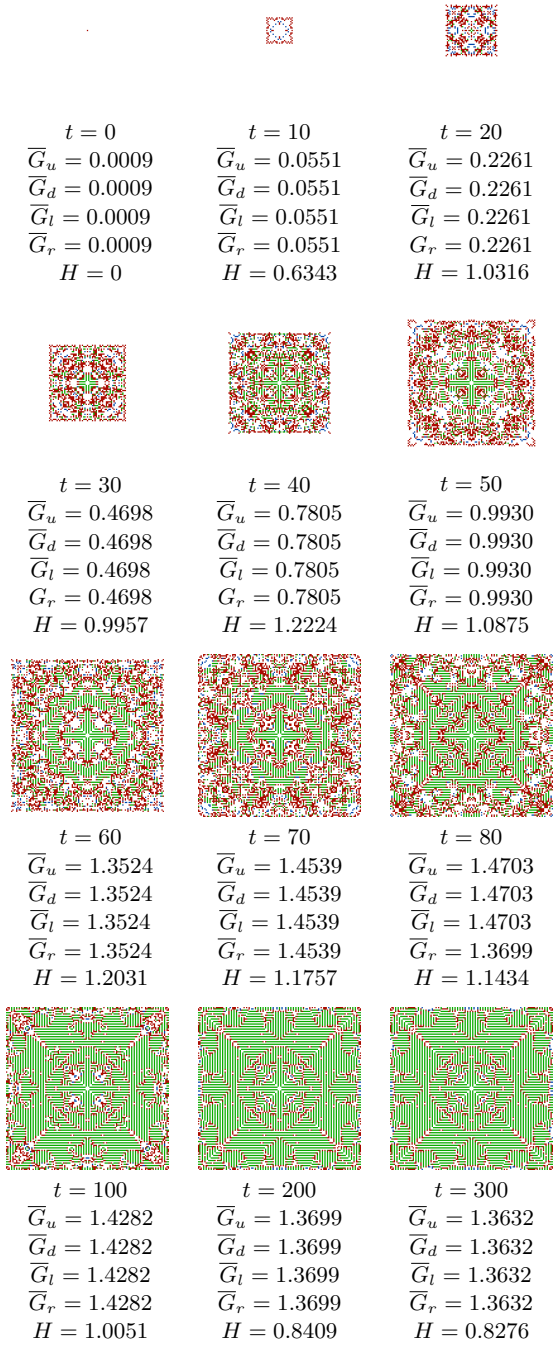


Figure 7. Patterns generated from a single cell as initial configuration and their corresponding $\overline{G}$ and H values

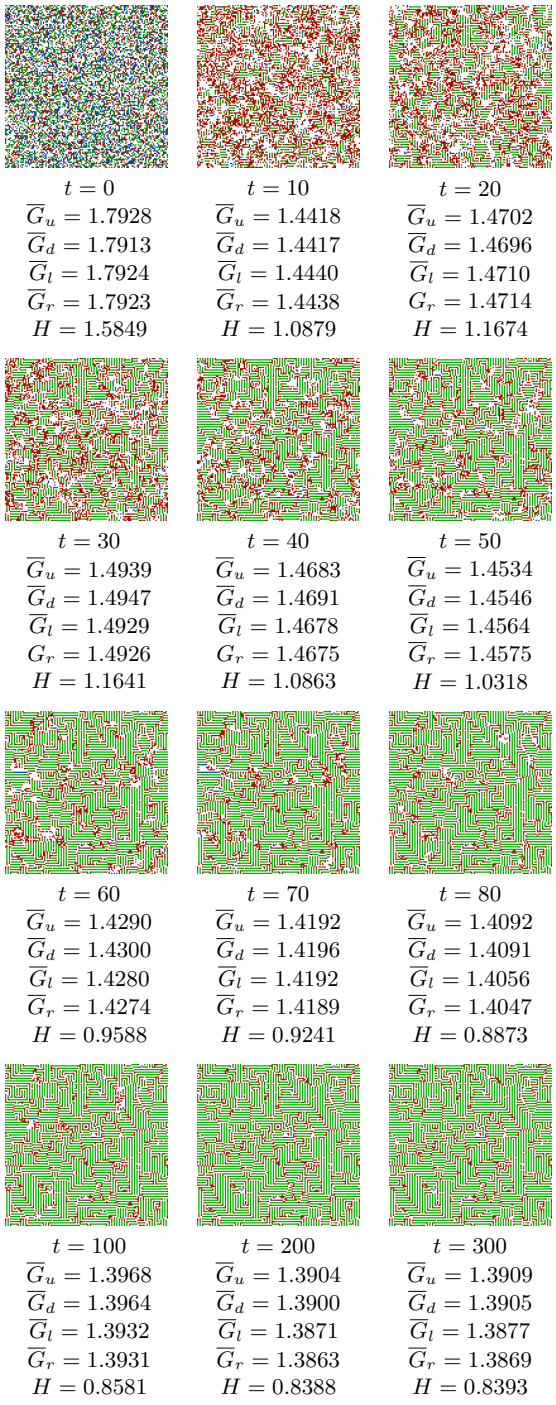


Figure 8. Patterns generated from a 50% seeded density as initial configuration and their corresponding $\overline{G}$ and H values

REFERENCES

[1] Andrienko, Yu. A., Brilliantov, N. V., and Kurths, J., 'Complexity of two-dimensional patterns', *Eur. Phys. J. B*, **15**(3), 539–546, (2000).

[2] Rudolf Arnheim, *Art and visual perception: A psychology of the creative eye*, Univ of California Press, 1954.

[3] Rudolf Arnheim, 'Towards a psychology of art/entropy and art an essay on disorder and order', *The Regents of the University of California*, (1966).

[4] Rudolf Arnheim, *Visual thinking*, Univ of California Press, 1969.

[5] Honor Beddard and Dodds Dodds, *Digital Pioneers*, V&A Pattern, V&A Publishing, 2009.

[6] M. Bense and G. Nee, 'Computer grafik', in *edition rot*, eds., Max Bense and Elisabeth Walther, volume 19, Walther, Stuttgart, (1965).

[7] Max Bense, *Aestetica: Programmierung des Schönen, allgemeine Textheorie und Textästhetik...*, Agis-Verlag, 1960.

[8] Max Bense, *Kleine abstrakte ästhetik*, E. Walther, 1969.

[9] Katie A Bentley, 'Exploring aesthetic pattern formation', in *Generative Art 2002 conference proceedings*, (2002).

[10] Daniel Ellis Berlyne, 'Conflict and information-theory variables as determinants of human perceptual curiosity.', *Journal of experimental psychology*, **53**(6), 399, (1957).

[11] G.D. Birkhoff, *Aesthetic Measure*, Harvard University Press, 1933.

[12] Paul Brown, 'Stepping stones in the mist', in *Creative evolutionary systems*, pp. 387–407. Morgan Kaufmann Publishers Inc., (2001).

[13] Alan Dorin, 'The virtual ecosystem as generative electronic art', in *Applications of Evolutionary Computing*, 467–476, Springer, (2004).

[14] Hans J Eysenck, 'An experimental study of aesthetic preference for polygonal figures', *The Journal of General Psychology*, **79**(1), 3–17, (1968).

[15] Hans Jürgen Eysenck, 'The empirical determination of an aesthetic formula.', *Psychological Review*, **48**(1), 83, (1941).

[16] Hans Jurgen Eysenck, 'The experimental study of the 'good gestalt' –a new approach.', *Psychological Review*, **49**(4), 344, (1942).

[17] Herbert W. Franke, 'A cybernetic approach to aesthetics', *Leonardo*, **10**(3), 203–206, (1977).

[18] John Frazer, *An evolutionary architecture*, Architectural Association Publications, Themes VII, 1995.

[19] Philip Galanter, 'Computational aesthetic evaluation: past and future', in *Computers and Creativity*, eds., Jon McCormack and Mark d'IInverno, pp. 255–293. Springer, (2012).

[20] Leo Geurts and Lambert Meertens, 'Crystallization', *Computers and Automation*, **19**(8), 22, (1970).

[21] Mohammad Ali Javaheri Javid, Mohammad Majid al Rifaie, and Robert Zimmer, 'Detecting Symmetry in Cellular Automata Generated Patterns Using Swarm Intelligence', in *Theory and Practice of Natural Computing*, eds., Adrian-Horia Dediu, Manuel Lozano, and Carlos Martín-Vide, volume 8890 of *Lecture Notes in Computer Science*, pp. 83–94. Springer International Publishing, (2014).

[22] Mohammad Ali Javaheri Javid and Rene te Boekhorst, 'Cell Dormancy in Cellular Automata', in *International Conference on Computational Science (3)*, eds., Vassil N. Alexandrov, G. Dick van Albada, Peter M. A. Sloot, and Jack Dongarra, volume 3993 of *Lecture Notes in Computer Science*, pp. 367–374. Springer, (2006).

[23] Kenneth C Knowlton, 'Explor-a generator of images from explicit patterns, local operations, and randomness', in *Proceedings of 9th Meeting of UAIDE*, pp. 544–583, (1970).

[24] S. Kullback and R. A. Leibler, 'On Information and Sufficiency', *The Annals of Mathematical Statistics*, **22**(1), pp. 79–86, (1951).

[25] Ming Li, *An introduction to Kolmogorov complexity and its applications*, Springer, 1997.

[26] Penousal Machado and Amílcar Cardoso, 'Computing aesthetics', in *Advances in Artificial Intelligence*, 219–228, Springer, (1998).

[27] Jon McCormack, 'Evolving sonic ecosystems', *Kybernetes*, **32**(1/2), 184–202, (2003).

[28] Abraham Moles, *Information theory and esthetic perception. Trans. JE Cohen.*, U. Illinois Press, 1968.

[29] Frieder Nake, 'Information aesthetics: An heroic experiment', *Journal of Mathematics and the Arts*, **6**(2-3), 65–75, (2012).

[30] Jaume Rigau, Miquel Feixas, and Mateu Sbert, 'Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity', in *Workshop on Computational Aesthetics*, eds., Douglas W. Cunningham, Gary Meyer, and Laszlo Neumann, pp. 105–112, Banff, Alberta, Canada, (2007). Eurographics Association.

[31] Ir Remko Scha, 'Kunstmatige Kunst', *De Commectie*, **2**(1), 4–7, (2006).

[32] L.F. Schwartz and L.R. Schwartz, *The Computer Artist's Handbook: Concepts, Techniques, and Applications*, W W Norton & Company Incorporated, 1992.

[33] Claude Shannon, 'A mathematical theory of communication', *The Bell System Technical Journal*, **27**, 379–423 & 623–656, (October 1948).

[34] Karl Sims, 'Interactive evolution of equations for procedural models', *The Visual Computer*, **9**(8), 466–476, (1993).

[35] Tomáš Staudek, *Exact Aesthetics. Object and Scene to Message*, Ph.D. dissertation, Faculty of Informatics, Masaryk University of Brno, 2002.

[36] Peter Struycken, 'Splash 1972/1974', in *Artist and computer*, ed., Ruth Leavitt, 30–31, Harmony Books, (1976).

[37] Renate Wackerbauer, Annette Witt, Harald Atmanspacher, Jürgen Kurths, and Herbert Scheingraber, 'A comparative classification of complexity measures', *Chaos, Solitons & Fractals*, **4**(1), 133–173, (1994).

[38] D. J. Wilson, 'An experimental investigation of Birkhoff's aesthetic measure', *The Journal of Abnormal and Social Psychology*, **34**(3), 390, (July 1939).

[39] Stephen Wolfram, *A New Kind of Science*, Wolfram Media Inc., 2002.

[40] Wojciech H Zurek, 'Algorithmic randomness and physical entropy', *Physical Review A*, **40**(8), 4731, (1989).

Four PPPPerspectives on Computational Creativity

Anna Jordanous [1]

Abstract. From what perspective should creativity of a system be considered? Are we interested in the creativity of the system's output? The creativity of the system itself? Or of its creative processes? Creativity as measured by internal features or by external feedback? Traditionally within computational creativity the focus had been on the creativity of the system's Products or of its Processes, though this focus has widened recently regarding the role of the audience or the field surrounding the creative system. In the wider creativity research community a broader take is prevalent: the creative Person is considered as well as the environment or Press within which the creative entity operates in. Here we have the Four Ps of creativity: Person, Product, Process and Press. This paper presents the Four Ps, explaining each of the Four Ps in the context of creativity research and how it relates to computational creativity. To illustrate how useful the Four Ps can be in taking a fuller perspective on creativity, the concepts of novelty and value explored from each of the Four P perspectives, uncovering aspects that may otherwise be overlooked. This paper argues that the broader view of creativity afforded by the Four Ps is vital in guiding us towards more encompassing and comprehensive computational investigations of creativity.

1 Introduction

A practical issue arises when considering the evaluation of a computational creativity system: from what perspective should creativity of a system be considered? Are we interested in the creativity of the system's output? The creativity of the system itself? Or of its creative processes? Creativity as measured by internal features or by external feedback?

The computational creativity community has traditionally considered creativity from the perspective of the creative output produced by a system, or the processes employed within creative systems (with notable exceptions, such as Saunders [48]). The call for this ICCC 2014 conference invites papers addressing the 'Process vs. product: addressing the issue of evaluating/estimating creativity (or progress towards it) in computational systems through study of what they produce, what they do and combinations thereof.'

This paper argues that to consider process and product is not enough; computational creativity should be considered and explored from four different perspectives, known as the Four Ps: the creative Person, Product, Process and Press (or environment) [43, 26].

The Four Ps have long been prevalent in creativity research relating to humans[2] and enable a more inclusive and encompassing approach to the study of creativity and accommodating multiple relevant perspectives. Here the Four Ps are presented and considered in the light of how they are relevant to computational creativity researchers.

1.1 The product/process debate in computational creativity evaluation

> 'As a research community, we have largely focussed on assessment of creativity via assessment of the artefacts produced.' [8, p. 1]

As illustrated by the ICCC 2014 call for papers, one important debate in computational creativity is about whether evaluation of a creative system should focus exclusively on the output produced by the system, or whether the processes built into the system should also be taken into account. Should both product and process should be included in evaluation [39, 8, 20], or should evaluation concentrate solely on the product of systems [45]? Ritchie [45] stated that examining the process is unimportant for creativity, arguing that humans normally judge the creativity of others by what they produce, because one cannot easily observe the underlying process of human creativity. Ritchie therefore advocated a black-box testing approach, where the inner program workings are treated as unknown and evaluation concentrates on the system's results. Later, however, Ritchie [46] conceded that it can be important to consider a system's 'mechnisms' in the case of 'more theoretical research'[46, p. 147].

While it is true that we can only use the material we have available to form an evaluation, evaluation experiments [36, 19] show that people often make assumptions about process in their judgements on product. As Hofstadter pointedout, '*covert mechanisms* can be deeply probed and eventually revealed merely by means of watching *overt behaviour* ... [this approach] lies at the very heart of modern science.' [15, quoted in p. 10, [39]]. Pearce & Wiggins [36] discussed how our interpretation of how something was produced is important, even if the actual method is unknown, and that such an interpretation can be derived if people are repeatedly exposed to the compositional systems (human or computational) that they are evaluating. Collins [6] discussed how making reasonable assumptions can assist the reverse-engineering[3] of program code from output, in scenarios where white-box testing (evaluation with access to the program code) is not possible.

Colton [8] acknowledged Ritchie's arguments but quotes examples from art to demonstrate that process is as important as the end product when evaluating creativity, at least in the artistic domain. As evidence, Colton cites conceptual art for details on conceptual art in the context of this debate, where the concepts and motivations behind the artistic process are a significant contribution of the artwork. Sol LeWitt defined Conceptual Art [25] as an art form where 'the

[1] University of Kent, UK, email: a.k.jordanous@kent.ac.uk

[2] Variants of these *Ps* also arise in slightly different guises in non-related areas, such as software project management [16] or education [2].

[3] Reverse-engineering is the process of identifying and perhaps replicating how a product is made, through analysis of that product.

idea or concept is the most important aspect of the work. ... The idea becomes a machine that makes the art.' Two examples are Tracey Emin's controversial exhibit *My Bed* (1999) and Duchamp's Fountain (1917). Jordanous [20] makes similar arguments for creativity in musical improvisation, finding that the process of improvisation is often seen as more relevant for creativity than the end result.

If assessing how creative a piece of conceptual art or a musical improvisation is, solely by evaluating the product, then there are two negative consequences:

1. The primary intentions of the artist/musician are ignored (their focus is on how the creative work is made rather than the end result).
2. The level of creativity presented will probably be underestimated, especially if the creative process results in producing something that might seem commonplace outside the context of that art installation/musical performance.

Colton [8] also posed a thought experiment that considers two near-identical paintings presented at an exhibition. In the first painting, the dots are placed randomly, whereas in the second, the dots' locations represent the artist's friendships with various people. Colton argued that the second painting would be more appealing to purchase than the first, though the end product is very similar, due to the process by which it was created. Colton's thought experiment illustrates how process can impact on our judgement of creative artefacts, though one could question if the experiment explores perception of creativity, or of quality/appeal.

The thought experiment described by Ventura [54] gives further evidence (perhaps unintentionally) on how knowledge of the creative process affects how we evaluate creativity. Two creative systems, the RASTER and iRASTER systems, were designed by Ventura to be decidedly non-creative. If these systems were implemented and their generated images were given to people to evaluate without telling the evaluators how they were produced, the evaluators may well rate the creativity of the system highly. Supplying the evaluators with details of how a program works, though, could have a detrimental impact on the subsequent evaluations [11, 8].

One issue with creativity is analogous to the adage that a magician never reveals their secrets. This adage is based on the fact that tricks do not appear so impressive once you have found out how the magician performed the trick. Similarly things can appear to be less creative when you know how they were produced:[4]

> 'it is not unknown for critics of AI to refuse to accept programs as creative (or intelligent) once the mundane mechanistic nature of the inner workings are revealed' [44, p. 4]

Colton [8] intentionally sidestepped this issue by reporting on his artistic system in high-level terms only, rather than giving details of the program [8, p. 8].

Until recently, computational creativity evaluation methodologies mainly looked solely at a system's *products* [45, for example] or at a combination of the *products* and the *process* [39]. Recently it has been acknowledged that there is more to creativity than process and product, with the Creative Tripod [8], whose evaluative framework is influenced by how an audience perceives the creativity of a system, SPECS [20] which requires the researcher to investigate what creativity means in the context of their system, and the FACE/IDEA models [9] which consider various aesthetic features and interactions between audience and system. Work on computationally creative societies has also developed in the last few years [48, is a significant example].

Along a similar broadening of perspectives, the next section brings in work from the wider creativity research community, examining further viewpoints - the creative *person* operating in a *press*/environment - and relating these viewpoints to a computational creativity standpoint.

2 The *Four Ps* of creativity

One major approach in creativity research is to break down creativity into four perspectives, commonly referred to as the *Four Ps* [43, 51, 34, 26, 49, 53, 35]:

- Person: The individual that is creative.
- Process: What the creative individual does to be creative.
- Product: What is produced as a result of the creative process.
- Press: The environment in which the creativity is situated.

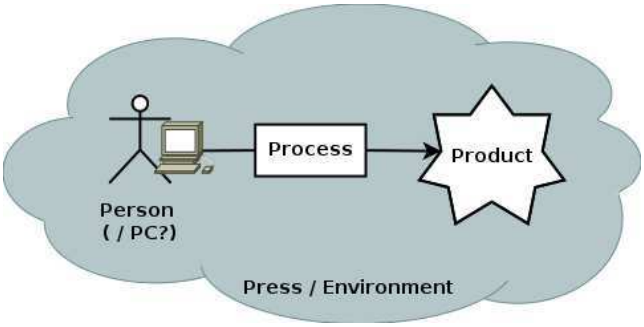


Figure 1. A simplified view of how the Four Ps fit together in creativity

Rhodes [43] was perhaps first to identify the four P perspectives. Rhodes collected 40 definitions of creativity and 16 definitions of imagination. The '*Four P*' dimensions of creativity emerged from analysis of these definitions.[5] Several people seem to have independently identified four similar themes of creativity [26, 51, 34, 35], boosting the credibility of the Four Ps.

Plucker, Beghetto & Dow [41] conducted a literature survey investigating the use (or absence) of creativity definitions in creativity research. As part of this review, Plucker et al. used their analysis to derive their own definition by identifying reoccurring themes and forming these into an inclusive definition which happens to account for each of the Four Ps:

> 'Creativity is the interaction among *aptitude, process, and environment* by which an individual or group produces a *perceptible product* that is both *novel and useful* as defined within a *social context*' [41, p. 90]

In reviewing Four Ps research, Kaufman [23] described addendums that have been suggested for the Four Ps: persuasion [49]

[4] If the inner workings of a program are very impressive, complex or novel, then we may still be impressed by the program, but this is a different perspective to whether or not we think the program is creative.

[5] As Rhodes' work appeared in a relatively unknown journal, many later advocates of a 'Four Ps'-style approach to creativity seem unaware of Rhodes' contribution (e.g. Odena, 2009, personal communications), so fail to cite him.

and potential [47]. In general, however, the Four Ps have been adopted as they were originally conceived by various researchers [43, 51, 34, 26].

2.1 The Four Ps: Person

This perspective addresses human characteristics associated with creative individuals or groups of people. Encouraged by Guilford's call in 1950 for studying the creative person, an abundance of different personal characteristics have been associated with creativity [43, 51, 24, 53, 35], ranging from personality traits, attitudes, intelligence and temperament to habits and behaviours (for example curiosity, persistence, independence and openness). Some of these are closely related; others are contradictory. Rhodes mentioned the relevance to creativity of people's personality traits, attitudes and habits, physique and intelligence and the identifiable features of creative people, as well as referring to people's temperament, habits, self-concept, value systems, defence mechanisms, and behaviour [43, p. 307].

Empirical studies up until 1968 were summarised by Stein [52] and were combined into a list of 18 distinct personality characteristics of a creative person, including aspects such as curiosity, persistence, independence and openness. Stein used these characteristics to identify creative individuals for study. There is a risk of circularity here, as the selection criteria for creative individuals chooses people to be studied, then the study involves examining those characteristics and criteria. Stein's work has not stood the test of time, with few current citations.

Several researchers subdivide the 'Person' category into finer-grained groups. Three sub categories of the 'pupil' perspective emerged during Odena & Welch's work [35]: personal characteristics of the pupil, their individual learning style (either adapting to new information or deriving new information themselves) and the influence of the pupil's background. Koestler [24] described three types of creative person: the *Artist*, the *Sage* and the *Jester*. Through Tardif & Sternberg's review of definitions of creativity, three main categories were identified with which to describe creative people: cognitive attributes, personality attributes/motivation and developmental influences. Tardif & Sternberg suggested three resultant modes of study of human creativity: cognitive psychology; psychometric testing; and study of human development.

These discrepancies and the sheer quantity of attributes together place an obstacle in the way of compiling a definitive list of attributes of a creative person and instead provoke disagreements on exactly which cognitive characteristics should be attributed to creative people. Tardif & Sternberg's review showed that as of 1988, different authors highlighted a variety of characteristics, with no general consensus and no characteristics common to all reports [53, Table 17.1, p. 434].

2.1.1 The Person in Computational Creativity

In computational creativity, the creative person could be analogous to the computer, or perhaps more accurately, to the computer program, software, or to a creative agent within a multi-agent system. Here the machine is the hardware hosting the creative agent, much as we might distinguish between physical and functional characteristics of a 'Person'.

Interesting work has been done on modelling creative agents, for example by Saunders [48], although the emphasis in computational creativity software tends to be on product generation and to some extent, process modelling, rather than on the modelling of characteristics of a creative Person in computational format. This is because computational creativity systems tend to be oriented towards a particular goal or domain, rather than being generally creative, as we can see by the plethora of domain-specific systems (as opposed to modelling of creative personal characteristics) in the various proceedings of ICCC conferences (International Conference on Computational Creativity). As argued in [20], different types of creativity require domain specific skills to some extent, so domain-specific computational creativity systems tend to be built around the most prominent necessary skills for that domain.

In terms of evaluating creative systems, Colton's Creative Tripod [8] emphasises the need for systems to demonstrate skill, imagination and appreciation before they can be considered as a candidate creative system, all three of which are alluding to personal characteristics.

Features, traits and aspects of the creative system can be studied, and it would be fascinating to explore how general creative personal characteristics could be specifically modelled within creative systems (see the *Process* section, next). Computational modelling of characteristics that encourage creativity could help us progress our systems to be able to be creative in more than one system which they were originally designed for; this would be significant progress in our pursuit of modelling creativity as a phenomenon which transcends different types of creative activity.

The 'Person' could also entail the individual(s) interacting with a creativity support system or co-creative system which interacts with people[27, 22]. Another possible interpretation of the 'Person' in computational creativity would be to acknowledge the role of the programmer(s), tester(s), researcher(s) and peers involved in shaping the project.

2.2 The Four Ps: Process

The creative process has been broken down into a series of sequential or cyclic stages occurring over time [42, 55] or subtasks [35].

In their work on student creativity in school music lessons, Odena & Welch [35] broke down the creative process into subtasks, identifying various types of process (e.g. different activities, group process, the structuredness or otherwise of a process and composition by improvisation) rather than tracing a linear progression of subprocesses.

It is often stressed that creativity is not just the first flash of inspiration, but is also the activity that validates, develops, and refines that first idea; rather than occurring at one point in time, creativity develops over a period of time [55, 42, 53]. Tardif & Sternberg [53] questioned whether creativity is a social or an individual process. The social view of creativity has notably been promoted by Csikszentmihalyi [12].

2.2.1 The Process in Computational Creativity

In computational creativity, the creative process might be that employed by a single piece of software, or the interactions between multiple machines or programs, or the interactions between machine and human users. As described above, the computational creativity community has given some attention to the concept of creative processes employed within computational creativity, with growing attention paid to this aspect in recent years. For example, the FlowR framework [5] is designed to facilitate creative computational workflows by chaining together processes in a linear pattern, and from personal communications with members of the project team, there are plans

to consider non-linear chains of processes as well. Additionally, the work by Joanna Misztal on poetry generation [31] specifically focusses on the processes required to generate poetry, at various levels of abstraction.

The generate-and-test [30, 38] or engagement-reflection approach [40] specifically models the creative process as a cycle of generating artefacts then improving the generation process via evaluating the generation phase. This is an approach which deserves broader adoption within computational creativity; evaluation is a critical part of the creative process [42, 12]. In terms of post-implementation evaluation, the FACE model for evaluation of creative systems [9] places importance on computational systems being able to report on the creative process (this report is referred to in the FACE framework as a *Frame*).

There are multiple theories about how human creativity processes are structured (see for example [42, 12, 23, 14]). Computational creativity research can provide a test-bed for these psychological theories and allow us to explore if implementing the theories result in creative behaviour. Conferences such as the Creativity and Cognition series showcase work that links between theory and practice to some extent, but further activity along these lines would emphasise the validity of computational creativity research, allowing computational work to contribute to human creativity research and vice-versa.

2.3 The Four Ps: Product

Many authors advocate that *proof* of creativity is necessary to be considered creative [21, 53, 41, 44]. The product-centric view adopted by computational creativity researchers such as Ritchie [45], that creative products are both necessary and sufficient for creativity, was present in earlier human creativity research [21]. But, inspired by Guilford's seminal 1950 address on creativity research, emphasis in human creativity research shifted from identifying creative individuals post-production of creative work, to predicting future potential for creativity in individuals. This change in emphasis is illustrated in the proliferation of psychometric tests [23, 19] within creativity research.

Tardif & Sternberg [53] considered the creative product more briefly than the other three 'Ps' in their review, deciding that while a creative product is essential for creativity, it is not enough merely to generate a product; the product should also be considered in a domain-specific context.

Computational creativity research has long acknowledged the importance of the output or artefacts generated by creative systems, as described above. To borrow a metaphor from human creativity research, it has been common (until recently) for computational creativity to follow the product-centric approach to creativity as advocated by Kagan: '*Creativity* refers to a product, and if made by a man, we give him the honor of the adjective' [21, p. viii].

2.3.1 The Product in Computational Creativity

Generating creative products has been an area of significant success for computational creativity. To see examples, one just needs to consult any year's proceedings of the International Conference on Computational Creativity where there are multiple examples to be found of systems which are reported in terms of the products they generate. The success of systems is often reported in terms of what kind of artefacts they generate, as noted in [18]. Some systems have been evaluated using Graeme Ritchie's empirical criteria [44, 45], which exclusively focuses on evaluating the products of computational systems without considering any of the other three Ps.[6]

2.4 The Four Ps: Press/Environment

The Press perspective encompasses a bidirectional perspective between the environment which influences the creator and receives the creative work, and the creator who publicises their work and is given feedback on what they produce. Tardif & Sternberg [53] considered both creative domains themselves and the social environments in which creative people are influenced as they employ creative process, advertise their creative products and receive feedback. Rhodes [43] concentrated on the role that the environment plays on a person during the creative process, rather than how the creative produce is judged by the external world after being created. Rhodes reflected on how everyone is different, so everyone perceives the world in a unique way and processes ideas according to their own contexts.

Of the Four Ps, this is the perspective that is often neglected when one takes an individualistic view of creativity. In general creativity theorists do however acknowledge the influence of the environment in which creativity is situated [49, 13]. If one concentrates on an individual's creativity, however, the Press perspective is often neglected, even if unintentionally. For example, although stating that '[t]o be appreciated as creative. a work of art or a scientific theory has to be understood in a specific relation to what preceded it' [3, p. 74], Boden's treatment of creativity mainly focused on different cognitive processes of creativity, rather than a detailed examination of social or environmental influences.

2.4.1 The Press in Computational Creativity

Some computational creativity researchers are starting to highlight the importance of the environment in which a creative system is situated [50, 17, 37, 48], with some of this work influenced by the DIFI (Domain-Individual-Field-Interaction) framework [12]. Social interaction between creative agents and their audience is an area which has been neglected by all but a few groups of researchers: for example nearly 75% of papers in the 2014 International Conference on Computational Creativity failed to make any reference to social or interactive aspects of creativity. But creativity cannot exist in a vacuum. A recent increase in development of the interactivity of creative systems (especially where this affects the way these systems works) is pleasing to see and deserves further attention [10].

There is a separate point to acknowledge regarding Press in computational creativity. As computational creativity researchers, we should stay aware of any potential biases that may be introduced, should an audience be aware that the creative agent of interest is computational rather than human [32, 19].[7]

2.5 Interaction between the Four Ps

Simonton [49] saw discrepancies between combining the Four Ps in theory and in practice:

> 'Now, in an ideal state of affairs, it should not matter which one of the four p's our investigations target, for they all will converge on the same underlying phenomenon. ... But reality is not so simple, needless to say. The creative process need not

[6] Recently proposed evaluation methods such as [8, 9, 19] place more emphasis on the other three 'Ps'.

[7] Many thanks to the anonymous reviewer who noted this point.

> arrive at a creative product, nor must all creative products ensue from the same process or personality type; and others may ignore the process, discredit the product, or reject the personality when making attributions about creativity.' [49, p. 387]

From this, one conclusion which seems to follow naturally is that an accurate and comprehensive definition of creativity must account for the (potential) presence of all four aspects, in order to be complete. Simonton, however, concluded that '[i]f we cannot assume that all four aspects cohesively hang together, then it may be best to select one single definition and subordinate the others to that orientation' [49, p. 387], with his natural research inclination leading him to focus his work on *persuasion*, his term for the Press/Environment aspect.

The mysterious impression often associated with creativity [56, 3, 23] can be explained to some extent when one or more of the Four Ps are not accounted for:

> 'Each strand [of the Four Ps] has unique identity academically, but only in unity do the four strands operate functionally. It is this very fact of synthesis that causes fog in talk about creativity and this may be the basis for the semblance of a "cult".' [43, p. 307]

Rhodes argued that creativity research should follow a specific path: 'from product to person and thence to process and to press.' [43, p. 309]

> 'Objective investigation into the nature of the creative process can proceed in only one direction, i.e. from product to person and thence to process and to press.' [43, p. 309]

Such a statement makes Rhodes's contribution less useful. For example, the Press (environment) in which one is creative has some influence on the creative Process, so one may prefer to study how Press and Person interact before looking at Process issues. Simonton viewed creativity as how a person's ideas emerge as influential when that person, by chance, has new ideas and promotes them to influence others. Creative people would not be equivalent to lucky people, by this interpretation, but chance would intervene in their success. Simonton refers to this as the 'chance-configuration theory' that 'outlines the general conditions that favor creativity' [49, p. 422].

Tardif & Sternberg [53] treated each of the Four Ps individually, 'as these really are separate levels of analysis, and it is from comparisons within levels that coherent statements about our knowledge of creativity can be made' [53, p. 429]. Tardif & Sternberg's summary is weakened somewhat by this as it does not make comparisons across the Four Ps, despite highlighting Simonton's emphasis on the interactions and relations between these four views [49]. In contrast Mooney [34] argued that the four approaches should be integrated in a model of creativity, proposing a model that 'puts together the four approaches by showing them to be aspects of one unifying idea' [34, p. 333]. While Mooney's claims become rather grandiose at points, Mooney's more specific contributions on creativity match neatly with the four Ps approach identified elsewhere at that date [43, 51]

2.5.1 Interaction between Four Ps in Computational Creativity

This paper argues that we can make significant progress in computational creativity by considering all four Ps in our computational creativity work. Tony Veale's tagline for the ICCC'2012 conference sums up current aspirations of computational creativity well; Veale characterises computational creativity research as 'scoffing at mere generation for more than a decade'. Generation of creative products is only a quarter of the full picture of creativity, only one of the Four 'Ps'. Granted, we have achieved much success in product generation, as exemplified by exhibitions, concerts and other demonstrations of creative products reported in various papers on computational creativity systems [18]. However, the more mature work and exciting potential comes from the incorporation of the other three Ps, at least to some extent, such as in [40, 48, 31].

3 Applying the Four Ps: examples of *novelty* and *value*

Novelty (originality, newness) and value (usefulness, appropriateness) form key parts of creativity [28, 3, 45, 20], often being identified as the two main aspects of computational creativity [39, 45, 4, for example].[8] Work in computational creativity illustrates both novelty and utility from each of the Four P perspectives, although some perspectives are represented more plentifully within computational creativity than others. To illustrate the discussions above, we can discuss novelty and value in computational creativity from each of the Four P perspectives. Considering novelty from each of the Four Ps:

Product Novelty is well associated with system outputs and products: how novel are the generated artefact(s)? The novelty of artefacts generated by computational creativity systems is a key consideration in Ritchie's empirical criteria for evaluating creative systems [45].

Process A creative process can take a novel approach or be implemented in a novel way, perhaps employing new algorithms or techniques or different approaches. Efforts at trying new processes and combinations thereof are being encouraged by systems such as the FlowR framework [5], which focuses specifically on enabling us to chain different processes together for creative purposes.

Person Creativity can be performed by a new creative entity, which demonstrates or uses novel characteristics relevant to that creativity. As is often encountered in computational creativity work, implementing or running a creative system on new hardware or in different software may also impact upon the system's performance and may have unexpected results. The number of new systems presented each year at the International Conference on Computational Creativity exemplifies how novel creative entities continually arise in computational creativity research.[9] (Also, the novelty of unexpected results is often unintentionally exemplified when live demos of these systems are attempted in unfamiliar computing setups.)

Press The creativity demonstrated by a system can be noted as being novel in a particular environment, even though it may be commonplace in other environments. The system may also exploit the surrounding press in previously unexplored ways. This was demonstrated neatly by the combination of two systems in [33], where a textual annotation system interacted with a system that generates emotion-driven music. The combination resulted in novel interpretations of fairy tales; such results would not have arisen were the systems operating in isolation.

Considering value from each of the Four Ps:

[8] It should be clarified that for this author, creativity consists of considerably more than novelty and value, though these are two key components of creativity. See [20].

[9] See `http://www.computationalcreativity.net/conferences`.

Product Value is also well associated with system outputs and products: how valuable or good are the generated artefact(s)? This is a highly current area of concern within computational creativity, with much evaluation concentrating on the quality of output [18].
Process The creative processes being incorporated within creativity can be useful in themselves for learning or studying how certain approaches and techniques work or for cross-application to new areas. Systems with an emphasis on modelling process, such as Misztal and Indurkhya's poetry generator [31] bring added utility by what they reveal about the processes being modelled.
Person Some creators become more valuable than others as a contributor in their field, based on their personal characteristics, experience and influence.[10] The same can be noted for creative systems to some extent; some are cited more often than others, for example Simon Colton's HR mathematical discovery system [7] (which provides a useful example of creativity in a non-artistic domain).
Press If creative activities benefit the external world in some way, then they have value to the press. As example, Harold Cohen's AARON colouring system has received much external attention, from media discussions [29] through to inspiring a screensaver for personal computers via `http://www.kurzweilcyberart.com`.

These above lists are not intended to be a full and conclusive portrait of novelty and value within computational creativity. What these lists illustrate is the different viewpoints that can be uncovered using the Four Ps as *signposts* with which to guide our thinking around computational creativity. The breadth of issues mentioned above shows aspects of novelty and value within computational creativity which may not always be accounted for if taking a product/process-oriented viewpoint; however it is argued here that those perhaps-overlooked aspects give us a closer rendition of creativity, guiding us away from incomplete viewpoints of creativity in the context of our computational work.

4 Summary

The difficulty of understanding what creativity is should not discourage us from such an attempt [43, 41, 8]. In creativity research, the *Four Ps* construct ensures we pay attention to four key aspects of creativity: the creative Person, the generated Products, the creative Process and the Press/Environment hosting and influencing the creativity. This framework helps us to consider creativity more broadly.

For example, if viewing *novelty* and *value* from the perspectives of *product*, *process*, *person* and *press*, we uncover various interpretations of these two key concepts within computational creativity which may otherwise have been overlooked. The *Four Ps* framework helps to highlight different perspectives on creativity, to portray creativity in a fuller context.

ACKNOWLEDGEMENTS

Many thanks to Carly Lassig for originally making me aware of the Four Ps and pointing me towards Rhodes [43]. Thanks also to various computational creativity researchers who provided feedback on earlier versions of these thoughts while in development, and the anonymous reviewers of this paper.

[10] This has been found, for example, in the recent Valuing Electronic Music project `http://valuingelectronicmusic.org` [1], where some people's endorsements can have a greater influence on the perceived value of an electronic musician and their work.

REFERENCES

[1] Daniel Allington, Byron Dueck, and Anna Jordanous. Networks of value in Electronic Dance Music: SoundCloud, London, and the importance of place, submitted.
[2] J. B. Biggs, 'Constructing learning by aligning teaching: constructive alignment', in *Teaching for quality learning at university: what the student does*, 11–33, SRHE & Open University Press, Buckingham, (2003).
[3] Margaret A. Boden, *The creative mind: Myths and mechanisms*, Routledge, London, UK, 2nd edn., 2004.
[4] David Brown, 'Computational artistic creativity and its evaluation', in *Computational Creativity: An Interdisciplinary Approach*, number 09291 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, (2009).
[5] John Charnley, Simon Colton, and Maria Teresa Llano, 'The flowr framework: Automated flowchart construction, optimisation and alteration for creative systems', in *Proceedings of the 5th International Conference on Computational Creativity*, (2014).
[6] Nick Collins, 'The analysis of generative music programs', *Organised Sound*, **13**(3), 237–248, (2008).
[7] Simon Colton, *Automated theory formation in pure mathematics*, Distinguished dissertations, Springer, London, UK, 2002.
[8] Simon Colton, 'Creativity versus the perception of creativity in computational systems', in *Proceedings of AAAI Symposium on Creative Systems*, pp. 14–20, (2008).
[9] Simon Colton, John Charnley, and Alison Pease, 'Computational Creativity Theory: The FACE and IDEA descriptive models', in *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 90–95, Mexico City, Mexico, (2011).
[10] Simon Colton and Geraint A. Wiggins, 'Computational creativity: The final frontier?', in *Proceedings of 20th European Conference on Artificial Intelligence (ECAI)*, pp. 21–26, Montpellier, France, (2012).
[11] David Cope, *Computer Models of Musical Creativity*, MIT Press, Cambridge, MA, 2005.
[12] Mihaly Csikszentmihalyi, 'Society, culture, and person: a systems view of creativity', in *The Nature of Creativity*, ed., Robert J. Sternberg, chapter 13, 325–339, Cambridge University Press, Cambridge, UK, (1988).
[13] Beth A. Hennessey and Teresa M. Amabile, 'The conditions of creativity', in *The Nature of Creativity*, ed., Robert J. Sternberg, chapter 1, 11–38, Cambridge University Press, Cambridge, UK, (1988).
[14] Beth A. Hennessey and Teresa M. Amabile, 'Creativity', *Annual Review of Psychology*, **61**, 569–598, (2010).
[15] Douglas Hofstadter, 'Creativity, brain mechanisms and the Turing test', in *Fluid concepts and creative analogies*, ed., Douglas Hofstadter, 467–491, Harper Collins, New York, (1994).
[16] Ivar Jacobson, Grady Booch, James Rumbaugh, James Rumbaugh, and Grady Booch, *The unified software development process*, volume 1, Addison-Wesley Reading, 1999.
[17] K. E. Jennings, 'Developing creativity: Artificial barriers in artificial intelligence', *Minds and Machines*, **20**(4), 489–501, (2010).
[18] Anna Jordanous, 'Evaluating evaluation: Assessing progress in computational creativity research', in *Proceedings of the Second International Conference on Computational Creativity (ICCC-11)*, Mexico City, Mexico, (2011).
[19] Anna Jordanous, *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application*, Ph.D. dissertation, University of Sussex, Brighton, UK, 2012.
[20] Anna Jordanous, 'A Standardised Procedure for Evaluating Creative Systems: Computational creativity evaluation based on what it is to be creative', *Cognitive Computation*, **4**(3), 246–279, (2012).
[21] *Creativity and learning*, ed., Jerome Kagan, Beacon Press Press, Boston, MA, 1967.
[22] Anna Kantosalo, Jukka Toivanen, Hannu Toivonen, and Ping Xiao, 'From isolation to involvement: Adapting machine creativity software to support human-computer co-creation', in *Proceedings of 5th International Conference on Computational Creativity, ICCC*, Ljubljana, Slovenia, (2014).
[23] James C. Kaufman, *Creativity 101*, The Psych 101 series, Springer, New York, 2009.
[24] Arthur Koestler, *The act of creation*, Danube Books, New York, 1964.
[25] Sol LeWitt, 'Paragraphs on conceptual art', *Artforum International Magazine*, **June 1967**, (1967).
[26] Douglas W. MacKinnon, 'Creativity: a multi-faceted phenomenon', in

Creativity: A Discussion at the Nobel Conference, ed., J. D. Roslansky, 17–32, North-Holland Publishing Company, Amsterdam, The Netherlands, (1970).
[27] Mary Lou Maher, 'Computational and collective creativity: Who's being creative?', in *Proceedings of the 3rd International Conference on Computer Creativity*, Dublin, Ireland, (2012).
[28] Richard E. Mayer, 'Fifty years of creativity research', in *Handbook of Creativity*, ed., Robert J. Sternberg, chapter 22, 449–460, Cambridge University Press, Cambridge, UK, (1999).
[29] Pamela McCorduck, *Aaron's code: Meta-art, artificial intelligence, and the work of Harold Cohen*, WH Freeman, New York, NY, 1991.
[30] G. McGraw and Douglas Hofstadter, 'Perception and creation of diverse alphabetic styles', *AISB Quarterly*, **85**, 42–49, (1993).
[31] Joanna Misztal and Bipin Indurkhya, 'Poetry generation system with an emotional personality', in *Proceedings of 5th International Conference on Computational Creativity, ICCC*, (2014).
[32] David C. Moffat and Martin Kelly, 'An investigation into people's bias against computational creativity in music composition', in *Proceedings of the 3rd International Joint Workshop on Computational Creativity (ECAI06 Workshop)*, Riva del Garda, Italy, (2006).
[33] Kristine Monteith, Virginia Francisco, Tony Martinez, Pablo Gervás, and Dan Ventura, 'Automatic generation of emotionally-targeted soundtracks', in *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 60–62, Mexico City, Mexico, (2011).
[34] Ross L. Mooney, 'A conceptual model for integrating four approaches to the identification of creative talent', in *Scientific Creativity: Its Recognition and Development*, eds., Calvin W. Taylor and Frank Barron, chapter 27, 331–340, John Wiley & Sons, New York, (1963).
[35] Oscar Odena and Graham Welch, 'A generative model of teachers' thinking on musical creativity', *Psychology of Music*, **37**(4), 416–442, (2009).
[36] Marcus Pearce and Geraint Wiggins, 'Towards a framework for the evaluation of machine compositions', in *Proceedings of the AISB Symposium on AI and Creativity in Arts and Science*, York, UK, (2001).
[37] Alison Pease and Simon Colton, 'On impact and evaluation in computational creativity: A discussion of the Turing Test and an alternative proposal', in *Proceedings of the AISB'11 Convention*, York, UK, (2011). AISB.
[38] Alison Pease, Markus Guhe, and Alan Smaill, 'Some aspects of analogical reasoning in mathematical creativity', in *Proceedings of the International Conference on Computational Creativity*, pp. 60–64, Lisbon, Portugal, (2010).
[39] Alison Pease, Daniel Winterstein, and Simon Colton, 'Evaluating machine creativity', in *Proceedings of Workshop Program of ICCBR-Creative Systems: Approaches to Creativity in AI and Cognitive Science*, pp. 129–137, (2001).
[40] R. Pérez y Pérez, A. Aguilar, and S. Negrete, 'The ERI-Designer: A computer model for the arrangement of furniture', *Minds and Machines*, **20**(4), 533–564, (2010).
[41] Jonathan A. Plucker, Ronald A. Beghetto, and Gayle T. Dow, 'Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research', *Educational Psychologist*, **39**(2), 83–96, (2004).
[42] Henri Poincaré, 'Mathematical creation', in *The Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method.*, volume Science and Method [Original French version published 1908, Authorized translation by George Bruce Halsted], chapter III of Book I. Science and the Scientist, 383–394, The Science Press, New York, (1929).
[43] Mel Rhodes, 'An analysis of creativity', *Phi Delta Kappan*, **42**(7), 305–310, (1961).
[44] Graeme Ritchie, 'Assessing creativity', in *Proceedings of the AISB Symposium on AI and Creativity in Arts and Science*, pp. 3–11, York, UK, (2001).
[45] Graeme Ritchie, 'Some empirical criteria for attributing creativity to a computer program', *Minds and Machines*, **17**, 67–99, (2007).
[46] Graeme Ritchie, 'Uninformed resource creation for humour simulation', in *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 147–150, Madrid, Spain, (2008).
[47] M.A. Runco, 'Creativity, cognition, and their educational implications', in *The educational psychology of creativity*, ed., John Houtz, 25–56, Hampton Press, New York, NY, (2003).
[48] R. Saunders, 'Towards autonomous creative systems: A computational approach', *Cognitive Computation*, **4**(3), 216–225, (2012).
[49] Dean Keith Simonton, 'Creativity, leadership, and chance', in *The Nature of Creativity*, ed., Robert J. Sternberg, chapter 16, 386–426, Cambridge University Press, Cambridge, UK, (1988).
[50] Ricardo Sosa, John Gero, and Kyle Jennings, 'Growing and destroying the worth of ideas', in *Proceedings of the 7th ACM Creativity and Cognition conference*, pp. 295–304, Berkeley, California, (2009).
[51] Morris I. Stein, 'A transactional approach to creativity', in *Scientific Creativity: Its Recognition and Development*, eds., Calvin W. Taylor and Frank Barron, chapter 18, 217–227, John Wiley & Sons, New York, (1963).
[52] Morris I. Stein, 'Creativity', in *Handbook of personality theory and research*, eds., Edgar F. Borgotta and William W. Lambert, 900–942, Rand McNally, Chicago, IL, (1968).
[53] Twila Z. Tardif and Robert J. Sternberg, 'What do we know about creativity?', in *The Nature of Creativity*, ed., Robert J. Sternberg, chapter 17, 429–440, Cambridge University Press, Cambridge, UK, (1988).
[54] D. Ventura, 'A reductio ad absurdum experiment in sufficiency for evaluating (computational) creative systems', in *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 11–19, Madrid, Spain, (2008).
[55] Graham Wallas, *The Art of Thought*, C. A. Watts & Co, London, UK, abridged edn., 1945.
[56] Raymond Williams, *Keywords: a vocabulary of culture and society*, Fontana/Croom Helm, Glasgow, UK, 1976.

How Many Robots Does It Take? Creativity, Robots and Multi-Agent Systems

Stephen McGregor and **Mariano Mora McGinity** and **Sascha Griffiths**[1]

Abstract. This paper seeks to situate computational creativity within the context of ongoing theoretical and practical investigations of environmentally situated and dynamic systems. Beginning with a consideration of the evidently goal directed nature of creativity, the problem of how teleological behaviour emerges in a fundamentally physical world. Creativity is reassessed as a search for goals in a dynamic environment rather than as a pursuit of a fixed goal in a stable and finite space of possible actions. A significant consequence of this evaluative shift is the impossibility of considering truly creative systems as anything other than embodied agents deeply entangled in an environmental situation. Two fields are discussed as potential habitats for such systems: robotics and multi-agent systems. Creativity from the perspective of ongoing research in these areas is considered, and some preliminary thoughts for future directions of enquiry are offered.

1 Introduction

This paper will address the question of the relationship between goals and creativity. Notions of purpose are so deeply ingrained in the standard view of creation that creativity itself is often defined in terms of the accomplishment of some expressive objective. Implicit in the problem of modelling creativity, however, is the emergence of end directed action in a reductionist world: how can something that is not in a physical sense present nonetheless contribute to the operation of a physically supervenient system?

Having posed the question of how a creative agent views its own objectives, the paper will turn to an exploration of the related problem of causality. In particular, the emergence of absent causes – which is to say, the influence of possible worlds, both historical and futuristic, removed from present reality – is addressed. This etiological inquiry is couched in terms of evolution by natural selection, with a brief consideration of this well researched process as a model of evidently goal directed and therefore potentially creative behaviour. A general hypothesis regarding the viability of explaining goals as emergent properties of complex systems, grounded in contemporary theoretical investigations of dynamic systems, will be put forward. *Contra* the idea that computationally creative agents must necessarily be handed a well defined goal by an external designer, dynamic processes are proposed as a basis for models that can discover their own goals through collaboration and environmental interaction.

This theoretical consideration is followed by a preliminary exploration of two compelling areas of research that move beyond what has been the *de rigueur* constraint satisfaction approach to computational creativity. First the topic of robotics will be considered from the perspective of the modelling of creativity, with particular attention to the problem of how a robot obtains, represents, and adapts its own goals. Robots are importantly embedded in a physical environment, and this situation opens the door to the possibility of the emergence of dynamic attractors that might be construed as new and unexpected goals outside any representation of an objective built into a robot's programming. The conclusion of this investigation will be that it seems reasonable to at least consider the possibility of an adequately flexible robot formulating goals that can be considered as evidence of its own creativity.

Next multi-agent models are considered, with particular attention to the ways in which complex patterns of activity with the trappings of intentionality can emerge from interactions within a population of agents individually following very basic sets of predetermined rules. As with robots in their environmental entanglements, swarms of agents have some prospect of generating collective behaviour that can be interpreted as being directed towards ends outside of the simple constraint satisfaction requirements programmed into the functioning of each independent agent. In the case of multi-agent models, the model becomes the environment, with the attractors that arise in the course of interactions becoming the handles for assessing the system in terms of the formulation and pursuit of goals. On the one hand, interpretation of action in a simulation of such a system presumably still falls back on the analysis of an external observer. On the other hand, the emergent properties of such systems potentially offer models of the parallel emergence of cognitive phenomena such as creativity in a physically grounded universe. Again, there seems to be scope for considering the implementation of multi-agent systems as a form of computational creativity that begins with a traditional programming task but that subsequently goes beyond mere constraint satisfaction.

The ideas proposed in this paper are at this stage indications of direction for future research. Exciting work is currently being done in the fields of both robotics and multi-agent systems, with some applications specifically towards modelling creative systems [1, 17]. This paper is intended to serve as a bellwether for further research in this direction, with the objective of moving beyond a constraint satisfaction approach to computational creativity. Traditional rule based implementations of creative agents have accomplished much in recent years, but reconsidering the emergence of goals within highly dynamic environments offers the grounding of a more robust argument for creative autonomy arising within the systems themselves. This reconsideration of the relationship between creativity, goals, and the environmental situation of creative agents can furthermore become a platform for extended discussions of interesting philosophical questions about causation and cognition.

[1] Queen Mary University of London, UK, email: s.e.mcgregor, m.moramcginity, sascha.griffiths@qmul.ac.uk

2 Creativity and Goals

> What redeems it is the idea only... and an unselfish belief in the idea – something you can set up, and bow down before, and offer a sacrifice to...
>
> JOSEPH CONRAD, *The Heart of Darkness*

If an agent cannot choose its own goals, can its "creative" behaviour really be considered creative? This question is familiar to anyone who has attempted to construct a creative agents, especially agents that produce works of art. "*Why* did the computer choose this word, this note, this color, and not another?" is generally asked by people confronted with an artistic piece produced by a computer. If the same question were put to an artist the answer would most likely be as unsatisfactory as that provided by the computer, yet it is the computer's response that is most troubling. The assumption is that art made by computers is something that needs to justify itself, whereas it is accepted that artists produce according to their inspiration, perhaps because works of art are seen as reflections of the mind or the personality of the artist who made them, and it is troubling to think of a computer as having a "personality". However, the fuzziness of the artist's response seems to reflect something essential in the creative process, something which can and should be exploited in the design of artificial creative systems: the goal is not fixed; it shifts and drifts and wanders; it might not even exist *a priori* but rather will emerge from the creative process itself.

2.1 Causality and Teleology

Aristotle's "Doctrine of the Four Causes" presents a framework for understanding the relationship between actions and motivations in a world of physical events [2]. Starting with the essentially reductive premise that the material nature of entities is at the root of actions in the world, the Doctrine outlines a hierarchy of relationships, culminating in the theory of how teleological – which is to say, goal directed – behaviour stands as the "final cause" that explains the regularity with which functional effects are produced in the natural world. Aristotle's four causes can be enumerated:

1. Material Cause - the behaviour indicated by the physical properties of matter
2. Efficient Cause - the consequences of the manipulation of physical material
3. Functional Cause - the reasoning regarding efficient causes that informs actions on materials
4. Final Cause - the goal that motivates functional planning

Conventional approaches to creativity generally descend Aristotle's causal ladder: there is a goal, a plan for achieving this goal, a set of actions carried out to realise that plan, and a world of physical relationships in which those actions have consequences. Indeed, a fundamental principle of a certain approach to aesthetics is that the perception of beauty involves the recognition of a function that defines an artefact and an appreciation of the creative process employed in the achievement of that functionality [46]. An alternative theory, rooted in the philosophy of Kant, considers aesthetic experience to unfold in a perceptual domain of its own, involving a detachment from any practical consideration of an object of beauty [20]. Even in this latter case, though, beauty, from the perspective of a creator, becomes an objective unto itself, with the elicitation of an aesthetic response in principle indicating the achievement of this goal. So, regardless of the theoretical grounding adopted by an analyst, creativity seems to be bound up in an end directed process.

Computational creativity has tended to adopt a similar line. Ritchie has characterised the creative behaviour of an information processing machine in terms of the identification of a class of existing artefacts that qualifies as a target domain, subsequent generation of artefacts that are expected to fall within this domain, and then evaluation on the part of the system of whether the creative goal has been achieved [32]. Output produced without some sort of goal criteria has been described as "mere generation", a ramble through a state space that, regardless of its consequences, cannot be properly considered as creativity [11]. This lines up well with Boden's description of levels of computational creativity, with high level transformations of state spaces trumping lower level recombinations of elements within a predefined space [10]. It is in this transformational degree of symbol manipulation, involving the delineation of a state space above a traversal of a known space, where the complexity of the goal directed aspect of creativity becomes evident. A fundamental challenge for a computer scientist interested in designing autonomously creative systems is therefore to understand what it would mean for computers to make decisions about the definition of their own search spaces.

But it is not even clear how teleological processes arise in the material world, reducible, as it is, to the interactions of physical fields on a very small scale. Deacon has taken Aristotle's Doctrine as a starting point for his own exploration of the emergence of goal oriented behaviour in material reality, beginning with the premise that modern philosophy has sometimes tended to use dualism and homunculi to obscure the hard question of final cause [13]. For Deacon the first step up from the tumult of pure physics is a consideration of thermodynamic processes, by which a tendency that is so reliable it has a nomic aspect emerges from the random interaction of particles. In fact, despite the regularity implicit in the terminology "laws of thermodynamics", there is no principle that requires systems to move towards entropic arrangements; it is just the overwhelmingly likely outcome of a stochastic process. The kernel of teleology might be discovered in the apparent lawfulness of entropy that arises in systems that are actually just complex and unpredictable.

Like Deacon, Kauffman recognises the seeds of emergence in the way that order can spontaneously come about in a dynamic system, giving rise to interpretable attractors [19]. The contemporary case for emergence maintains that nested hierarchies of interacting attractors can be extrapolated into apparently teleological behaviour. At the higher end of the scale exist cases like evolution by natural selection, which, while it has been grasped through an astounding act of reductionist interpretation, can nonetheless only really be understood as a process directed towards the goal of fitness—and in fact it has been argued that evolution itself should be treated as a creative process. To put it simply, an evolved organism is a confluence of functions that result in their own perpetuation. Taking an example offered by Millikan, the biological operation of an animal can only be understood in terms of the functional role that the creature's various organs play in sustaining life, and these functions have been determined through an assiduous process of evolutionary trial and error: a lion's heart exists in order to pump blood through the lion's body, even though the genetic and developmental process that resulted in the existence of the organ cannot have been somehow aware of that outcome [24].

But, Millikan asks, what happens if a fully developed lion comes into existence spontaneously? While the lion might be considered an operational organism, it is tempting to conclude that its organic components have no function in the sense of having been selected

because of a goal they accomplish. An evolved lion inherits properties of goal directedness from the generational history of organisms that has contributed to its fitness. This extension of the lion's emergent identity into the past corresponds to a converse projection of the functional properties of its components towards the accomplishment of future goals, specifically the goals of the lion surviving and reproducing. The spontaneous lion, on the other hand, while it also has some hope of coincidentally surviving and replicating, has simply happened: it cannot be interpreted as the fulfillment of a goal that has emerged in the unfolding of events in a complex and unpredictable environment. In terms of Aristotle's efficient cause, the lions are identical, but in terms of final cause they seem to be completely different.

Bickhard has responded to Millikan's case for a connection between causal history and functionality, however, by arguing that the history of a system cannot be a part of its ongoing operation [9]—history is, presumably, a contextualised interpretation of a present situation. Instead, Bickhard proposes, function should be understood in terms of the contribution a functional component makes to its system's persistence in a state that defies the entropic tendencies of the universe [8]. It is the case that the dynamics of complex and chaotic systems result in the emergence of processes that, in their regularity, seem to have a sense of following some kind of rule. This shift away from the basic laws of physics begins with processes such as thermodynamics, where the regularity lies precisely in the predictable breakdown of structure in systems, and moves out towards the further from equilibrium states that characterise the process of evolution, or more explicitly cognitive apparati such as representational symbols.

So by the emergentist account, causation is understood in terms of nested layers of dynamically coupled, intricately entangled processes, with each emerging attractor becoming an element in a higher level of interactions. This view escapes the paradoxes of trying to incorporate some representation of the system's past into its current operation, and at the same time seeks to explain the evident gravity of future outcomes in the workings of higher order complexes. The upshot of this is that teleological processes are necessarily associated with systems that are highly non-linear on several levels, an insight that sits well with the enactivist world view of Varela, Thompson and Rosch, who suggest that a mindful agent – which is to say, one capable of the planning and execution inherent in creativity – must be situated in a deeply interactive relationship with a dynamic and unpredictable environment [42].

There is a gravely concerning ramification to this conclusion from the perspective of a computer scientist interested in designing autonomously creative agents, however: if teleological processes only emerge in the context of complex interaction with a chaotic environment, it is difficult to imagine how a symbol manipulating machine could hope to creatively flourish in its rule based domain. Considering that even computational processes modelled non-deterministically can be reduced to linear operations, the case for a strictly algorithmic system producing output that would be judged even basically creative seems doomed. Two possibilities immediately present themselves as the beginning of a solution to this challenge: the modelling of dynamic interactions between rule following agents, and the physical construction of environmentally situated robots.

3 Robots

The classic intelligent agent concept [34] entails that an agent should be able to use actuators to manipulate its environment, which it monitors via perception. The agent has goals which it is trying to satisfy via its actions. Whether these goals have been reached is subject to an evaluation which the agent achieves by applying a metric. In classic AI, the agent's environment is understood much less literally than it is in robotics. Robots exist in a physical world that they actively manipulate and that directly affects their actions. Also, they share this world with humans. What follows offers a brief survey of contemporary approaches to robotics as they relate to creativity, followed by some thoughts on the future exploration of robots as creative agents.

From the perspective of the description of creative systems, the great appeal of robots lies in their situation in the same highly non-linear environment from which human creators have emerged. As a first approximation, robots might be considered to have goals that are handed to them by a designer, grounded in external observations: in this case, the robot becomes an expression of its own creator's stance towards the world, and even in this basic instance a dynamic emerges where the robot's behaviour can become an element in a larger creative system, with the designer responding to the robot's successes and failures through subsequent design decisions. In what follows, this scenario will be case in terms of robots as a form of creative expression. More complexly, robots might be modelled as adaptive agents involved in a feedback loop with their own environments. In this case, while there may be overarching goals handed to a robot by a designer, it is the behaviour that emerges in the pursuit of this goal that may be considered creative. Ultimately, it is conceivable that robots or perhaps even more compellingly networks of robots might be involved in processes with unpredictable outcomes that can be interpreted as the emergence of truly goal oriented causation.

It has recently been argued [27] that real progress in natural language processing will depend on a more human-like machine which has a situated knowledge embodied in its own physical form. This presence [26] is necessary for a cognitive architecture which is more human-like and therefore capable of a human-like command of natural language. This may just as well be just as true for other cognitive abilities.

Feldman [16] sees two possible ways in which a robot can fully understand human subjective experience. One is a full simulation of the human body to gain insights into human experience. The other would be a new type of grounding that builds up an understanding of the world through the robot's own sensors and bodily experience.

Creative automata and machines which exist in the physical world have been built for centuries. There is, for instance, the case of von Kempelen's speaking machine [43, 15], which was a hybrid between a research project on the human vocal apparatus and an entertainment tool similar to a musical instrument.

Creativity in the domain of robotics can be conceptualised in terms of creative activities that are performed by intelligent agents capable of performing a full action-perception loop which takes the environment into consideration. Within this action-perception loop the, agent must have some "creative goal".

3.1 Agents and Embodiments

In order to understand what it means for a robot to be creative we will now describe a few systems which do in some way fulfill the criterion of being "deemed creative" if they were "performed by a human" [44]. Creative robots come in two flavours currently: they are either presented as being creative themselves or they are used as tools for expressing a human's creativity. We will fist deal with the later kind of robot for creative tasks.

Robots as a form of creative expression are teleoperated, which is to say their actions are determined by the perceptions and decisions

of a human performer. Ogawa et al. [30] report on a teleoperated robot called the "Geminoid" [36] being used for a task in which the android and an actor performed a play live on stage together. This robotic agent had the following properties:

- The android takes the shape of a physical body which is modelled on an actual female human. The body has 12 degrees of freedom (DoF). These are mainly used for its facial expressions which closely copy the operator's facial movements. It also has loudspeakers which transmit the operator's voice to the audience.
- Perception is accomplished through a camera system which lets the operator see the machine's view of its environment.
- The machine's processing of the environment is realised by feeding the video back to the operator, and its actions are hence based on receiving "commands from the human operator".

So the robot's body itself is used for artistic expression. The authors conclude, based on experiment, that the robot actually improved the audience's sense of immersion in the performance. This is a surprising result but shows that the embodiment through the artificial agent can actually generate a different level of "meaning", as the authors suggest. It is actually the human-like but not-human body that generates this added meaning.

Robots as creative agents are autonomous to a certain extent. Tresset and Deussen [40] report on a robot, named e-David, which creates visual art through painting on a canvas. This agent had the following properties:

- e-David is not anthropomorphic (human-like). It is an industrial robot that only consists of an arm. The arm is also its actuator, with which it manipulates a pencil or brush.
- The perceptive apparatus is a camera system.
- The system performs the action-perception loop by creating an image it intends to paint and then monitoring its output by perceiving the painting as it emerges through its own actions applied to a canvas.

Embodiment is crucial in the case of e-David. The authors list thirteen ways in which e-David's embodiment has a direct impact on the final result of the visual art it produces. These include the velocity at which the arm moves, the pressure it applies to the painting, and control of the amount of paint on the brush. All of the factors have a direct effect on the visual appeal of the product which e-David produces. Thus, this robot demonstrates the importance of considering the physical presence of an artificial creative system in the creation of visual art.

Both the Geminoid and e-David illustrate how important the actual physicality of an intelligent agent it is and how their individual embodiments shape their creative output. However, the processing system in each case is actually quite different. Whereas e-David is autonomous in its actions to a large extent, the Geminoid is operated by a human. Thus, these two specific robots have different levels of autonomy and one needs to debate what "responsibilities", in the sense of Colton and Wiggins [11], they take on within the creative process.

3.2 Goals

As already illustrated, robotic agents that use their physical appearance and structure to pursue creative objectives can differ in their goals. Whereas the Geminoid in the study discussed above tries to evoke emotional response in an audience, e-David monitors its own output on a canvas via a visual feedback system.

Similarly, musical robots have goals which they pursue. In this case, the environment is typically the musical instrument with which the robots interact physically.

A robot coordinating its own body in a creative process

Batula and Kim [6] present a system which plays the score of a two-finger piece on a piano. The robot in this case is a small humanoid. Its environment is a keyboard. Its perception relates to the monitoring of its own motions and audio-feedback.

The robot's goal is to play the piece it has been assigned correctly. The authors frame their research as an investigation into the motorics required for musicianship. The robot's goals are simple: it detects mistakes in its own playing. This is very much in line with our argumentation. The system's physicality comes from the control of its own limbs in relation to the velocity of its playing. The robot controls its own motion, and the decisions of how to play rely solely on its own bodily control.

A robot coordinating with another body in a creative process

A contrasting approach is presented by Mizumoto et al [25]. In their approach the focus is on ensemble performance. The goal is for the agent to ask: "Am I creating the same output as another agent?"

This is a different question because the machine is no longer in control of the speed at which the product is created. The robot plays a theremin while the human plays drums. The robot's perception is used to actually calculate the action of the actuators, in contrast to the actuators acting independently to exert force on the physical environment. The required processing relies on a coupled-oscillator model.

3.3 Environments

What kind of environments do robots encounter in the course of creative processes? The comedic robot is one recent concept which has been implemented. Thus far, these robots are the only agents which actually treat an audience as their environment. They do exactly what an intelligent agent does by monitoring what effects their actions have on the environment.

Audience Monitoring

Other agents with which robots interact may be artificial (see section 4) or human audiences. Knight [21] analyses the impact of embodiment on performances in robot theatre. Knight et al [22] present a system which tells jokes to an audience.

In the system described a small humanoid robot is the comedic agent. Its goal is to make the audience laugh. It monitors levels of audience interest and attention (more precise methods are further described in [23]). The robot presents jokes and will choose the sequence of jokes in accordance with the audience's reaction. This is a direct application of the action-perception loop. The quality of the creative output is measurable in the sense that the audience reaction is the operationalisation of what the output should achieve.

Interacting with the Audience

Katevas at al [18] also use a humanoid robot as a stand-up comedian. In their performance, however, the robot actively engages with the audience and directly addresses individual members of the audience. In this way, the robot influences the outcome of the creative process. The goal is an active audience reaction, so the robot tries to improve the outcome and generate more laughs by engaging the audience.

As such the robot is not only relying on its output in the form of jokes, but also actively and preemptively shapes the audience's reaction and hence its environment's reaction to the jokes. This can be

considered a different approach. If joke telling is considered an artistic and creative process, then the audience's reaction is the measure by which one can tell whether the result of the activity is of good quality. The robot here imitates the practices of human stand-up comedians by actively inducing a reaction in the audience. It does not just rely on the humorous value of the verbal stimuli it presents to the audience.

3.4 Creative Robots

In line with the theoretical points raised above, entertainment robotics is a growing market [7, 33]. The potential here is vast. A robot can use the principles outlined above to become an active companion [12, 3], giving itself an advantage over static media such as television broadcasting or film.

The three principles addressed here, embodiment, goals, and environments, will play a crucial goal in developing systems that can be deemed creative. This section has illustrated differing approaches to all three of these topics. In designing creative robotic systems, the human designer will have to think carefully about how the agent will pursue its goals within the given environment.

In line with the argument in this paper, for a robot to be truly creative it must be able to show adaptive behaviour. Embodiment will obviously be given from the outset in a robotic system, influencing the system's actions, perception, and interaction with the environment in a non-trivial way. However, real adaptivity for creativity will arise only if the robotic agent is able to define its own goals. An approach to robotics which includes this kind of behavioural autonomy is evolutionary robotics [29]. This approach assumes that the agent has some kind of overall goal such as playing a musical piece or amusing an audience via comedic practices. The sub-goals upon which the system operates would have to be adaptable. One way of implementing such a strategy would be to devise methods that allow the robot to choose between the goals outlined above (see section 3.2), or, with respect to interacting with the environment, choosing between the two strategies of, for instance, interaction with an audience as described above (see section 3.3).

4 Multi-Agent Systems

It is sometimes easy to forget that artists are not totally isolated from their environment: they come into contact with other artists who are tackling problems and trying to reach goals very similar to their own. Artists, scientists, chess players, normal people trying to make ends meet—creative people are influenced by other people, and they themselves influence other people, very often people with whom they are in no direct contact. Think of the generations of musicians influenced by Beethoven or of mathematicians working on problems formulated by Gauss.

In fact, one would be justified in thinking that creative processes are never the work of one individual alone, no matter how visionary and illuminating her thinking might be: every creator stands on the shoulders of giants. The intention here is to discuss how this interaction might be modelled through artificial agents, and how such an interaction might influence the behaviour of the agents towards, ultimately, determining the goal of the creative process itself.

As they relate to the imperative of creative goals as behavioural causes, the appeal of multi-agent systems is their potential for producing emergent attractors which cannot be understood as components of any single agent's behaviour. Agents themselves may be goal oriented – indeed, their processes are typically modelled in terms of the satisfaction of very basic criteria – but these goals are simplistic, whereas the operation of the overall system is nuanced. The power of simple agents collaborating to develop and realise complex goals can be observed in various real-world contexts, from the swarm behaviour of certain insects to the efficacious productivity of financial markets and indeed the homeostatic condition of entire ecological systems. This paper considers the question of how computers might be used to model multi-agent systems and then to analyse the potential for considering these systems as generators and executors of creative objectives.

4.1 Interacting agents

Multi-agent systems have been used extensively to model the origin and evolution of an impressive array of different social constructs [14, 31, 35], from ant or termite colonies to computer networks to economic markets. Agents are assigned a more or less rigorous set of beliefs, desires and intentions which determines their interaction. Agents are goal-oriented: their actions are determined by a desire to maximise a reward function, and it is through their interaction that the system evolves.

Most interestingly, multi-agent systems can show emergent properties: interaction between the agents allows the self-organisation of system properties that were not originally part of the system. Self-organisation, i.e. the lack of a centralised element imposing structure on the emergent property, is an important characteristic of such systems, revealing how organised properties can arise from simple interactions alone. These kinds of systems have been used, for instance, to model the self-assembling of biological complex structures[28], or to model the origin and evolution of language [37, 5, 4, 38]. In Steels' work, agents create and agree on a lexicon to name a series of objects in their environment. Their interaction follows a protocol specified in a "language game", similar to the language games described by Wittgenstein [45]. Van Trijp [41] shows that the "Naming Game" will converge towards a stable lexicon if certain requirements are met.

According to Tomasello [39]:

> The current hypothesis is that it is only within the context of collaborative activities in which participants share intention and attention, coordinated by natural forms of gestural communication, that arbitrary linguistic conventions could have come into existence evolutionarily...

This hypothesis seems to validate the modelling approach. An effort to understand creative processes as an attempt at collaborative behaviour by intelligent agents might prove to be very fruitful.

4.2 Creative processes as collaboration: a thought experiment

We propose a thought experiment which could help to illuminate the relationship between goal-seeking behaviour and creativity. Agents of different physical or cognitive characteristics are placed in an environment and forced to collaborate in order to achieve a series of tasks. To simplify things, we propose the following interaction rules:

1. All interactions are one-to-one: two agents are chosen and made to interact.
2. Agents are chosen at random: the system does not show a topology, i.e. it is a mean-field system.

3. One of the agents adopts the role of the demonstrator; the other is the observer.

Both agents have a clear idea of the task that is to be carried out. However, their different physical and cognitive skills require them to adapt their own actions to the task: some agents are better equipped to carry out the task in one way, whereas others must find efficient ways to carry out the task. During the interaction, the demonstrator performs the task in the most efficient way it can. Obviously, this way depends on all the previous experience of the agent. More particularly, it depends on what it has learned from all its previous interactions with other agents.

Following this demonstration, the observer must decide whether it is fit to perform the action in the same way. It does this by attempting to imitate the demonstrator. If it cannot, it must try to find a way to perform the action in a way that will resemble the demonstrator's actions, only adapted to its own abilities. If it succeeds in carrying out the action, then the observer will include this action into the set of actions it is capable of carrying out to perform the assigned task; the game is successful and two new agents are chosen to play new game. If, on the other hand, after a fixed number of attempts the agent is incapable of performing the action in a satisfactory way, then the game starts again, only now a new task is chosen: the goal changes. The new task should be similar to the previous one, if possible, so that agents might be able to identify properties of the task that are difficult for them, and perhaps learn to avoid them or find a way around them.

The hypothesis offered here is that such a system would become stable, i.e. it would reach a point after which all interactions would be successful. At this point all agents would have learned how to behave when forced to carry out a collaborative task. Every agent would have learned to adapt its own goal, according to its capabilities, to fulfill the task in a cooperative manner. Every agent would have learned how to work around what it cannot do.

5 Conclusion

This paper has examined the idea that creativity can be understood in terms of a process of adaptation on the part of agents attempting to accomplish a set of goals in complex and unpredictable environments. The hypothesis presented here is that agents dynamically coupled with their environments might become involved in the instigation of higher level emergent features that can be interpreted as potentially surprising and valuable new goal directed behaviours. There is scope for hoping that a network of multiple environmentally situated agents, each independently working towards their own micro-goals, will remit a systemic shift that in turn can become a target for discovery of new possible goals available to agents. From an external perspective, such a system offers the overall impression of being directed towards goals that are not in any way present in the programming that defines the behaviours of its components. In the physical universe, definable as it is in terms of a few simple rules of interaction, has nonetheless become a cauldron for such complex emergent systems evolution and cognition. In the same sense, a system of simple, interactive, environmentally oriented computational agents might have a chance of developing patterns of behaviour that can collectively be considered creative.

Existing work in the pertinent fields of robotics and multi-agent systems has been briefly discussed. The embodied situation of robots invites a consideration of the development of goal directed behvaiour in an unpredictable environment. And the dynamics of multi-agent systems present a platform for investigating the possibility of treating the attractors that emerge unexpectedly in the course of interaction as unanticipated creative objectives. The juxtaposition of these two topics in the context of computational creative naturally suggests an amalgamation: a potential project developing swarms of individually adaptive robots, treating their own community of robotic co-agents as an environment embedded in the physical world, with each robot adapting its behaviour based on its interaction with its peers. On an individual level, the robots would update their procedures based on observations of other robots and with the pre-programmed objective of accomplishing simple goals. On a collective level, the robotic system as a whole might very well take on an emergent aspect, with unexpected intimations of higher level organisation. The question raised by such a model is whether the system's proclivity for organising itself in a surprising and potentially effective way can be considered the creative discovery of a new objective.

ACKNOWLEDGEMENTS

Griffiths is supported by ConCreTe: the project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733. McGregor is supported by EPSRC grant EP/L50483X/1.

REFERENCES

[1] Mohammad Majid al Rifaie, John Mark Bishop, and Suzanne Caines, 'Creativity and Autonomy in Swarm Intelligence Systems', *Cognitive Computation*, **4**(3), 320–331, (2012).

[2] Aristotle, *The Works of Aristotle Volume II: Physica*, The Clarendon Press, Oxford, 1930. Translated by R. P. Hardie and R. K. Gaye.

[3] Ruth S Aylett, Ginevra Castellano, Bogdan Raducanu, Ana Paiva, and Mark Hanheide, 'Long-term socially perceptive and interactive robot companions: challenges and future perspectives', in *Proceedings of the 13th international conference on multimodal interfaces*, pp. 323–326. ACM, (2011).

[4] A. Baronchelli, V. Loreto, L. Dall'Asta, and A. Barrat, 'Bootstrapping communication in language games: Strategy, topology and all that', in *The Evolution of Language, Proceedings of the 6th International Conference (EVOLANG6), edited by A. Cangelosi, A. D. M. Smith & K. Smith, World Scientific Publishing Company*, (2006).

[5] J Batali, 'Computational simulations of the emergence of grammar', in *Approaches to the Evolution of Language: Social and Cognitive Bases*, eds., J R Hurford, M Studdert-Kennedy, and Knight C., 405–426, Cambridge University Press, Cambridge, (1998).

[6] Alyssa M Batula and Youngmoo E Kim, 'Development of a mini-humanoid pianist', in *10th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2010)*, pp. 192–197. IEEE, (2010).

[7] George A Bekey, *Autonomous Robots*, MIT Press, Cambridge, MA, 2005.

[8] Mark Bickhard, 'The Emergence of Contentful Experience', in *What Should Be Computed to Understand and Model Brain Function?*, ed., Tadashi Kitamura, World Scientific, (2001).

[9] Mark Bickhard, 'The Dynamic Emergence of Representation', in *Representation in Mind, Volume 1: New Approaches to Mental Representation (Perspectives on Cognitive Science)*, eds., Hugh Clapin, Phillip Staines, and Peter Slezak, Elsevier, (2004).

[10] Margaret A. Boden, *The Creative Mind: Myths and Mechanisms*, Weidenfeld and Nicolson, London, 1990.

[11] Simon Colton and Geraint Anthony Wiggins, 'Computational creativity: The final frontier?', in *Proceedings of 20th European Conference on Artificial Intelligence (ECAI)*, eds., L. De Raedt, C. Bessiere, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, and P. Lucas, pp. 21–26, Montpellier, France, (2012). IOS Press.

[12] Kerstin Dautenhahn, Sarah Woods, Christina Kaouri, Michael L Walters, Kheng Lee Koay, and Iain Werry, 'What is a robot companion-friend, assistant or butler?', in *Intelligent Robots and Systems,*

2005.(IROS 2005). 2005 IEEE/RSJ International Conference on, pp. 1192–1197. IEEE, (2005).
[13] Terrence W Deacon, *Incomplete nature: How mind emerged from matter*, WW Norton & Company, New York, NY, 2011.
[14] Mark D'Inverno and Michael Luck, *Understanding Agent Systems*, Springer Verlag, 2nd edn., 2004.
[15] Homer Dudley and Thomas H Tarnoczy, 'The speaking machine of Wolfgang von Kempelen', *The Journal of the Acoustical Society of America*, **22**(2), 151–166, (1950).
[16] Jerome A Feldman, *From Molecule to Metaphor: A Neural Theory of Language*, MIT Press, Cambridge, MA, 2006.
[17] Petra Gemeinboeck and Rob Saunders, 'Creative Machine Performance: Computational Creativity and Robotic Art', in *Proceedings of the Fourth International Conference on Computational Creativity*, (2013).
[18] Kleomenis Katevas, Patrick G.T. Healey, and Matthew Tobias Harris, 'Robot stand-up: Engineering a comic performance', in *Proceedings of the 2014 Workshop on Humanoid Robots and Creativity at the 2014 IEEE-RAS International Conference on Humanoid Robots (Humanoids 2014)*, Madird, Spain, (2014). Available: $http://cogsci.eecs.qmul.ac.uk/humanoids/Katevasetal\_2014.pdf$.
[19] Stuart Kauffman, *At Home in the Universe: The Search for Laws of Complexity*, Oxford University Press, 1995.
[20] Gary Kemp, 'The Aesthetic Attitude', *British Journal of Aesthetics*, **39**(4), 392–399, (1999).
[21] Heather Knight, 'Eight lessons learned about non-verbal interactions through robot theater', in *Social Robotics*, eds., B. Mutlu, C. Bartneck, J. Ham, V. Evers, and T. Kanda, 42–51, Springer, (2011).
[22] Heather Knight, Scott Satkin, Varun Ramakrishna, and Santosh Divvala, 'A savvy robot standup comic: Online learning through audience tracking', in *International Conference on Tangible and Embedded Interaction*, Funchal, Portugal, (2010).
[23] Heather Knight and Reid Simmons, 'Estimating human interest and attention via gaze analysis', in *2013 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4350–4355. IEEE, (2013).
[24] Ruth Millikan, *Language, Thought, and Other Bilogical Categories*, MIT Press, Cambridge, MA, 1984.
[25] Takeshi Mizumoto, Takuma Otsuka, Kazuhiro Nakadai, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno, 'Human-robot ensemble between robot thereminist and human percussionist using coupled oscillator model', in *, 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1957–1963. IEEE, (2010).
[26] Roger K Moore, 'Presence: A human-inspired architecture for speech-based human-machine interaction', *Computers, IEEE Transactions on*, **56**(9), 1176–1188, (2007).
[27] Roger K Moore, 'From talking and listening robots to intelligent communicative machines', in *Robots that Talk and Listen – Technology and Social Impact*, 317 – 336, De Gruyter, Boston, MA, (2014).
[28] Radhika Nagpal, Attila Kondacs, and Catherine Chang. Programming methodology for biologically-inspired self-assembling systems, 2002.
[29] Stefano Nolfi and Dario Floreano, *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-organizing Machines*, Cambridge, MA, 2000.
[30] Kohei Ogawa, Koichi Taura, and Hiroshi Ishiguro, 'Possibilities of androids as poetry-reciting agent', in *RO-MAN, 2012 IEEE*, pp. 565–570. IEEE, (2012).
[31] Liviu Panait and Sean Luke, 'Cooperative Multi-Agent Learning: The State of the Art', *Autonomous Agents and Multi-Agent Systems*, **11**(3), 387–434, (2005).
[32] Graeme Ritchie, 'Some Empirical Criteria for Attributing Creativity to a Computer Program', *Minds and Machines*, **17**, 67–99, (2007).
[33] Florian Röhrbein, Sascha Griffiths, and Laura Voss, 'On Industry-Academia Collaborations in Robotics', Technical Report TUM-I1338, Technische Universität München, (2013).
[34] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall International, Harlow, third edn., 2013.
[35] Yoav Shoham and Kevin Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*, Cambridge University Press, 2008.
[36] S.Nishio, H.Ishiguro, and N.Hagita, *Geminoid:Teleoperated android of an existing person*, chapter 20, 343–352, I-Tech, Vienna, Austria, 2007.
[37] L. Steels, 'A Self-Organizing Spatial Vocabulary', *Artificial Life*, **2**(3), 319–332, (January 1995).
[38] L. Steels, 'Self-organization and Selection in Cultural Language Evolution', in *Experiments in Cultural Language Evolution*, ed., L Steels, John Benjamins, Amsterdam, (2012).
[39] M. Tomasello, *Origins of Human Communication*, MIT Press, Cambridge, MA, 2008.
[40] Patrick Tresset and Oliver Deussen, 'Artistically skilled embodied agents', in *Proceedings of AISB2014*, Goldsmiths, University of London, (1st - 4th April 2014).
[41] R van Trijp, 'The Evolution of Case Systems for Marking Event Structure', in *Experiments in Cultural Language Evolution*, ed., L Steels, 169–205, John Benjamins, Amsterdam, (2012).
[42] Francisco J. Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind*, MIT Press, Cambridge, MA, 1991.
[43] Wolfgang von Kempelen, *Über den Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*, J.V. Degen, Vienna, 1791. Facsimile re-print from 1970, Stuttgart: Frommann-Holzboog.
[44] Geraint A Wiggins, 'A preliminary framework for description, analysis and comparison of creative systems', *Knowledge-Based Systems*, **19**(7), 449–458, (2006).
[45] L. Wittgenstein, *Philosophische Untersuchungen*, Joachim Schulte Wissenschaftliche Buchgesellschaft, Frankfurt, 2001.
[46] Nick Zangwill, 'Aesthetic Functionalism', in *Aesthetic Concepts: Sibley and After*, eds., Emily Brand and Jerrold Levinson, Oxford University Press, Berlin, (2001).

The Creativity of Computers at Play

David C. Moffat
Department of Computing
Glasgow Caledonian University, UK.
D.C.Moffat@gcu.ac.uk

Abstract. There are many domains where creative software is being energetically developed, from writing and art to music and mathematics. These domains are open, without clear measures of value, and usually depend on humans to judge the creativity. While such research is obviously relevant to the nature of creativity, it may be that another creative domain is relatively overlooked; namely, that of puzzles.

This paper proposes the game of chess as a good domain in which to demonstrate, investigate and develop computational creativity. It shows some initial comparisons on two chess puzzles, one of which novices or even non-players could follow. The results support the case for computational creativity of programs that play in this domain. In conclusion, all puzzle or strategy games are suitable research testbeds for creativity, both natural and artificial.

1 Introduction — Creative domains

There are many domains where software has been tested for creativity, and is being energetically developed, from writing and art to music and mathematics. These domains are open, without clear measures of value, and typically depend on humans to judge the creativity. While such research is obviously relevant to the nature of creativity, it may be that another creative domain is relatively overlooked; namely, that of puzzles and play.

1.1 Games as a domain for computational creativity

Within the subject of games, AI has been able to make several large contributions. Most of them are general AI techniques, but one or two belong more specifically to the sub-field of computational creativity. First, let us recall that solving problems can be a creative activity, even if the solution is already known to somebody else.

Some researchers take the position that video games are highly relevant for the field of computational creativity. Liapis et al [7] go as far as to call games the "killer app for computational creativity." I certainly agree with their promotion of this perspective; but even they limit themselves in this recent position paper to matters which are generally forms of procedural content generation. My argument here pushes into the different role of computer as *player*.

1.1.1 Solving problems can be creative.

It is often said, at least in passing, that it takes creativity to solve (hard) problems. Engineering and design are creative endeavours, after all; and they consist largely in solving problems. They are not considered to be part of the "creative industries" however: they are not *called* "creative" (in the English-speaking world), and so they tend to get passed over in favour of the more overtly artistic domains. Even engineers themselves (such as AI researchers) tend to have this bias, as is evident in the field of computational creativity.

That is unfortunate, it seems to me, because the arts are in some ways still too challenging for the research field of computational creativity. In particular, to assess the quality of the supposedly creative products (computer generated art, music, jokes and poetry) requires human judgement; and that is extremely slow compared to computer speeds. Research could progress very much faster if only computers were set to work in a creative domain that did not depend on human reaction (at least not in real-time).

The suggestion of this paper is that we do have such a creative domain, and that it is relatively overlooked so far. The domain is that of games; and in particular the *playing* of them. Games are often puzzles in their own right, or they include puzzles within them, as modern video games do. In a typical story based video game, the player is expected to make decisions without having enough information to be sure, and without being able to foresee all the consequences. That is in essence a form of puzzle. There are puzzles placed throughout such games in their "levels" or areas within the virtual world where part of the story takes place. The player has to solve these puzzles before being able to move on through a door, or to the next level.

1.1.2 Games in computational creativity today.

Games are in fact a domain for the field of computational creativity, in the form of video games, and that is because it takes a great deal of labour to make the content for such games with their virtual worlds for player's characters to wander around in.

In order to save costs, video game programmers naturally make specialist software tools to help the designers generate the so-called "levels" of the game. The levels are virtual spaces filled with objects like: trees and houses, roads and walkways, obstacles and vehicles, and computer-controlled "non-player" characters, and the instructions they need to help them navigate around the space in an apparently intelligent way. In the bigger games there are many levels or areas with whole farms, fields and forests, and the virtual towns and cities have to be planned out just as real cities have town planners. To generate so much content for games is only feasible because of the specialist software that takes up much of the burden.

These software tools are increasingly automated, and able to make more appropriate design decisions, to better help the human designers. What the tools do is called "procedural content generation" (or PGC).

1.1.3 PGC is not play.

PGC is an increasingly important part of the industry, as well as an active area of academic research in computational creativity (or AI). Because it helps in the creative process of game design, PGC is obviously a part of the field of computational creativity. But PGC is AI for the making of games, not the playing of them; and it is *play* that is the focus of this paper.

There are other common AI contributions to games, including the use of finite state machines, fuzzy logics, decision trees, search algorithms, and occasionally even neural networks and genetic algorithms. These are AI, but are not part of the field of computational creativity. Neither are they uniquely applied to games, but are rather general techniques developed for and applied to other domains.

The work on search algorithms for games is a healthy and exciting research area these days, especially with the recent developments in Monte Carlo Tree Search (MCTS). Search algorithms like this are used to plan moves in puzzles and adversarial games, usually, like chess. In other words, search algorithms are used to make computers *play* games, but are seen as a mainstream AI technique that is useful for games, rather than as belonging to the sub-field of computational creativity. If that is an oversight, then it is the aim of this paper to correct it.

As other authors have recently noted then, PGC is an active and rich area for computational creativity [7] and [3]. However it is the computer as player that is of interest to me here, and is the area that is still treated relatively lightly, in my view.

1.2 Games and puzzles in AI history

While games have some overlap with computational creativity, they have been far more important to AI in general. It could be asserted that no other domain has been more important to AI, in fact. Let us first consider why that might be so, and then go on to reconsider creativity in that context.

1.2.1 AI has been at play since it began.

In a curious parallel to human development, the field of AI began playfully, before turning to more serious matters as it matured.

Even before modern digital computers existed, thinkers like Turing [10] and Shannon [9] were designing chess playing algorithms, and speculating that computers would one day play chess well enough to beat human players. If only they could have seen how right they were!

Rather like a child, AI in the early days was fed on challenges that led its development, including games like chess and checkers, and puzzles like trying to plan how to put childrens' toy wooden blocks on top of each other in a certain order. These tasks are usually called "toy problems" but they surely count as puzzles as well.

Games and puzzles were chosen as development challenges because they are formally and concisely specifiable, with clear goal conditions, and yet only humans could play them. Being thus characteristic of human intelligence, they were naturally seen as natural aims for computers (AI) to tackle. In the very name of AI, the early preoccupation with intelligence is clear to see. However, the related concept of *creativity* was mentioned much less often than intelligence. It still is, to this day, and indeed the research effort that declares its interest in creativity is tiny compared to the world's AI research.

On the other hand, when humans play, they are often said to be creative, in the way they develop interesting strategies or styles of play, or in finding novel but useful solutions to problems. Before we dismiss the possibility that computers might be creative in the way they play games, or solve problems, we should examine how humans are creative in play, if they are.

2 Creativity and play :

2.1 Play is creative for humans

Children and young animals are naturally playful. They play as part of growing up, in order to learn about their world. Humans are especially busy with play of all kinds, as first recognised by the Dutch historian Huizinga in his classic book asserting the layful nature of man, *Homo Ludens* [4]. Especially for humans, games are used to structure interactions and provide a context in which children (and adults) can play. This leads their cognitive and social development.

2.1.1 Play also encourages creativity.

This is partly because of the nature of the playground, which is a place of safety, but where different roles can be acted at the same time. Players can pretend to perform actions that in real life would be dangerous or impossible. For example, little boys often love to play with toy guns, and pretend they are shooting at each other. Later on, they may play first-person shooter video games like "Medal of Honor". Although they are bigger boys by then, or even full grown men, and the game has more "adult content", they are nevertheless still essentially playing as they did when they were little boys, with pretend guns. It is the safety of the game situation, and the pretence of it, that encourages a creative approach. Because there can be no serious consequences, and the danger is only pretend, and not real, it allows experimentation with different acts, from the illegal to the lethal and from tabu to terrorism.

Experimental thinking is necessary to creativity, as is taking the chance of being wrong. Making poor decisions in real life can have grave consequences, but in games failure is an opportunity to learn by trying again. Trying more risky actions, or a wider variety of actions, means that a there is more chance of discovering actions or decisions that lead to success eventually, even if initially they did not seem to. The style of thinking or problem solving in games or puzzles is thus ideally suited to finding new ways to achieve the desired goals. After more playing, more and better ways to win may be found. Eventually the player or puzzle-solver can discover the best and most elegant solutions: and these can properly be called "creative."

2.1.2 When is a puzzle solution creative?

The two most typical characteristics of creative products are commonly held to be *novelty*, and *quality* (or value). That is by now approaching a consensus [12]. We may question the novelty and the quality of a solution then, but is the "solution" the answer to a puzzle, is it something else, like the way that the answer was found?

To simplify the discussion at this point, let us consider the creativity of the *product* of thought, and not of the *process*, nor of the *producer*. The thinker of thoughts (the producer) is either a human or a computer, but we do not want our assessment of creativity to fall into a confusion about the nature of the thinker, such as whether it is warm to the touch, or as cold as metal. A definition of creativity that depends on body temperature has clearly gone wrong somewhere.

The way that thoughts are produced may be called creative with more legitimacy; but as some other authors do [12] I shall exclude this matter from the discussion, at least for this paper. That leaves the

question of whether the *product* of the thought processes (or calculations or algorithms) can be creative.

In the case of the solving of puzzles then, and of the playing of games which are often sequences of problems, we wish to know whether any solutions that are found can be called creative. If they are, then we should call those solutions creative, no matter who or what found them (e.g. human or computer).

2.1.3 *On the* novelty *of solutions*

Certainly for games and puzzles, the notion of creativity is immediately under threat here, because the solution must already be known by the person who sets the puzzle. Any game must have a way to win, and there must be a way to solve any puzzle, and there must be a way to check when the players have solved it correctly. Otherwise, they will get frustrated with wasting their time if there is no solution for them to find.

Following Boden's distinction between H-novelty (historical novelty) and P-novelty, we note simply that puzzle solutions are not H-creative, because the solution was already known [1, 2]. However, as the puzzle solver did not know it yet, the solution is new to him or her or it, so it is P-novel (for psychological novelty).

In a research strategy where we wish to study the psychological processes of creativity, this P-novelty is the ideal notion for us. It means that we can evaluate how well different algorithms perform in finding solutions that we already know about. To study algorithms that are aimed at H-novelty would be to apply our knowledge of creative processes, excitingly but would be appropriate only *after* we have gained the knowledge; and that can be arrived at best by studying P-novelty first.

Note that the creative *process* has just returned, uninvited but naturally enough, in that last point.

2.1.4 *On the* quality *of solutions*

As well as P-novelty, we need our problem-solving algorithms to produce *good* solutions, before we can call them creative. Here again, it is an advantage to research into games and puzzles as problem-solving domains. The evaluation of solution quality is typically built into the game or puzzle as part of its specification, usually in the form of a points score.

2.2 Is AI at play creative?

Although we left the issue of *process* behind, and attempted to make the final *product* bear the test of creativity alone, consideration of the extra criterion of *surprise* brought the *process* issue in again through the back door. It might be that the character of the *process* is what will ultimately determine whether we think that an algorithm is creative.

The source of creativity is still disputed in the field, with some researchers such as Indurkhya [5] including the audience or culture and society at large as co-contributors. That is an interesting view, but here we focus on the cognitive process as a determinant of creativity.

First let us consider playful algorithms as candidates for computational creativity. If people can be creative in the way they play games, then when AI plays games, and solves puzzles, is it being creative as well? Let us take the game of chess as an example.

3 Chess for (creative?) computer play

There is a deep history of chess in AI, which makes it a potentially rich domain for the field of computational creativity if it can be shown to be relevant in that regard. The world of chess is itself rich, and includes many forms of chess play, and other playfulness. Let us focus here on chess puzzles, or "compositions."

Iqbal and Yaacob [6] reported an extensive study on chess puzzles, and their aesthetics for human observers. They showed some of the major components of a chess puzzle that people would see as beautiful. This is interesting and innovative work on the *beauty* of chess, and related to, but not the same as, my concern here; which is the potential for *creative play* in chess. Let us turn to a couple of example chess puzzles or "compositions" that are beautiful, but also can be called creative.

In a composition, a strong player (such as a chess Grandmaster) sets up a position on the chessboard and challenges us to find the winning play. An example is shown here, in Fig. 1, with "white to play and mate in two moves." The composition is by the famous chess player Susan Polgar, who was a child prodigy and the first ever female player to become a full Grandmaster in her own right.

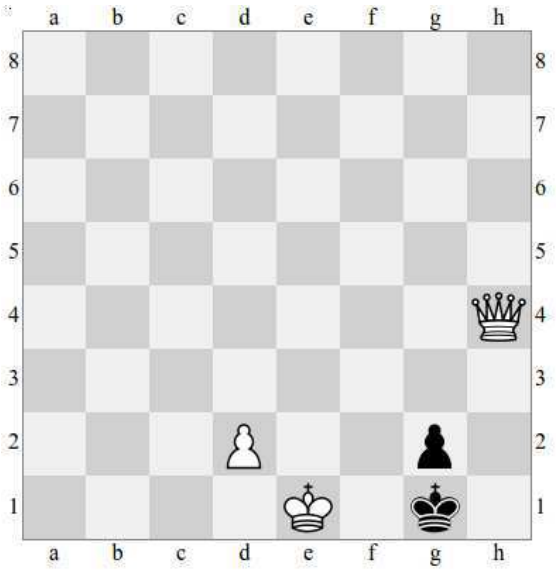


Figure 1. White to play and mate in two moves. From: [8].

A more complex composition, in Fig. 2, (from [11]) is also by Susan Polgar.

This is quite a difficult puzzle, which Polgar has specifically asked people to try to solve themselves, without using the help of a computer. The author of the article is a chess columnist, who loves chess compositions, but took a whole evening to solve this one. The solutions to both of these puzzles are in the next section, in case readers wish to try to solve them on their own first. That will help to give a sense of any creativity needed or involved in solving the compositions.

In both cases, the common characteristics of good chess compositions are on show. The puzzles are difficult to solve, intriguing because the obvious attempts are not correct, and therefore contain an element of misdirection. It is as if the composer anticipates the thought processes of the solver and baffles them. To solve such puzzles quickly is therefore an impressive feat, and shows some deeper understanding of the chess positions.

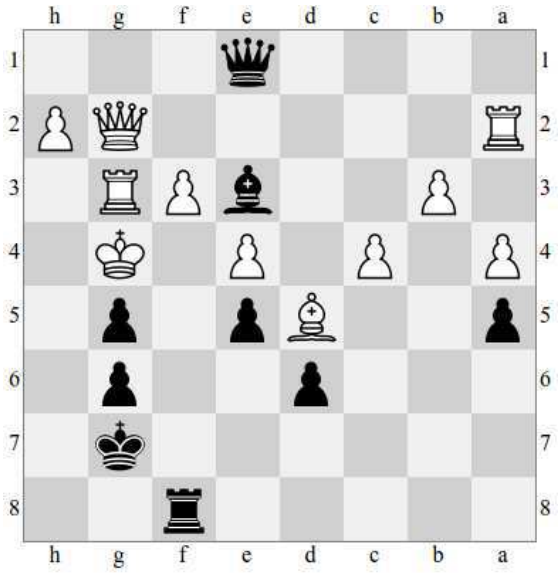


Figure 2. Black to play and mate in three. From: [11].

The upshot is to create a feeling of surprise in the solver, when the solution is finally shown; or else if the solver finds the solution himself, there is a feeling of satisfaction, and appreciation of the artistry in the composition if it is a good one.

3.1 Computer performance on the puzzles

While it takes a human player some time to solve the puzzles, computer programs can solve them much quicker. To illustrate this, a modest but convenient computer player was tested with both puzzles (available at http://www.apronus.com/). It runs in a Javascript browser, and was timed on a small notebook computer with only 1GB of RAM memory and a 1.6GHz Intel Atom CPU.

The first puzzle is relatively easy, and a fair player might find the solution in well under a minute. The computer found the solution in 200ms. (It is to move the white king away from the black king, giving him space to move out, which is then his only option; but luring him into a trap. The queen swoops down next to him and it's checkmate. `1. Kd1, Kf1. 2. Qe1 #`).

The second puzzle is more serious, and even most Grandmasters would probably take at least five to ten minutes to solve it. The same computer took only 700ms. Weisenthal gives a nice walk-through of the thought processes of a typical player trying to solve the puzzle, which even a novice player could follow. He shows how such compositions are constructed to mislead and tease the solver [11]. (The trick is to see the second move, which is a relatively quiet one, not suggesting itself to the typical chessplayer; and that white is then oddly helpless against the quiet threat. `1. ..., Rf4. 2. K x g5, Bb6. 3. ..., Bd8 #`).

3.2 Assessment of the computer's creativity

Can we say that the computer algorithm that solved the two compositions is creative? Well it finds the correct solution, which it did not know beforehand, so its product is both novel (to itself) and valuable. Indeed the computer is exactly as creative as any human solver by this reckoning; but as the computer is so much faster, it is that much more "creative", in the terms given above.

What about the extra criteria of creativity mentioned earlier, namely that os *surprise*? The surprise is built into the puzzle by the composer, in the sense that it was designed to have a non-obvious solution that would thus be hard to find. This property is again equal for both computer and human solver; but again the computer's great speed tells in its favour.

Objectively then, by the criteria of creativity laid out in this paper, and on the results of this limited test of two puzzles, the computer is more creative than any human expert player.

That may be an astonishing and unwelcome conclusion for some readers, especially given that the chess algorithms were never written in order to specifically address the question of computational creativity in the first place.

3.3 Possible objections and resolution

One common objection to this claim of computational creativity will be to complain that computers and only calculating their way to a solution. In this case they are executing a "brute force" search. This is an appropriate term for chess algorithms, and indeed it is exactly how it was envisaged from the beginning of AI by founders like Shannon and Turing that computers would come to play chess. The ironic wit in the term is deliberate — the computer is displaying only a brute form of intelligence, and yet with such power that it gives an uncanny impression of genuine intelligence.

This objection of brute force, or of mere calculation, is a classic objection to AI in all its forms, and is immediately persuasive to ordinary people, as well as many experts. However, it is not quite fair as a supposedly unfavourable comparison with human cognition, for the following reasons at least:-

1. computer "cognition" is apparently very different, but that does not make it necessarily inferior or worse. To assume that anything different from us must be inferior is characteristic of racism and xenophobia, and is outwith science.
2. human cognition is itself not well understood in any case. This makes it too tempting to overstate any claim that other cognition is different from it, without having any solid basis.

While it is true that we feel that our human thought processes are often intuitive, and not to be explained, they are also successful at the same time. This gives our own creativity a mystique that we cannot attribute to algorithms once we understand how they work. But again, to rest on a vague concept like "intuition" as the key distinction between two supposedly different kinds of cognition, seems too hasty and unsound.

4 Conclusion

Starting from a commonly shared notion of what creativity is, we have taken a tour through some chess puzzle territory, to explore the possibility that chess algorithms might be good models for computational creativity. We found that computer performance in this respect is high, and that we are thus bound to accept that computers are creative, or else we have to re-examine our conceptions and definitions of creativity.

Computers in this domain can easily exceed human performance, which is already a contribution to the field of computational creativity. However the main intention of this paper is to establish the viability and even suitability of computer games, with chess as an example, as a research domain for the field. It appears in conclusion that this potential may have been generally underestimated to date. Reasons for this might include a general prejudice against rational reasoning as being creative; or against computers especially. But whatever reasons for it there may be, the point remains that computers and algorithms, as game players and puzzle solvers (not only composers), are not yet fully appreciated by the field, which continues to devote more attention to the arts. As the area of games and puzzles is more tractable however, for evaluation especially, we should expect better progress with this as a research domain.

5 Acknowledgements

I would like to thank the two anonymnous reviewers for their comments and suggesting a couple of references.

REFERENCES

[1] Margaret A. Boden, 'Creativity and artificial intelligence', *Artificial Intelligence*, **103**, 347–356, (1998).
[2] Margaret A. Boden, 'Computer models of creativity', *AI Magazine*, **30**(3), 23–34, (Fall 2009).
[3] Michael Cook and Simon Colton, 'ANGELINA – coevolution in automated game design', in *Proceedings of the Third International Conference on Computational Creativity*, p. 228, Dublin, Ireland, (may 2012).
[4] J. Huizinga, *Homo Ludens: A Study of the Play-element in Culture*, Beacon paperbacks ; 15 : Sociology, Beacon Press, 1955.
[5] Bipin Indurkhya, 'Whence is creativity?', in *Proceedings of the Third International Conference on Computational Creativity*, p. 62–66, Dublin, Ireland, (May 2012).
[6] Azlan Iqbal and Mashkuri Yaacob, 'Advanced computer recognition of aesthetics in the game of chess', in *WSEAS Transactions on Computers*, p. 497–510, (May 2008).
[7] Antonios Liapis, Héctor P. Martínez, and Georgios N. Julian Togelius, 'Computational game creativity', in *Proceedings of the Fifth International Conference on Computational Creativity*, (2014).
[8] "Susan Polgar", Wikipedia, 2015, January 12.
[9] Claude E. Shannon, 'Programming a computer for playing chess', *Philosophical magazine*, **41**(314), 256–275, (1950).
[10] "Alan Turing", Pearson; and the A.M.Turing Trust, 1953.
[11] Joe Weisenthal. Here's a delicious chess problem that had me scratching my head for hours. Business Insider website. http://www.businessinsider.com/black-to-move-and-mate-in-3-problem-2013-8, August 14 2013.
[12] Geraint Wiggins, 'Searching for computational creativity', *New Generation Computing*, **24**(3), 209–222, (2006).

Towards a Computational Theory of Epistemic Creativity

Jiří Wiedermann[1] and **Jan van Leeuwen**[2]

"The creative act is not an act of creation in the sense of the Old Testament. It does not create something out of nothing: it uncovers, selects, re-shuffles, combines, synthesizes already existing facts, idea, faculties, skills. The more familiar the parts, the more striking the new whole.
A. Koestler [9]

Abstract. We investigate the computational process of creativity from the viewpoint of our recent thesis stating that computation is a process of knowledge generation. Rather than considering the creativity process in its full generality, we restrict ourselves to so-called epistemic creativity which deals with the processes that create knowledge. Within this domain we mainly concentrate on elementary acts of creativity — viz. drawing analogies. In order to do so using the epistemic framework, we define analogies as certain relationships among linguistic expressions and we state what knowledge must be discovered in order to resolve a given incompletely specified analogy. We assume analogies are formed in a natural language and also require that a solution of each analogy must contain an explanation why the resulting analogy holds. Finally, the difference between non-creative and creative computational processes is discussed. Our approach differs from the majority of previous approaches in stressing the knowledge discovery aspects of computational creativity, in requiring explanations in analogy solving and, last but not least, in including theory-less domains serving as knowledge base for knowledge discovery process.

1 INTRODUCTION

Creativity is an activity producing knowledge in the form of ideas, artifacts or behavior that is new for its creator and in some way valuable or important for him or her. Without creativity, no artificial system can aspire to be on par with human intelligence. In its most developed form creativity permeates all human activities. It has been subject of studies in many academic disciplines, among them in psychology, cognitive science, education, philosophy (particularly philosophy of science), technology, theology, sociology, linguistics, economics, and in arts. While all of these disciplines have defined creativity according to their own paradigms and needs, hardly any of them made a serious effort to reveal the underlying mental mechanisms supporting and enabling the process of creativity. This is perhaps due to the fact that the anticipated nature of these mechanisms has been assumed to lay outside of the disciplines at hand. But there is one exception to this rule, and this is the field of artificial intelligence, and especially artificial general intelligence (AGI). Mechanisms of artificial creativity have been intensively studied in cognitive science as well. Due to its omnipresence in many fields of study, the literature concerning creativity is immensely rich and too extensive to be discussed, summarized or referenced fully here.

When inspecting definitions of creativity in whatever discipline, AGI included, two things strike the eye: first, the definitions are very informal, given in a natural language, and second, the definitions hardly ever mention the term knowledge. Especially the latter fact is quite surprising since, perhaps with the exception of artistic creativity, the ability to create new knowledge permeates all domains of creativity. In such domains the primary purpose of creativity is to generate or to demonstrate new knowledge in whatever form — be it conventional knowledge used in everyday life, or scientific knowledge, or a skill, behavior, or a "materialized knowledge" (i.e., knowledge embedded into objects, their functioning, shape or appearance). This kind of creativity is called *epistemic creativity*. Mokyr (cf. [11]) describes it as "actually creating new knowledge or combining existing fragments of knowledge in altogether new ways", as part of his more general view of productive creativity. How can the functioning of epistemic creativity effectively be understood?

It is true that the research field called "knowledge discovery" has become quite popular since the 1990s. Knowledge discovery describes the process of automatically searching large volumes of structured (databases, XML) and unstructured (text, documents, images, multimedia) data for patterns that can be considered knowledge about the data. When compared to what one expects from epistemic creativity, the field of knowledge discovery, despite its name, merely extracts knowledge about the data without having the ambition to create new knowledge other than that which can be straightforwardly extracted from data. This, by the way, can be illustrated by the fact that in the research papers in this field, the word "creativity" is used quite rarely.

It is also true that at the intersection of the fields of artificial intelligence, cognitive psychology, philosophy, and the arts there is a flourishing multidisciplinary endeavor called "computational creativity" (also known as artificial creativity or creative computation). Its roots go back to the nineteen sixties. The field is concerned with theoretical and practical issues in the study of creativity. Here the situation seems to be fairly opposite to the previous case: while the field teems with the word "creative" and all its derivatives, the notion of "knowledge" is much less frequent here. In part this could be due to the fact that the field very often seeks its inspiration in artistic creativity. The field is looking for its theoretical foundations.

In our opinion this frequent overlooking of the connection between (computational) creativity and knowledge generation - where the latter obviously is the main sense of epistemic creativity - may have been caused by an insufficient understanding of what compu-

[1] Institute of Computer Science of CAS and Czech Institute of Informatics, Robotics and Cybernetics of CTU, Prague, Czech Republic email: jiri.wiedermann@cs.cas.cz
[2] Center for Philosophy of Computer Science, Utrecht University, the Netherlands email: J.vanLeeuwen1@uu.nl

tation is. In our recent works [17],[18], [19], [15] we coined the idea that the classical view of computation, based on the ways information is processed by all sorts of machine models (typically by Turing machines), prevents us from clearly seeing the main purpose of computations. The classical view favors the view of HOW computations are performed, instead of WHAT they are doing, i.e. of what is their sense. We hold the view that computation is any process of *knowledge generation*, as we have demonstrated in our previous works. Note that the notion of knowledge generation is machine-independent: we are not interested how, by what means, knowledge is generated, be it in a serial, parallel, interactive, or any other way. What counts is what knowledge is generated.

Changing the view of what computation is may have dramatic consequences. For instance, in the past, various authors have argued that cognition is not computation (cf. [3], [14]), where they have viewed computation in its classical sense, through classical models and scenarios of computations. Under the new view, cognition becomes knowledge generation, and thus, computational, independently of the underlying machine model and computational scenarios. The previous problem vanishes thanks to a new apprehension of computation.

Seeing computations as knowledge generation processes does not automatically turn every computation into a creative process. Intuitively, epistemic creativity requires more than producing knowledge according to some rigid schema (program), counting with some fixed number of alternatives each of which corresponding to a certain prespecified circumstance. For creativity, we require more: new, original alternatives (pieces of knowledge) satisfying as many required constraints as possible must be discovered within the existing knowledge and combined in a novel way under whatever circumstance that cannot be known beforehand. From the candidate alternatives, the one best fitting the constraints must get chosen. This leads to a computational view of epistemic creativity.

The ideas described in the last paragraph answer the often posed question why people have ideas and computers don't. The reason why computers are not creative can have two reasons. The first one is that in the majority of cases when an average person is using a computer, creativity is not required by the application (e.g., in looking for a train schedule). The second answer concerns the so-far quite rare cases where creativity is required — e.g., when consulting symptoms of a disease, or asking for a nice analogy. In such situations a computer will probably not be as creative as we would like to see because it is programmed without understanding how creativity works and what its prerequisites are. Nonetheless, the essence of epistemic creativity has been described in the last two sentences of the previous paragraph. Can we say more about the respective creative processes? Can we be more specific in describing which knowledge generating processes can be seen as creative processes? What are the prerequisites for computational creativity? (Note that we are using the term "computational creativity" in a new, broader sense than mostly used in the eponymous research field.)

In this paper we will answer the last three questions from the epistemic viewpoint of computations. As it turns out, answering the last three questions in their full generality is not easy. Therefore, in what follows we will first investigate but a specific case of creativity. We will concentrate on one of the simplest cases of creativity, and this is analogy solving. Solving an analogy can be seen as an elementary creativity act that calls for discovering and displaying new relations between known pieces of knowledge. Then we will extend our study to a general case of new knowledge discovery.

The structure of the paper is as follows. In Section 2 we present our view of computation as knowledge generation that will offer a unified framework for our further consideration of computations. Special attention is paid to computations in theory-less domains corresponding to natural languages. Section 3 contains the main contribution of the paper. After some preliminaries in Subsection 3.1. analogies and their formal definition in the epistemic framework is presented in Subsection 3.2. The "hard to vary" principle is described, enabling a "quality" judgment of explanatory analogies. In Subsection 3.3. metaphors and allegories as variants of analogies are considered. Subsection 3.4. deals with the efficiency issues in analogy solving. The entire Section 4 is devoted to the general problem of knowledge discovery. Finally, Section 5 contains a general discussion, also paying attention to the difference between creative and non-creative knowledge generation. Conclusions are given in Section 6.

The contribution of the paper to the present state of the art of the theory of computational creativity can be seen in several planes. First, the epistemological view of computations offers a natural unified framework for studying problems related to epistemic creativity. Second, this framework, being machine independent, allows the consideration of theory-less knowledge domains. Third, pertaining to analogy solving, the requirement for a computation to be accompanied by evidence that it works as expected is mirrored in the definition of analogy by a similar demand for analogy explanations. Fourth, explanations attached to each solution of explanatory analogies allows one to judge their explanatory power via the "hard to vary" principle. Finally, our considerations shed further light on the general problem when a computational process is a creative process.

2 COMPUTATION AS KNOWLEDGE GENERATION

Viewing computation as knowledge generation as described in [17], [18] and [19], requires certain ingredients that we first describe informally.

Knowledge in our framework is knowledge in the usual sense of this word. This, of course, does not look like a definition of knowledge, but we need not be very specific. For illustration purposes only, we cite the following definition from Wikipedia: *Knowledge is a familiarity with someone or something, which can include facts, information, descriptions, or skills acquired through experience or education. It can refer to the theoretical or practical understanding of a subject. It can be implicit (as with practical skill or expertise) or explicit (as with the theoretical understanding of a subject); it can be more or less formal or systematic.* Obviously, knowledge according to this definition is observer–dependent.

Any knowledge is a part of a so-called *epistemic domain,* or *domain of discourse,* corresponding to the kind of knowledge we are interested in. Such a domain can be given formally — as in mathematical or logical theories (e.g., theory of recursive functions) or entirely informally, in a natural language, as all sentences describing phenomena in a real world. Intermediate cases (like physical, chemical or biological theories) described in part formally and in part informally are also acceptable. In any case, we must have means to describe the so-called *pieces of knowledge* (e.g., axioms, sentences or formulae in formal theories, or words and linguistic expressions in informal theories described in a natural language).

The final ingredient we require are so-called *inference rules* applicable to the pieces of knowledge in a given domain allowing constructing, generating new pieces of knowledge that will still belong to the domain at hand. Again, in the case of formal theories these rules are also formal rules (like deductive rules in logic), but we also

allow entirely informal ones, corresponding to "rational thinking" in the case of informal theories.

The epistemic domain together with the corresponding inference rules form the *epistemic theory*.

Each computation we will consider will generate knowledge from some epistemic domain with the help of the corresponding computational process. We will say that such a computation will be *rooted* in this domain. Starting from the so-called *initial knowledge* the computational process will generate *output knowledge* within the given epistemic domain. Depending on the epistemic domain, initial knowledge is given in the form of axioms, definitions, observations, facts, perceptions, etc. The output knowledge may take the form of propositions, theorems or proofs in the case of formal theories, and statements, hypotheses, scientific laws, or predictions in the case of natural sciences. In the case of informal theories (like theory of mind, arts, etc.) the generated knowledge takes the form of conceptualization, behavior, communication, utterances in a natural language, thinking, and knowledge about the world formed mostly in a natural language or in a form of scientific theories and other writings.

From what has been said above one can see that the epistemic domains range from so-called *theory-full* domains corresponding to formal, abstract theories to *theory-less* domains that admit no formal descriptions for capturing e.g. behavior in common life situations (cf. [13]).

In order for a computation to generate knowledge there must be evidence (e.g., a proof) that explains that the computational process works as expected. Such an evidence must ascertain two facts: (i) that the generated knowledge can be derived within the underlying epistemic theory, and (ii) that the computational process generates the desired knowledge.

The latter is the key to the following more formal definition (cf. [18]). In this definition we assume that the input to a computation is part of both the underlying epistemic domain (and thus of the theory) and the initial data of the computational process. Do not forget that although the notation used in the definition formally resembles the notation used in the formal theories, we will also be using it in the case of informal epistemic domains.

Definition 1 Let T be a theory, let ω be a piece of knowledge serving as the input to a computation, and let $\kappa \in T$ be a piece of knowledge from T denoting the output of a computation. Let Π be a computational process and let E be an explanation. Then we say that process Π acting on input ω generates the piece of knowledge κ if and only if the following two conditions hold:

- $(T, \omega) \vdash \kappa$, i.e., κ is provable within T from ω, and
- E is the (causal) explanation that Π generates κ on input ω.

We say that the 5-tuple $C = (T, \omega, \kappa, \Pi, E)$ is a computation rooted in theory T which on input ω generates knowledge κ using computational process Π with explanation E.

When considering epistemic creativity in the sense of human mental ability, one usually thinks of it in the context of a natural language. How could the corresponding computation (seen as knowledge generation) be captured by the above definitions?

First of all, one must bear in mind that the underlying knowledge domain is a domain comprising, in principle, all human knowledge. This knowledge can be seen as a union of various specific knowledge domains which vary from theory-full to theory-less domains. The respective knowledge is thus *heterogeneous knowledge* and natural language serves as an important, and in fact, the only one known mediator among the respective theories. The less formal the knowledge is the more it relies on the natural language. The "inference rules" for heterogeneous domains are a mix of informal and formal rules. That is, when one speaks within theory-less domains, the informal rules of "rational thinking" are used. Otherwise, speaking within theory-full domains one makes use of the rules corresponding to that domain. Natural language provides not only a tool for initial forming and describing a theory, it also provides a unified tool for understanding all theories and "moving" among them. Last but not least, natural language and its semantics provide a link between a theory and the physical world. Only due to natural language and only within a theory one can explicate meaning of the expressions of a natural language, i.e., their semantics. Namely, in our framework the meaning of any expression of a natural language is given by knowledge pertinent to this expression within a certain domain of discourse. This knowledge comes again in the form of a theory stating all contexts and relationships among them in which the expressions at hand can be used. That is, this theory captures the ways in which usage of an expression makes sense in various contexts. Semantics is knowledge and therefore it can be generated by a computation. From this viewpoint all computations, including the computations that generate knowledge based on understanding natural language, bear a homogeneous structure despite the fact that the underlying knowledge as a whole covers many epistemic domains.

The knowledge framework behind a computation over the domain of a natural language will normally be based on *cooperating theories*. This is an extremely complex system since in principle to each word a theory (in our general sense) is attached, controlling the proper use of this word. In general, such a theory depends not only on the word at hand, but also on the context in which the word is being used. In the case of embodied cognitive systems the context does not only refer to the grammatical context, but also to the entire perceptual situation (cf. [16]). All this leads to a complex intertwining of the respective theories working of the internal models of the world. If realized along the lines sketched above, the underlying cooperative theories should display understanding. The problem of understanding is the central problem of AGI and our approach to computation seems to offer a versatile tool for capturing the related issues. This is because it concentrates on the specification of WHAT the sense of understanding is, while postponing the questions HOW this can be realized. Nevertheless, it is fair to state that so far we do not know much about cooperating theories leading to computational understanding.

Second, what computational process is behind a natural language? It is the process running in our heads. Although we do not know the details of how it works, we do know that it generates knowledge that we can describe by natural language as indicated above. And finally, what corresponds to the explanation? Again, it is an explanation in a natural language.

To summarize, we see that natural language is used here as a means for describing the underlying theory-less domain and the inferences over such domain, as well as for explaining the respective computations as performed by the human brain. Note the analogous situation in classical computing where, for example, λ-calculus is used both as a programming language and as the underlying model of computation.

3 COMPUTATIONAL CREATIVITY

3.1 Preliminaries

Any computation as defined in the previous section generates knowledge. Nonetheless, as remarked in the Introduction, this does not

necessarily mean that any computation should be seen as a creative process, as a process that generates something new, original, unexpected, surprising, deserving a special interest or having some worthy value as required in epistemic creativity. This "surprise effect" does not happen when an output of a computation can routinely be produced in a straightforward way, following pre-programmed paths corresponding to a priori envisaged circumstances. The majority of current computer programs works in this way. Typical examples include the computation of a function. Such a process can be seen as generating explicit knowledge (i.e., a function value corresponding to the input value) from implicitly described knowledge that is given in the form of an algorithm. There is no room for creativity in such a process. Note that, e.g., various editors and spreadsheets belong to the category of such computations. Operating systems can serve as an example of an interactive non-creative computational process. What they do can be subsumed as an iteration of the following activity: *"if so and so happens, do so and so"*. In computations of this kind no creativity is assumed, since it is not required by the applications at hand.

What about database searches? Here, pieces of knowledge are sought by searching a finite amount of data ("knowledge items") using a specified criterion. Is here some room for creativity? Now the answer is not so simple as in the previous case. In "old fashioned" databases as used in the early days of computing that used to seek an item satisfying a certain condition within the set of structured data, the situation was similar to the previous case. But think about the following case: a "database" (or rather: a knowledge base) containing all knowledge possessed by an average person (whatever it might mean), i.e., knowledge contained in the mind of that person. The query would be as follows: *"name me an animal living in a desert having the same relation to its living environment as has a shark to the ocean"* (the example taken from [16]). In this case, we can obtain an answer "I don't know" (e.g., from a child), or "a camel" (from an average educated person), or "a desert lion" (from an informed animal rights activist), or even "Cataglyphis bicolor" (a desert-dwelling ant also called "the Sahara desert ant"), from some joking entomologist. Now, were there some aspects of creativity in delivering any od these answers? Which of these answers is the best? And, last but not least, what was the mechanism enabling the answering of such a query?

3.2 Analogies

The last example has been an example of analogy solving. Discussions and studies of analogies go back to the ancient philosophers, since analogies have always played in important role in reasoning in logics, science, law and elsewhere. The role of analogy has been intensively studied for years in cognitive science (cf. [8], [10]). The notion of analogy is rarely formally defined. What one can find in the literature, vocabularies and on the web are informal definitions serving to the purpose of the underlying field. Thus, one can find definitions like *"analogy denotes a similarity between like features of two things, on which a comparison may be based;* or *"a comparison between one thing and another, typically for the purpose of explanation or clarification"*, or *"analogy is a figure of language that expresses a set of like relations among two sets of terms"*. In logic, *"analogy is a form of reasoning in which one thing is inferred to be similar to another thing in a certain respect, on the basis of the known similarity between the things in other respects"*.

There are many variants of analogies. For the purpose of knowledge generation we will be especially interested in so-call *explanatory analogies*. Such analogies create understanding between something unknown by relating it to something known. They provide insight or understanding by relating what one does not know with what one knows. Thus, these analogies may be seen as providing elementary creativity steps in deriving new knowledge. This approach where knowledge is not obtained by simply composing pieces of old knowledge has to be contrasted with the classical epistemological procedures of knowledge generation. Such procedures are usually described as extrapolations of repeated observations, or of known facts, as some variants of an induction process. In this process, there is no creativity aspect: knowledge is merely transformed from one form to an other. However, it is reasonable to expect that the ability to create new knowledge must also include the ability to create new explanations, not merely extrapolating or generalizing the past experience.

In order to better understand explanatory analogies, we will need a more formal definition of analogy that will enable us to see the finer details of the envisaged computational process of creating knowledge leading to analogy solution. Therefore, for our purposes the desired definition should fit into the framework of epistemic computations.

The starting point will be to choose a suitable theory in which the respective computations will be rooted. In this respect, note that all informal definitions of analogies involve direct or indirect reference to natural language. Moreover, they are using linguistic expressions like features, relations, similarity, comparison, or explanations. Therefore, a natural choice for such a theory would be a natural language $\mathcal{NL}$ possessing the richness of linguistic expressions needed to understand and resolve analogies. The (informal) rules corresponding to $\mathcal{NL}$ would be those of "rational thinking", and the corresponding computational process will be that produced by the human brain (cf. Definition 1) and the discussion thereafter.

In the following definition (taken from [16]) the adjective *linguistic* will mean that the corresponding expressions, predicates or relations are not described in any formal logical calculus or theory — rather they are described by expressions of a natural language $\mathcal{NL}$ corresponding to the respective pieces of knowledge. These pieces of knowledge form the *knowledge base* of $\mathcal{NL}$. Their validity usually cannot be proved formally but can be known from experience, empirically or from hearsay.

Definition 2 *Let $S = (s_1, \ldots, s_k)$ and $T = (t_1, \ldots, t_k)$ be two sequences of linguistic expressions from $\mathcal{NL}$. If there exists a linguistic k-ary predicate $P \in \mathcal{NL}$ such that both $P(S)$ and $P(T)$ hold and linguistic relations $R_1, \ldots, R_k \in \mathcal{NL}$ such that $R_i(s_i, t_i)$, for $i = 1, \ldots, k$ holds, then we say that S* is analogous to *T w.r.t. predicate P and relations $R_1, \ldots, R_k$.*

Parameters $s_1, \ldots, s_k$ and $t_1, \ldots, t_k$ are called attributes *of S and T, respectively. Relations R_i's are called* similarity relations*.*

Note that the linguistic expressions, predicates and relations are all described as expressions of a chosen natural language $\mathcal{NL}$.

Definition 3 *Using the notation from Definition 1, given S and T,* analogy solving *is a knowledge generating process whose purpose is to find linguistic predicate P and linguistic relations $R_1, \ldots, R_k$ such that S is analogous to T w.r.t. predicate P and relations $R_1, \ldots, R_k$.*

We say that P is a conjecture *and $P(S)$, $P(T)$ and $R_i(s_i, t_i)$ are the* explanation *of this conjecture.*

To illustrate the use of the introduced formalism, consider again the example from Subsection 3.1. Excerpting from [16]: If $S =$

$(shark, ocean)$ and $T = (camel, desert)$, then we may define predicate $P(x, y)$ as"x lives in y" and R_1 as "both *shark* and *camel* are animals", R_2 as "both *ocean* and *desert* are living environments". Then the claim "x lives in y" is the conjecture and the facts that "*camel* lives in a *desert*" , "*shark* lives in *ocean*" , "both *shark* and *camel* are animals" and "both *ocean* and *desert* are living environments" are its explanation. The previous task is often described as "the relation of *shark* to *ocean* is like the relation of *camel* to *desert*" and abbreviated as *shark : ocean :: camel : desert*.

If all expressions in S are known and only some expressions from T are missing, then S is called the *source* and T is called a *target* of the analogy. Then the whole analogy inclusively of its explanation can be seen as an *explanatory analogy.* The task of finding both the conjecture and its explanation is an act of knowledge discovery. This is because in general the predicates corresponding to the conjecture and the explanations must be discovered among the pieces of knowledge that are at one's disposal.

We have already noted that an explanatory analogy might admit more than one solution. For instance, the solution of analogy *shark : ocean :: ? : desert* could have been either a camel, or a desert lion, or a Sahara desert ant. Under some circumstance, the answer "I don't know" could also be correct. In order to judge the quality and validity of an answer, we must also know the respective explanation. If all explanations are evaluated by an observer as valid, then what answer is the best? In such a case, the best answer would be the one which maximizes the number of relations between the source and the target (i.e., maximizes number k in Definition 2). For instance, in our case, the answer "desert lion" is to be preferred, because in addition to similarity relations R_1 and R_2 it also satisfies relation R_3 "both $shark$ and $desert\ lion$ are predators". The more similarity relations the candidate solution of the incomplete analogy satisfies, the harder it is to come with a different solution. We say that the solution at hand is "hard to vary". According to Deutsch [6], such a solution has a better "explanatory power" than the other competing solutions.

The multitude of answers points to the fact that the answer is observer dependent. The "less knowledgeable" observer might not know about the existence of desert lions and therefore the answer "camel" would sound more plausible to him or her. An observer not knowing any animal living in a desert obviously must answer "I don't know".

3.3 Variants of analogies

Analogies also occur in a number of different forms which can be seen as generalizations or specific cases of our definition of analogy. Let us mention but a few of such instances of analogy.

A more general case is the case of so-called *incomplete analogies*, in which one has to find an analogy between two (or even more) linguistic notions S and T but not all (possible none of the) attributes of neither notion are given. That is, a part of a solution must also be the discovery of the respective attributes of T and S whose pairs correspond to the similarity relations, and the maximization of the number of such pairs. Such problems occur, e.g., in taxonomy dealing with classification of things or concepts based on sharing similar features. In such cases the degree of creativity seems to be higher than in the cases described by Definition 2.

In the opposite direction, a metaphor is a special type of analogy. A *metaphor* is an expression of language that describes a subject by comparing it with another unrelated subject resembling the original subject only in some semantic aspects, on some points of comparison. Both subjects then share the same semantic property which is not immediately apparent from the names of both subjects (cf. the metaphor *"time is money"*) (cf. [10]).

An extended metaphor is *allegory,* in its most general sense. Allegory has been used widely throughout the histories of all forms of art, largely because it readily illustrates complex ideas and concepts in ways that are comprehensible to its viewers, readers, or listeners. Allegories are typically used as literary devices or rhetorical devices that convey hidden meanings through symbolic figures, actions, imagery, and/or events, which together create the moral, spiritual, or political meaning the author wishes to convey (cf. Wikipedia). Re-casting allegory into our framework, allegory usually establishes similarity relations between the narrative story and its possible interpretations in a real or imaginary world. Discovering such relations is a task for allegory creation as well as their projection into the solution of the allegory at hand. The idea is that this projection is not usually obvious at the first sight and its discovery is a task for the observer. In this sense, the similarity relations are "indirectly defined" and depend on the individual taste and knowledge of the observer. Aesthetics and emotions can play an important role in this process. In this way, both creating an allegory by its creator as well as its "deciphering" by an observer are creative acts.

Finally, let us mention the most general and the most important case that plays a crucial role in scientific discovery, and this is the case of the resolution of a "flaw" in a theory. In our framework (cf. Definition 1), the scenario of such a situation is as follows: consider theory T working well over some epistemic domain until one day an input ω to T is found delivering output κ_1. This output is different from output κ_2 which for some reason was expected (e.g. κ_1 disagrees with observations or with experiments). Now the question is, what is the minimal adjustment of theory T such that it would predict output κ_2 on input ω while retaining its ability to work correctly for all other inputs? Clearly, this is another variation on the theme of analogy. This time however, the epistemic theory T itself has become the source and the new theory T' the target of the analogy, and one has to invent new attributes of the target theory preserving as much of the old theory as possible while repairing its flow w.r.t. input ω. Of course, it may happen that theory T is "irreparable" and T' will be completely different from T. History of science knows a lot of such examples (cf. the clash of Darwinism and creationism).

3.4 Efficiency issues in analogy solving

In the framework of epistemic computations one cannot speak about complexity of computations in the classical sense. This is because, in this case, no concrete computational model is used. What can be done for a computation generating complex knowledge, is to describe what partial knowledge or pieces of information are needed in order to generate the knowledge.

Consider the case of solving an explanatory analogy in the form as described by Definition 2 and 3.

The input knowledge for our computation consists of two linguistic predicates $S, T \in \mathcal{NL}$ from a natural language $\mathcal{NL}$, respectively, with $S = (s_1, \ldots, s_k)$ and $T = (t_1, \ldots, t_k)$. Since we are dealing with explanatory analogy we will assume that some (but not all) attributes t_i's in T are left unspecified. Let $\mathbb{K}$ be the knowledge base that is at the disposal of our computation.

In order to resolve such analogy, we need to discover the following knowledge:

1. we have to check whether an object corresponding to predicate S does exist in $\mathbb{K}$. If not, the answer would be *"I don't know"* and we are done.

2. for each object $T' \in \mathbb{K}$ satisfying predicate T in specified attributes, we check whether in $\mathbb{K}$ there exists

   (a) a k-ary predicate P satisfying $P(S) = P(T')$ (i.e., we are looking for a conjecture). If there is no such P the answer would again be *"I don't know"* and we are done.

   (b) next, we look for linguistic relations of similarity $R_1, \ldots, R_k \in \mathbb{K}$ such that $R_i(s_i, t_i)$, for $i = 1, \ldots, k$ holds. If such relations are found then the answer would return object T', conjecture P and explanations R_i's. Otherwise the answer would again be *"I don't know"* and in either case, we are done.

   If no object T' is found then the answer is *"I don't know"* and we are done.

A more involved procedure would be needed in case the necessary knowledge is not found and we don't want to "give it up". If this happens then it is possible to consult "external sources" such as the web, encyclopedias, monographs, experts, etc. In any case, one can see that resolving explanatory analogies is a quite demanding task, requiring in the worst case knowledge of all items in the underlying knowledge base.

Can we say at least something about the computational complexity of solving explanatory analogies? Well, in any case, when we are dealing with analogies over finite knowledge bases, the previous "algorithm" of finding a solution (if it exists) solves in fact a combinatorial problem over a finite domain and therefore can be solved in finite time.

Obviously, the solution of an analogy problem, and in general, of any creativity task depends on all items in the underlying knowledge base. In order to address the essence of the problem of knowledge discovery in terms of the size of the underlying knowledge base we also use a metaphor, viz. the *metaphor of a mosaic*. Namely, a simplistic view of knowledge discovery is that we seek a piece or pieces of knowledge that fit into a certain unfinished mosaic composed from pieces of knowledge possibly from various domains. Here, "fit" means that the new pieces of knowledge are related to the existing pieces by a certain set of known eligible relations that can be either of a syntactic or a semantic nature. (Note that this was also the case of analogies and metaphors.) Then the *creativity problem* is the task of composing a solution of a problem from finitely many pieces of knowledge that have to be related in a logical way in order to come up with the desired solution. It is interesting to observe that a mosaic where only few pieces are missing can be seen as a hypothesis, or a conjecture. In the case of explanatory analogy solving the size of the mosaic is bounded by the number of attributes of both source and target predicate (parameter k in Definition 2). If n is the size of the knowledge base then solving the mosaic problem requires inspection of at most $\binom{n}{k} = O(n^k)$ subsets of the knowledge base. This means that for sufficiently large n and a fixed k the mosaic problem is of polynomial complexity and thus fixed parameter tractable in k.

A problem similar to the creativity problem — the so-called *domino problem* — has been studied in classical complexity theory (based on Turing machines). In 1966, Berger [2] proved that the domino problem is (classically) undecidable if the pieces of knowledge can be used an arbitrarily number of times. The basic idea of the proof is to have a mosaic to encode a halting computation of a Turing machine.

On the one hand, this explains the difficulty of finding new knowledge in general: there is no (Turing machine) algorithm solving such a task. On the other hand, solutions with a small number of pieces are relatively easy to find by a combinatorial search. It is interesting to note that the unrestricted creativity problem seems to be one of the few known undecidable problems of practical significance.

4 Discovering Knowledge

In [5] Barry Cooper asked, whether information can increase in a computation. Indeed, how could a computation produce information which has not already been somehow encoded in the initial data? This does not seem to be possible. An exhaustive answer to this problem has been given by S. Abramsky in [1]. He concludes that, while information is presumably conserved in a total (closed) system, there can be information flow between, and information increase in, subsystems. Note that in our definition of computation we have considered computational processes rooted in the underlying epistemic domains. This can be viewed as though computations are "observing" their "environments" as captured in their knowledge bases, and indeed, some of them even update the underlying knowledge or gain information from cooperating theories under an interactive scenario. In this case it is possible for such a computation to discover new knowledge.

More precisely, it is possible to go beyond the current knowledge explicitly represented in a knowledge base. This can be done by discovering new relationships among the elements of knowledge, or to discover an element or elements of knowledge that satisfy a required relationship to the existing pieces of knowledge, or to gain a new piece of knowledge from "external sources". By "discover" we readily mean to make something explicitly known, i.e., to obtain explicit knowledge of something for the first time. As an example of new knowledge one can take the resolution of a given analogy.

When speaking about creativity in the sense of knowledge generation one must take into account that knowledge can only be generated from knowledge — this is in fact the essence of our definition of computation. Thus, there exist two opposite processes related to knowledge processing: knowledge acquisition, and knowledge generation.

There are many ways of knowledge acquisition: by reason and logic, by scientific method, by trial and error, by algorithm, by experience, by intuition, from authority, by listening to testimony and witness, by observation, by reading, from language, culture, tradition, conversation, etc.

The purpose of *acquisition processes* is to let the information enter into a system and to order it — via computation — into the existing theory or theories over the pertinent knowledge domains (and represent it in a knowledge base). Such domains take various forms of conceptualizations which are part of the respective theory. A *conceptualization* is a simplified, abstract view of the world representing the given knowledge domain. It captures the objects, concepts and other entities and their relationships existing within the knowledge domain at hand (cf. Wikipedia). Obviously, any knowledge acquisition process builds and updates the existing theories.

The purpose of the *knowledge generation process* is to produce knowledge in reaction to the external or internal requests. One can distinguish two basic principles of knowledge generation: syntactic and semantic knowledge mining. Both methods make use of specific inference mechanisms whose purpose is to discover hidden patterns in the data.

Syntactic knowledge mining works solely over the data representing knowledge. It takes into account only the syntax of the respective data, not their meaning, and also the syntactic inference mechanisms of the underlying theory. Syntactic knowledge mining is the compu-

tational process of discovering patterns mainly in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Finding a pattern corresponding to a certain relationships among data not previously known certainly counts as knowledge discovery.

Semantic knowledge mining is the main engine of creativity. It also looks for hidden patterns in data (knowledge representations) which are semantically, rather than syntactically, related. Usually, based on the semantics of one pattern (the base) and a semantic relation an other pattern (the target), possibly in a different knowledge domain, with a similar semantic structure as the base pattern, is sought satisfying the required relation.

Very often the task of semantic knowledge mining is formed in a natural language. That is, the items to be sought and the relations to be satisfied are described in linguistic terms (as was the case of analogy solving). This complicates the searches since the meaning of linguistic terms must be known. The meaning of each term is, in fact, a piece of knowledge — a theory describing (the properties of) the notion at hand in various contexts. Thus, semantic knowledge mining calls for detecting similarities between theories usually related to different knowledge domains.

Discovering a, in a sense, "parallel" theory to a given one contributes to a better understanding of either theory since it enables to expect relations holding in one theory to also hold in its pendant theory. This is an important element of insight, explanation and understanding. Insight, understanding and explanation make only sense within a theory. They must follow from known facts and rational thoughts.

Unfortunately, general mechanisms of semantic knowledge discovery, explanations and understanding are largely unknown. What we have described here are but the first steps along the respective road.

5 DISCUSSION

We have already remarked at several occasions that, although we see every computation as a knowledge generating process, we cannot automatically consider any computational process to be a creative process. Stating this, when will a computation become a creative process? This seems to be a "million dollar question" of the entire field.

As an example, consider computing x^2, given x. Is this computation a creative process? Compare this with solving the incomplete analogy *shark : ocean :: ? : desert*. Where is the difference? We even know for both cases an algorithm leading to an answer. So where is a difference? Why is finding a solution of the former task considered to be a "non-creative" task whereas finding a solution of the latter is considered to be a creative task?

Well, there seem to be two important differences. In solving the first task one can compute directly with x and manipulate it as the computation requires. In the second case, the computation requires to discover other notions contained in the knowledge base, and the answer depends on what knowledge is stored in the knowledge database at that time. As we have seen, the solution of the second task need not be unique. And the second difference is in the complexity of explanations. While in the first case we must offer an explanation as required by Definition 1, in the second case, in principle, we must offer two explanations: one as asked by Definition 1 related to the correctness of the computation, and the second one required by Definition 2 concerning the correctness of analogy drawing. In general, is there a clear cut between non-creative and creative computational knowledge generation? Nevertheless, the extreme cases can be distinguished.

One might think that there is one more difference. Namely, that in the first case we do not need to know the semantics of x whereas in the second case it is necessary to know the semantics of the "parameters" of the analogy. But this is not true — both computations proceed without knowing the semantics of the respective notions.

Thus, as it appears, in creative knowledge generation (i.e., in computational creativity) the resulting knowledge depends, in addition to the discovery algorithm, on the contents of the underlying knowledge base. The result need not be uniquely defined and in some case need not be defined at all. The respective "creative computation" must work over whatever complete or incomplete knowledge base over the domain of the natural language at hand.

An aspect that seems implicit in "epistemic creativity" is that it isn't driven by the search for a pre-determined answer. In other words, creativity seems be synonymous with "unanticipated solution". In this context the underlying computational process is divergent since it leads to many answers, solutions, knowledge items even from domains that are not internal already but may be imported from elsewhere. Thus creativity seems to involve the generation of options that do not follow by mere deterministic reasoning. (If an artist has found a style that he can repeat, the question is whether it remains a creative process after the first time.)

The bottom line seems that a creative process is not a special type of computation to begin with but a whole collection of computations as also seen from the schema of resolving explanatory analogies in Subsection 3.4. This process is guided at best by some overall triggering process. This latter process may also hold a criterion for judging the computations or rather, the knowledge they come up with, a kind of objective function (which may not be well determined nor a "function" either). In the case of analogies we opted for the "hard to vary" criterion. In any case, this would mean that the question "when is a computation creative" is perhaps not the proper one to ask, if we accept that it is rather a complex process of "divergent computations".

Interestingly, in [12] the author attributes the difference between creative and non-creative *mental processes* not to the underlying computational/functional mechanisms, but rather to the way in which the mental process is experienced. This, however, throws no light on the nature of the underlying mechanisms.

In the context of computational creativity our analysis of analogy solving has revealed that the larger the knowledge base the greater the potential for discovering new knowledge. In order to have the knowledge base as large as possible it must potentially involve all the existing human knowledge and the creative agent must have a command of natural language in order to be able to navigate among various knowledge domains. Along these lines it appears that among the main obstacles of the progress in AGI is our insufficient knowledge of natural language processes concerned with the interactions between computers and human (natural) languages and representation of knowledge accessible to natural languages. Automatic procedures building the respective knowledge bases must be sought (cf. [16]).

Finally, one remark regarding the series of recent writings and interviews of one of the world top thinkers, the prominent British physicist David Deutsch (cf. [7]). At these occasions he has repeatedly stressed that *"The field of artificial (general) intelligence has made no progress because there is an unsolved philosophical problem at its heart: we do not understand how creativity works."*

In spite of what is known about computational creativity (cf. [4]) and despite of the enormous activities in this field, there is something in Deutsch's statement. What is still missing in all known ap-

proaches, are the phenomenal issues related to creativity. The "phenomenal component" of creativity seems to be required for a genuine understanding and realization of creative acts. In our approach we have covered up this problem by the requirement of a full mastering of the natural language. This appears to be impossible without engagement with issues around consciousness and free will, and this is why we have stressed the central role of natural language in epistemic creativity processes.

6 CONCLUSION

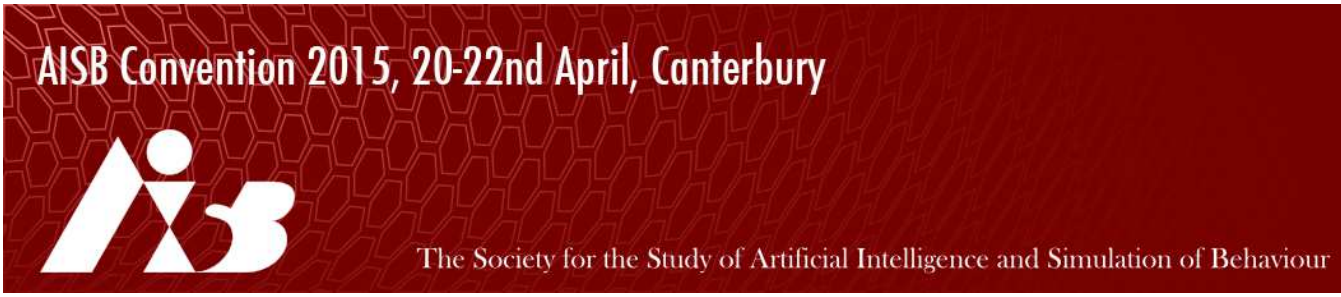
Our approach is consistent with the modern philosophical view accepted since ancient times that creativity is a form of discovery of new knowledge rather than some kind of inspired guessing. In this discovery process the role of natural language is indispensable since it serves as a universal language bridging various theory-less knowledge domains serving as knowledge base for a knowledge discovery process. Our approach to the problems of computational creativity via the epistemic view of computations offers a natural and uniform framework for the investigation of such problems. Under this view, computational creativity is simply seen as a specific kind of computational knowledge discovery in the underlying knowledge base. The richer the knowledge base the higher the potential for creativity is possessed by the corresponding computations. From this viewpoint, the classical, "non-creative" computational processes are but a special, in a sense "degenerated" kind of computations that do not make use of epistemic theories corresponding to knowledge domains described by explicit knowledge. The epistemic view of computations points to the full capability of computations by revealing their creative potential already in their very definition.

ACKNOWLEDGEMENTS

The research of the first author was partially supported by ICS AS CR fund RVO 67985807 and the Czech National Foundation Grant No. 15-04960S. We thank the anonymous reviewers for their insightful comments that have helped to improve the final version of our paper.

REFERENCES

[1] S. Abramsky. Information, processes and games. In J. van Benthem and P. Adriaans, editors, Handbook of the Philosophy of Information, pages 483549. Elsevier Science Publishers, Amsterdam, 2008
[2] Berger, R.: The undecidability of the Domino Problem, Memoirs of the American Mathematical Society 66, 1966.
[3] Bishop, J.M., (2009), A Cognitive Computing fallacy? Cognition, computations and panpsychism, Cognitive Computing 1:3, pp. 221-233
[4] Boden, M.: The Creative Mind: Myths and Mechanisms (Weidenfeld/Abacus & Basic Books, 1990; 2nd edn. Routldge, 2004)
[5] Cooper, B.: Turing centenary: The incomputable reality Nature 482,465, 23 February 2012
[6] D. Deutsch, *The Beginning of Infinity. Explanations That Transform the World.* Penguin, 2011, 496 pages
[7] D. Deutsch, "Creative Blocks", AEON Magazine, 02 October 2012
[8] G. Edelman, *Second Nature: Brain Science and Human Knowledge.* Yale University Press, 2006, 224 p.
[9] Koestler, A.: The Act of Creation, Penguin Books, New York, 1964.
[10] G. Lakoff, and M. Johnson, *Metaphors We Live By.* Chicago, IL: The University of Chicago Press, 1980
[11] Mokyr, J.: Mobility, Creativity, and Technological Development: David Hume, Immanuel Kant and the Economic Development of Europe. Session on Creativity and the Economy, German Association of Philosophy, Berlin, Sept. 18, 2005. In G. Abel, ed., Kolloquiumsband of the XX. Deutschen Kongresses fr Philosophie, Berlin 2006, pp. 1131-1161.
[12] Nanay, B.: An experiential account of creativity. In: Elliot Paul and Scott Barry Kaufman (eds.): The Philosophy of Creativity. Oxford: Oxford University Press, 2014, pp. 17-35.
[13] Valiant, L.: Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World, Basic Books, (2013)
[14] van Gelder, T.: What might cognition be, if not computation? The Journal of Philosophy, Vol. 92, No. 7. (Jul., 1995), 345-381.
[15] J. van Leeuwen, J. Wiedermann: Knowledge, representation and the dynamics of computation. To appear in: G. Dodig-Crnkovic, R. Giovagnoli (Eds): Representation and Reality: Humans, Animals and Machines, 2015, Berlin: Springer
[16] Wiedermann, J.: The creativity mechanisms in embodied agents: An explanatory model. In: 2013 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2013, pp. 41-45.
[17] Wiedermann, J. van Leeuwen, J.: Rethinking computation. In: *Proc. 6th AISB Symp. on Computing and Philosophy: The Scandal of Computation - What is Computation?*, AISB Convention 2013 (Exeter, UK), AISB, 2013, pp. 6-10
[18] Wiedermann, J., van Leeuwen, J.: Computation as knowledge generation, with application to the observer-relativity problem. In: *Proc. 7th AISB Symposium on Computing and Philosophy: Is Computation Observer-Relative?*, AISB Convention 2014 (Goldsmiths, University of London), AISB, 2014
[19] Wiedermann, J.,van Leeuwen, J.: What is Computation: An Epistemic Approach. In: G. Italiano *et al.* (Eds), *SOFSEM 2015: Theory and Practice of Computer Science*, Proceedings, Lecture Notes in Computer Science Vol 8939, Springer, 2015, pp. 1-13.



AISB Convention 2015

University of Kent, Canterbury

Proceedings of the AISB 2015 Symposium on Computing and Philosophy: The Significance of Metaphor and Other Figurative Modes of Expression and Thought

Edited by John Barnden and Andrew Gargett

ISBN: 978-1-908187-49-9

Introduction to the Convention

The AISB Convention 2015—the latest in a series of events that have been happening since 1964—was held at the University of Kent, Canterbury, UK in April 2015. Over 120 delegates attended and enjoyed three days of interesting talks and discussions covering a wide range of topics across artificial intelligence and the simulation of behaviour. This proceedings volume contains the papers from the *8th AISB Symposium on Computing and Philosophy: The Significance of Metaphor and Other Figurative Modes of Expression and Thought*, one of eight symposia held as part of the conference. Many thanks to the convention organisers, the AISB committee, convention delegates, and the many Kent staff and students whose hard work went into making this event a success.

—Colin Johnson, Convention Chair

Introduction to the Symposium

Communication and expression in language, pictures, diagrams, gesture, music etc. is rich with figurative aspects, such as metaphor, metonymy, hyperbole and irony. People engage in such communication and expression in a variety of contexts and with a range of effects. Modelling figurative patterns of communiction/expression is a key aim of academic disciplines such as linguistics, philosophy, discourse studies, and psycholinguistics, and automatically understanding such phenomena is a long-standing and now expanding endeavour within Artificial Intelligence, with metaphor generation also receiving more attention. In addition, some researchers have suggested that metaphor can be an intrinsic part of thought, not just an aspect of external communication/expression.

Specific topics of interest for the Symposium included the following:

- How philosophical thinking on figurative expression and thought can/should be exploited/heeded by relevant AI researchers
- How computational attempts to model figurative expression can aid philosophical thinking about it
- How the production of figurative expression reflects speakers' conceptualisations, goals and commitments
- How to model/analyse/understand the emotional and evaluative content of figurative expression
- The intersection of issues of figurative expression and issues of embodiment, enactivism, cognitive simulation, etc.
- Whether thought, as opposed to external expression, can be metaphorical, ironic, etc., and if so what this amounts to (philosophically, computationally, psychologically, ...)
- How figurative and especially metaphorical thinking might be involved in introspection, and therefore be bound up with the nature of consciousness
- Links between figurative thought/expression and the nature of creativity
- Figurative aspects of philosophical theorizing (about any topic), especially as uncovered by detailed technical analysis of figuration
- Figurative aspects of notions of computation…and even: could the notion of computation be irreducibly metaphorical?

It was a specific aim of the Symposium to encourage speculative thought, provisional proposals, and provocative question-raising based on careful analysis of issues. The papers in this volume serve that aim well. Given the broad scope of the Symposium we could not hope to cover all the topics listed above, but the papers range widely and bravely over the remit of the Symposium, which brings together disciplines in an unusual way.

We thank Colin Johnson as Convention chair, and the Committee of the AISB, for providing the opportunity to hold a symposium with the above remit. We thank our co-organizers Mark Bishop and Yasemin Erden for their help, and moreover for entrusting us with an event in a symposium series that they have played leading roles in over the years. We also thank our Programme Committee, which had wide international and disciplinary reach, for their hard work. Apart from the organizers, the committee contained Tony Beavers, Jerry Feldman, Eugen Fischer, Mark Phelan, Mihaela Popa, Mark Sprevak, Tony Veale and Yorick Wilks.

—John Barnden and Andrew Gargett, Symposium Organisers

Contents

Metaphor, Fiction and Thought

John Barnden [1]

Abstract. I will set out various un/underdeveloped opportunities for AI, philosophy and metaphor research to interact, with prospects for distinctly new lines of research and approaches to old problems. The opportunities I address in this paper are on the following topics: fiction-based accounts of metaphor, and a potentially resulting radical holism as regards the way metaphorical meaning arises from discourse; an anti-analogy-extension thesis, supporting unlimited non-parallelism between source and target in metaphor; the idea that thought can be metaphorical, and perhaps even more deeply than already mooted; deploying metaphor to solve a difficult problem in propositional attitude theory, which inludes the "meaning intention" problem as a special case; the "cognitive addition" of metaphor in language understanding, possibly leading to radical changes in how one thinks of the semantics even of non-metaphorical sentences.

1 INTRODUCTION

I will set out various un/underdeveloped opportunities for AI, philosophy and metaphor research to interact, with prospects for distinctly new lines of research and approaches to old problems. The opportunities I address in this paper are on the following topics, with the numbering corresponding to the sections of the paper.

2. Fiction-based accounts of metaphor, developed independently and under different names in various disciplines. One issue arising here is a possible radical holism as regards the way metaphorical meaning arises from discourse.
3. An anti-analogy-extension thesis, supporting unlimited *non*-parallelism between source and target in metaphor.
4. The idea that thought can be metaphorical, and perhaps even more deeply than already mooted.
5. Deploying metaphor to solve a difficult problem in propositional attitude theory (the problem being a generalization of the so-called "meaning intention" problem).
6. Something I call the cognitive addition of metaphor in language understanding, possibly leading to radical changes in how one thinks of the semantics even of non-metaphorical sentences.

There are threads strongly linking these topics. The dependencies will be summarized in the Conclusion section (section 7).

The paper draws heavily from already published papers and a journal paper under review (these will be cited below). In some places I incorporate partially-reworked extracts from those papers. However, the ideas have not all been drawn together before, or presented in a Computing and Philosophy venue, and some suggestions in sections 1 and 3 are new.

[1] School of Computer Science, University of Birmingham, UK, email: j.a.barnden@cs.bham.ac.uk

2 FICTION-BASED APPROACHES TO METAPHOR

I take a metaphorical expression such as "Ideas were whizzing around in his mind" to talk about a *target scenario* (here, a particular state of the mentioned person's mind and ideas) using the resources of a *source* subject matter (here physical objects and space).[2]

In various disciplines, researchers have suggested variants of an approach to metaphor that rests on what we can call *fictions*. Roughly and briefly, under such an approach the hearer of a metaphorical sentence uses the literal meaning of the sentence in context to (begin to) construct a fictional scenario expressed partly in source subject-matter terms. The fictional scenario is similar to a partial world as depicted by an ordinary fictional narrative such as a novel. The hearer may then elaborate (fill out) the fictional scenario by means of inference, using knowledge of the source subject matter. Metaphorical meaning arises when the hearer takes aspects of the fictional scenario and converts them into (alleged) aspects of the target scenario.

The fictional-scenario aspects that are so converted may either have been put there directly by the literal meaning of the metaphorical sentence, or may have arise through elaboration of the scenario. The created information about the target scenario forms part of the meaning of the sentence for the hearer. "Conversion" includes the case where an aspect is simply copied over to the target scenario without change, in the sense illustrated below.

This general characterization fits fiction-based approaches to metaphor in philosophy (see notably [42]), a recent enrichment of Relevance Theory accounts of metaphor developed in the field of linguistic pragmatics [18], and aspects of the "blending" or "conceptual integration" developed within cognitive science [21]. It is similar to the use of imaginary worlds for poetry understanding [31].

The characterization also fits the *ATT-Meta* approach to metaphor understanding that I have been developing and that is partially realized in a working computer program. I will describe this approach, as this will enable certain issues to arise in this section and other sections of this article.

The ATT-Meta approach makes an assumption that is contentious. Taking the above example of "Ideas were whizzing around in his mind," the approach does say that there can be a fiction in which an idea can do things like whizzing. Some may find this unintelligible. But perhaps this feeling can be allayed by the following. The approach in fact says that the stated whizzing implies that the ideas are indeed physical objects, in the fiction, as well as being ideas. In effect, the real-life fact that ideas are not physical objects is suppressed from becoming part of the fiction. (An important sector of the technicalities developed in the ATT-Meta computer program is for ensuring such suppression.) Another way of putting it is that it is presumably

[2] This statement is just a comment about metaphor, not a definition of it—and I am sceptical that it can be rigorously defined [7].

intelligible to state a counterfactual such as "If ideas were physical objects, then they could [do things like whizz around]." Fictions used in metaphor, at least according to the ATT-Meta approach, are akin to the bogus scenarios entertained in understanding counterfactuals. When compared to the real of ordinary fictions (novels, short stores, films, etc.) they are perhaps most akin to fanciful, fantasy tales.

However, a more moderate approach could have it that in the fiction there are physical objects that merely correspond to ideas outside the fictions rather than themselves also being ideas within the fiction. The rest of this paper would not be much disturbed by this alternative approach. (In fact, ATT-Meta probably needs to be extended to include the alternative anyway, in order to handle similes properly, such as in "The idea was like a balloon that was flying around the room." Currently, ATT-Meta would have to treat this in the same way as "The idea was a balloon that ...'.)

2.1 The ATT-Meta Approach

The ATT-Meta approach [2, 5, 6, 8, 12] is mainly geared towards cases involving familiar metaphorical views, such as the view of the mind as a physical region. However, the approach is not much concerned with conventional metaphorical phraseology based on such views, as in "The idea was at the back of his mind." Rather, it is geared towards open-ended forms of expression that transcend familiar metaphorical views. This is best brought out by examples such as the following:

1) "The managers were getting cricks in their necks from talking up [to some people in power over them] and down [to the managers' subordinates]."[3]

It is common for abstract control relationships, especially in organizational settings, to be metaphorically viewed in terms of relative vertical position of the people concerned. However, someone having a crick in their neck is not a matter addressed by this view. Thus the sentence transcends the metaphorical view in question.

For purposes of (1), the fictional scenario is seeded with the premise that the managers literally got cricks in their necks from continually looking in two different physical directions, upwards and downwards to the mentioned sets of people.[4] This scenario gets elaborated, for instance by addition of propositions that the cricks cause the managers to have pain, emotional stress, difficulty in continuing such head-turning, and dislike of continuing it. These propositions follow just by ordinary commonsense knowledge about neck-cricks, etc. Some of these elaborated aspects of the fictional scenario get converted to become target-scenario propositions such as that (a) the managers experience annoyance and other emotional stress, and (b) it is difficult for the managers to continue the conversations.

Note especially that (1) does not just convey (a). The sentence is richer than if it had merely said that the managers were "getting annoyed" at their conversations. Annoyance does not in general imply difficulty of continuing, though it may imply reluctance to continue. However, in the fictional scenario, having a crick in their necks not only causes pain but also *makes it difficult for the managers to continue turning their heads,* and therefore difficult to continue the conversations. This difficulty is simply copied over to the target-scenario (by a mechanism to be mentioned below).

Example (1) and others have been analysed under the ATT-Meta approach (see for instance [3], [4], [6] and [9]). In example (1) the only "conversion" of fictional-scenario aspects into target-scenario ones were actually change-free transfers: difficulty of continuing the conversation in the fiction is converted to provide the same difficulty in the target scenario. But in general, genuine conversions are needed. This is illustrated by the following example:

2) One part of Mary was insisting that Mick was adorable.

I take (2) to rest on two very general metaphorical views that are often used about the mind. First, there is the view of a person or a person's mind as having parts, where furthermore these parts are persons with their own mental states. I call these the "subpersons" of the person, and I call the view *Mind as Having Parts that are Persons*. (Note carefully that the parts are themselves a metaphorical fiction—the view *not* about objectively-existing parts of the person being metaphorically viewed as subpersons.) If a part (a subperson) of a person P believes (desires, intends, ...) X then, intuitively, the whole person P could be said to partly believe it. But what does it mean to partially believe something? The way I cast it is to say that the real person has a mere *tendency to* believe X. [5]

One main point of *Mind as Having Parts that are Persons* is that it allows different subpersons to have different beliefs or other types of mental state, and may even have beliefs that conflict with each other. This can rise explicitly in sentences that have a form such as "One part of P believes X, but another part believes Y" where X and Y conflict. In such a case the whole person P has tendencies to believe various conflicting things, without really *believing* any one of them. But I will also claim that the case of conflicting tendencies can arise implicitly, and in fact arises in (2).

The second metaphorical view comes into play when, as in in (2), the subpersons are portrayed as communicating in natural language. Since what is communicated is some idea that the whole person is entertaining, the additional metaphorical view here is that of *Ideas as Internal Utterances*. This is a very widely used metaphorical view that also often arises independently of *Mind as Having Parts that are Persons*. I will address the internal-utterances aspect of (2) shortly.

Now, there is a need to convert aspects of a fictional source scenario in which one or more "parts" of a person have particular mental states into aspects of the whole person's mental states in the target scenario. To handle fiction-to-target conversions, ATT-Meta borrows in part from conceptual metaphor theory (see [29], though more closely from [26]). A conceptual metaphor consists of a set of mappings—or as I will say, *correspondences*—between aspects of the source subject matter and aspects of the target subject-matter. These mappings constitute an analogy. The ATT-Meta approach broadly adopts this idea, though the correspondences are considerably different in form and function from those in conceptual metaphor theory and in analogy theory, as will be clarified below.

A metaphorical view in ATT-Meta involves a small number of very general, high-level, view-specific correspondences. In the case of *Mind as Having Parts that are Persons*, only two correspondences appear to be needed for a large array of examples. I just discuss one of them here. It can intuitively be expressed as follows.

(C) A person *having some tendency* to believe/desire/intend/fear/like/... something corresponds metaphor-

[3] Cited in [25, p.162]. The example is from the *Daily Telegraph* newspaper.

[4] In discussing ATT-Meta previously I have usually used a weak notion of pretence rather than fiction, and have called the fictional scenario the pretence scenario. For present purposes the notion of a fiction is more revealing.

[5] Elsewhere I have cast this as the person having a "motive" to believe X, in a very general sense of a reason or some other factor. This is on the assumption that a tendency to believe something is underlain by a motive to believe it. Here I revert to an earlier, more theoretically neutral formulation in terms of tendencies.

ically to *at least one subperson* of that person having a tendency to (respectively) believe/desire/intend/fear/like/... it.

C can be deployed by the hearer of (2) as follows. Taking sentence (2) literally, the hearer puts the premise that (literally) the mentioned part of Mary insists that Mick is adorable into the fictional scenario. This fictional claim is used to infer that (by default) the part is a subperson inside Mary. Given the general default that when people claim things they believe them, the hearer can then infer that, still in the fictional scenario, *that subperson believes that Mick is adorable*. It follows *a fortiori* that that subperson has a *tendency* to believe that Mick is adorable. Then hearer converts that fictional-scenario claim using (C), to become the target-scenario claim that *Mary* has some tendency to believe that Mick is adorable.

But also the insistence in (2) can be used to infer within the fiction that actually there is a subperson of Mary that believes that Mick is *not* adorable. This is because of the real-world nature of insistence. Typically, someone insists something when there is a conversation with a person who denies it. Thus, the presence of a subperson who claims that Mick is not adorable can be inferred by default. This new subperson presumably *believes* that Mick is not adorable. Hence, again using (C), we conclude that Mary has a tendency to believe that Mick is *not* adorable, as well a tendency to believe that he is.[6]

A final comment on (2) is that it crucially involves the notion of insistence by fictional subpersons, but this notion does not need to have its own correspondence to any non-metaphorical notion about the person's (Mary's) mental states. In short, insistence as such does not need to be handled by any correspondence associated with the two metaphorical views mentioned above. The insistence was used merely to generate, within the fictional scenario, certain conclusions that could be mapped by (C). If insistence does not have its own tailor-made correspondence associated with any metaphorical view the hearer knows, it is a view-transcending aspect of (2).

However, assuming that an utterance by a subperson is (metaphorically speaking) an utterance inside Mary, and assuming that *Ideas as Internal Utterances* involves a mapping of such utterances to thoughts of Mary's, then there is an additional line of processing leading to conclusions that Mary is entertaining certain thoughts.

One difference between ATT-Meta's approach and (other forms of) conceptual metaphor theory is that in ATT-Meta there are two broad sorts of correspondence: (i) *view-specific* correspondences such as (C), associated with particular metaphorical views, and (ii) *view-neutral mapping adjuncts* that apply by default in any case of metaphorical understanding, irrespective of what metaphorical views are in play, and that build upon the effects of, and indefinitely extend the reach of, the view-specific correspondences. Returning to the neck-crick example, (1), how can the hearer create target-scenario conclusions such as that the managers, in the target scenario, experience negative emotions, caused by the conversations, and find it difficult to continue their conversations? Such conclusions arise within the fiction, but they need to be transferred to the target scenario. The crucial observation here is that there are general qualities about metaphors' fictional scenarios that are very often copied in metaphor to the target scenarios no matter what the specific metaphorical view is. Amongst such qualities are the following:

- Emotional/attitudinal states, value-judgments, etc. (of typical observers such as the hearer to the target scenario, or of agents within the scenario itself).
- Mental states, such as believing, intending, wanting.
- Time-Course, incl. starting, continuing, ending, immediacy, smoothness/intermittency, rates at which episodes occur, temporal relationships between episodes, etc.
- Causation, prevention, enablement, ability, attempting and tendency relationships, and related qualities such as effectiveness.
- Ease/difficulty properties.

For each of these qualities there is a *View-Neutral Mapping Adjunct* (VNMA) that allows transference of aspects of a suitable fictional scenario to the target scenario. In our neck-crick example, one VNMA delivers a correspondence between emotional distress of the managers about the conversations, in the fiction, and emotional distress of the managers about the conversations, in the target scenario. The VNMA concerned with causation allows the inference that the fact that the conversations *cause* the emotional distress in the fiction is inferred to correspond to their also doing so in the target scenario. Equally, the within-fiction difficulty for the managers of continuing with the conversations transfers to the target scenario, because of VNMAs handling time-course (a case of which is the continuation of a situation) and difficulty. The continuation of a situation is one case of a qualitative temporal attribute.

While (1) only involves the use of VNMAs and (2) uses only view-specific correspondences, both types of conversion mechanism are needed in general. Both types are defeasible, so their results can be defeated in specific circumstances by other evidence.

One important facility currently missing from ATT-Meta is an ability to discover novel analogy between two scenarios. In a minority of cases of metaphor, and quite often with cases of so-called image metaphor (resting largely on physical appearance), there are no existing correspondences that will deliver useful results. However, a novel-facility could readily be added without disturbing the existing nature of the approach.

2.2 Issues for Fiction-Based Approaches

By virtue partly of having been realized in a working computer program, it is fair to say the inference and conversion mechanisms in ATT-Meta have been worked out much more specifically and completely than in fiction-based approaches developed in non-computational research endeavours, even though much more work needs to be done on ATT-Meta itself (both theory and program). The work of computationally operationalizing fiction-based theory has thrown some general issues into relief, all of which I believe need further research and, more particularly, could benefit from collaborative research between philosophy, metaphor theory and AI.

First, it is not rare for ordinary fictional narratives to meld several entities, such as people or places, in the real world into a composite entity in the fictional world. Ordinary fictional narrative can also do the reverse, i.e. have several different entities in the fictional world correspond to one entity in the real world. Such violations of one-to-one mapping between fiction and what lies outside the fiction raise philosophical issues—e.g., about the nature of fictional entities and about cross-world correspondences more generally—and detailed computational issues as regards representation and inference, while also possibly being important in metaphor. However, they have been little studied in the metaphor area. This may be partly because they are rare in metaphor—but the matter has not seen much explicit exploration. That it may not be rare is suggested by the *Mind as*

[6] As pointed out by a reviewer, (2) suggests that Mary is actually having conscious, occurrent thoughts about Mike. This addition to the interpretation of (2) can be handled by assuming that (C) covers such thoughts, and recognizing that when someone claims something X, insistently or otherwise, they have a conscious, occurrent thought that X.

Having Parts that are Persons view. Although ATT-Meta does not currently in fact postulate a mapping between the actual person and the fictional subpersons (as opposed to the above partial correspondence (C) between the mental states of the actual person and those of the subpersons), this might be a valid basis for analysis. Conversely, utterances such as "The country wants to abolish slavery," when analysed as metaphorical, could perhaps be cast as metaphor that puts one thinking agent in the fiction (that agent being the country) in correspondence with a large number of thinking agents in the country.[7]

Notice here in passing that, again, an element of the target scenario can also appear in the fictional source scenario, either with merely its properties from the target scenario or with a partially different set of properties. The country in the slavery example just mentioned is in both the target scenario and the source scenario, but in the latter it is a thinking agent as well as a country. We saw an analogous phenomenon when discussing ideas whizzing around in someone's mind: the ideas were in the source scenario as well as the target scenario, but in the source scenario they were physical objects as well as ideas. This use by a fiction of elements from outside it, with possibly a warping of the nature of those elements, is familiar from ordinary fictional stories.

Secondly, I have argued elsewhere [11] that metaphor understanding can be facilitated by "reverse" conversion steps, i.e. ones in the target-to-fictional-scenario direction, as well as ones in the normal, forwards direction. Such reverse conversion is in fact implemented as standard in the ATT-Meta system. The most interesting basis for wanting reverse conversion is a claim that it is sometimes easier to find coherence between related metaphorical utterances in a discourse and surrounding or interspersed utterances by looking to the fictional scenario rather than to what the fictional scenario says about the target scenario. Reverse conversion brings fiction-based theory of metaphor closer to the theory of fiction in general, given that it is standard for ordinary stories to bring in information about the real world. For instance, if we know that a certain fictional character is intended to correspond to a real person, we would tend to import our knowledge of that person into the fiction (if not contradicted there) suitably amending it to fit the circumstances of the fiction. Yet reverse conversion is not extensively considered in metaphor research. (It has been mooted without extensive detail in the context of Interaction theories of metaphor [41], and has been discussed in some applications of the blending approach)

Thirdly, I have also argued elsewhere (e.g., in [13]) that a metaphorical sentence sometimes cannot readily be given its own meaning in terms of the target scenario. Rather, it may conspire with surrounding literal or metaphorical sentences to convey something about the target. This is a form of holism about discourse meaning. The general point is that several sentences in a discourse might need to contribute to building up a fictional scenario (perhaps with the help of reverse conversion, if literal sentences are involved) and to allow appropriate elaborations that lead to fruitful opportunities for fiction-to-target conversion. However, following traditional assumptions about literal sentences , language researchers in many disciplines appear to assume virtually without argument that every sentence, including metaphorical ones, must be assigned its own meaning in terms of the situation actually being talked about. However, I conjecture that it is merely a *typical* case that a sentence taken alone can be assigned such a meaning. Rather, meaning can act much more holistically across sentence (or clause) boundaries, and there is no hard syntactic limit as to what sort of segment of discourse might in a particular case be treated most naturally as a unit bearing specific meaning.

An example I use in [13] is

3) "Everyone is a moon, and has a dark side which he never shows to anybody." [attributed to Mark Twain by [17, p.74]]

Note that the example could just as well have been in the following multi-sentence form, which is just as comprehensible:

3a) "Everyone is a moon. Everyone has a dark side which he never shows to anybody."

I suggest that it is misguided to suppose we must first derive a metaphorical meaning for the clause/sentence "Everyone is a moon" and a metaphorical meaning for the clause/sentence "[Everyone] has a dark side which he never shows to anybody" and then combine these meanings. Rather, the second clause indicates what it is about being a "moon" that we should attend to (this isn't provided by the first clause), while it is the first clause that brings moons into the picture (the second clause doesn't do this). I claim the best approach is to form a fictional scenario on the basis of both clauses, and only then extract implications for the target scenario. In the fiction, the moon aspect reinforces the never-showing aspect of the second clause.[8]

Now, the second clause in (3) or second sentence in (3a) could plausibly have been given a metaphorical meaning even if the first clause/sentence hadn't been uttered. The fiction would have just cast the person as *some* physical object that has a dark side not shown to anyone else. So, for (3/3a) itself, one can imagine a process whereby the hearer works out that metaphorical meaning for the second clause/sentence and only later refines or strengthens it in some way by means of the first clause/sentence.

But the main point I wish to make is that it would be quite hard to give the first clause its own metaphorical meaning, and therefore quite hard to form an integrated understanding by taking a metaphorical meaning for the sentence and a metaphorical meaning for the second and combining them. . Either it would involve using the second clause for guidance as to what the first one means, in which case there hardly seems any point considering the first clause at all by itself, or the operation would involve taking the clause in isolation of the second, in which case (unless surrounding discourse context could help) we have the usual problem of the indeterminacy of metaphor (see, e.g.,[39]). Without the second clause it is wide open what the first clause is getting at. For example, it could be construed as saying that everyone is somehow subservient to something that is being metaphorically portrayed as the Earth, or as saying that everyone serves as a source of illumination for the world in times of darkness, or ...

Actually, the first clause has a deeper effect than just reinforcing the never-showing in the second clause. The moon also has a bright side, at least some of which we can normally see, and which is extremely salient in a clear night sky. Thus, a more elaborated interpretation of (3) or (3a) could include the notion that everyone also has a side that is (in part) usually very much apparent. This new message cannot come from just the second clause, because although the mention of a dark side weakly suggests a non-dark side, there

[7] Sentences such as "The country wants to abolish slavery" would typically be analysed as involving a *metonymic* step from country to (some/many) people in the country. But the metaphorical analysis route has also been mooted (see, e.g., [32]), and would gain weight in a richer case such as "The country is sweating with the effort of getting rid of slavery."

[8] (3) apperas to assume that Earth's physical moon has a dark half that cannot be seen. here there seems to be a mistaken supposition that the dark side is a fixed part of the moon, rather than changing as the moon orbits the Earth. Also, the passage may be mistakenly equating the dark side with the side facing away from the earth.

is no warrant for taking that side to be bright and salient. But, the fact that the message cannot come just from the second clause alone is a not a reason for saying that the first clause should be given its own metaphorical meaning, but is rather a reason to say that a unified fictional scenario should be constructed from both clauses, and then target-scenario meaning should be extracted from that scenario as appropriate. However, I do not have a specific theory about how hearers are pressured to adopt this more holistic approach across clauses/sentences and when they give them separate metaphorical meanings.

Thirdly, I have sought to explain chained metaphor (where something A is viewed as B and something about B is viewed as C) in terms of nesting of fictions within each other. I have treated some real examples elsewhere, but a simple, chained variant of (1) would be "The managers had cricks chewing into their necks ...." where the managers' state is metaphorically cast as having a crick in their necks but the cricks are in turn cast as being animals. This would be handled by having the fictional scenario discussed above, but now there would be, nested within it, a fiction in which the cricks are animals. This nesting is of course similar to the common phenomenon of stories-within-stories. It would appear that this matter needs further attention in the philosophy of fiction (not least because of the question of whether or not it is merely fictional that the inner fiction exists, and how one formally cashes out that potential meta-fictionality), while on the other hand metaphor research has been slow to come up with detailed theories of chained metaphor.

3 AN ANTI-ANALOGY-EXTENSION THESIS

In the ATT-Meta approach, as in conceptual metaphor theory, metaphor is based on familiar analogies. An ATT-Meta metaphorical view involves a set of entrenched analogical ncorrespondenceXX rules, and VNMAs are additional analogical correspondence rules. Nevertheless, a key point about the ATT-Meta approach can be called the *Anti-Analogy-Extension Thesis*.[9] This says that open-ended view-transcending elements of the source subject matter (e.g., the crick in (1), the insisting in (2)) should *not*, normally, be given target-scenario parallels, and in particular that existing analogies should not be extended to encompass those elements—they should be left unparallelled. ATT-Meta seeks to get away with the least amount of analogy possible, *contra* other theories such as Structure-Mapping Theory [22, 15], which assume that the task is to maximize the extent of analogy.

In contrast to such theories, the ATT-Meta approach claims that the hearer tries to connect view-transcending to within-fiction content that *can* be converted via already-known correspondences (view specific or view-neutral). This is on the theoretical principle that, typically, the unparalleled items are proposed by a speaker not as individually standing for aspects of the target scenario being addressed, but rather to build a fictional scenario that holistically illuminates the target side using correspondences that the hearer is expected already to know.

In particular, in the neck-crick example (1), the cricks and resultant physical pain have no parallel in the target scenario. The cricks are only there to convey emotional distress, difficulty in continuing the conversations, etc. Similarly, there is no need at all to propose that for (2) the mentioned part corresponds to an identifiable aspect of the real person, or to propose that there is some internal, real mental action that can be clearly held to correspond to the action of insisting in the sentence. Rather, the mentions of a part and of insisting are *merely* tools towards constructing a rich fictional scenario, which in turn conveys in an economical, accessible and vivid manner the possession of a particular sort of mental state by Mary.

The Anti-Analogy-Extension Thesis goes hand in hand with a form of holism about the fictional scenarios and the metaphorical sentences leading to them, related to the holism of the previous subsection. The fictional scenario is to be regarded not as having a detailed analogy to a target scenario but rather something that *holistically* conveys information about the target scenario. This conveying is, to be sure, done by the action of correspondences that pick on specific aspects of the fictional scenario. But the ultimate intent here is to transfer information, not specify an analogy. And any specific aspect of the fictional scenario that is grabbed by a correspondence may be the result of inference over large amounts of information within the scenario. In particular what this means is that there may be no specific part of the metaphorical sentences that can be said to correspond to a given aspect of the reality scenario (although this can happen in simple cases of metaphor). For example, going back to (2), an aspect of its meaning not detailed above (but explained in [9]) is that Mary lacks the belief that Mike is adorable (she merely has a tendency to believe it, and indeed also has a tendency to disbelieve it). This lack does not correspond to any one aspect of (2) but rather to the whole of (2).

Another work that emphasizes both frequent holism of metaphor (in this subsection's sense) and the lack of need for, or indeed the frequent undesirability of, analogy-extension is Langlotz's treatment of idioms [30], including metaphor-based ones.

4 METAPHORICITY OF SOME THOUGHT

The anti-analogy extension thesis has interesting consequences for the nature of thought, consequences that have barely been addressed in AI or philosophy and need more work in metaphor theory itself. Within the cognitive linguistics field, it is typical to think of metaphor as something that is somehow fundamental in the mind, not just in communication and external expression, and in particular to think of many concepts, particularly abstract ones, as in some way structured by metaphor (i.e., by being linked by metaphorical mappings to source concepts). See [40] and [33] for critical discussion of some of the main points here. One reason for the hypothesis is that metaphor occurs in media other than language, such as in graphical media. One might try to account for this in a number of ways, but an one parsimonious option is that metaphor is inherently a mental as opposed to purely communicative or externally-expressive phenomenon. I will take the point to basically be that, when thinking but not externally communicating about some subject matters, we are at least sometimes mentally using metaphorical mappings between those subject matters and suitably-related source subject matters. There is no implication here that this mental activity is conscious. I assume here that it may well be unconscious.

The Anti-Analogy-Extension Thesis leads to an especially strong claim: namely, that major portions of a metaphorical thinking episode may not individually have *any* translation into non-metaphorical thoughts within the person's mind. This is because extensive areas within a metaphorical fiction may not have any analogical correspondence to the target scenario, but rather just serve indirectly to support those limited aspects of the fiction that are in analogical correspondence to the target. Open-ended elaboration of fictional scenarios could exist in mind just as much (or more) than in language and other external expression. For example, someone thinking (but not communicating) about the managers in (1) may

[9] The account in this section is based on [8].

mentally develop the fictional scenario in creative ways as above, such as imagining pains in many parts of the managers' bodies, not just their necks, imagining the managers massaging those parts, contorting themselves, etc. These could have consequences about the intensity of the emotional states, their longevity and difficulty of eradication, and the desires of the managers. These conclusions can be mapped to reality. But most of the fictional scenario is *not* mapped.

I also wish to make a more radical conjecture. In the discussion so far, even if some thoughts are in an unparalleled region of a fictional scenario, their function in the mind is nevertheless to support fiction-to-target conversions that produce mental representations directly in terms of the target subject matter. One might say that the latter representations are literally about the target scenario—so the unparalleled parts of the fiction are indirectly connected to those literal representations. But it is possible that there are metaphorical representations in the mind that have *no* connection to a literal description of the target scenario, even indirectly. For instance, one can conceive of a person whose only resource for thinking about electricity is that it is a liquid flowing within wires, etc. She knows nothing about electricity other than what can be approximately captured by these resources, and she has no translation of the liquid-based thoughts about electricity into any other terms. Many of our concepts about relatively abstract matters, such as time, electricity, money, love, mental states, ... at least *include* metaphorical views, and I am now supposing that a concept could consist *only* of such a view. So, the person's concept is *irreducibly metaphorical*. (This does not mean either that it is irreducible in principle or that for some other person it is not irreducible.)

Yet the person might agree, if asked, that electricity isn't *really* a liquid. If she knows about metaphor, she might more specifically agree that electricity is only metaphorically a liquid. So, we as observers, and even the person herself, should not take her to think that electricity really is a liquid, but rather as metaphorically thinking about electricity as a liquid, perhaps unconsciously. As long as her liquid thoughts are adequately linked to relevant actions she needs to take in the world (e.g. actions on switches, carefulness about cutting wires, etc.) she can operate in the world perfectly well for everyday purposes. While this sort of possibility falls naturally out of standard cognitive linguistic considerations (even if not yet fully developed in that field), it appears not to be catered for in detailed theories of representation and mind in AI and philosophy.

5 ATTACKING AN ESOTERIC NETTLE WITH THE SCYTHE OF METAPHOR

I believe considerations of metaphor can help with a long-standing philosophical problem about the nature of propositional attitudes (broadly, contentful thoughts) and the meaning of propositional attitude reports—reports of mental states, with sentences of the form "John believes ..." as the simplest sort of example. Metaphor could provide a radically new, and subversive, solution. I call the problem one of *esoteric imputation*. It has been noted in different forms by various philosophers, such as Clapp, Richard, Schiffer and Soames (see citations below), and often arises with attempts to provide theories of propositional attitudes (PAs) and the meaning of PA reports. The problem is that theories are in danger of imputing, to ordinary people, thoughts that implausibly involve esoteric aspects of non-commonsensical explications of thought that are postulated by the theories.[10]

For example, one common type of theory is roughly that the meaning of "John believes that spies are evil" is that John is in a certain relation BEL to the proposition that spies are evil, via some "mode of presentation," "way of thinking" or "guise" for that proposition. Such a theory involves some specific, technical notions of matters such as what a proposition is, what a mode of presentation (etc.) is, what it is for a mode of presentation to present something, what BEL is, and what it is for a proposition to refer to the world. Typically, while some aspects of these technical notions might be reasonably intuitive, the whole package is so esoteric that it is unimaginable that anyone other than philosopher could entertain them in their thoughts.[11] (See [36, 37] for complaints along these lines, in discussion of the "meaning intention problem." See also [1].) Lest someone think that what one calls the meaning of a PA report or any other sentence needn't be the same thing as the content that a hearer grasps when encountering it, I should point out that the problem arises also in iterated attitude reports such as "Mary believes that John believes that spies are evil." Here, one's theory of PAs and PA reports should not have as a consequence that Mary has a belief that is couched in terms of of the esoteric explication of John's belief that the theory would assign as the meaning of "John believes that spies are evil." or more broadly as the scientifically accurate nature of what it is for John to believe that spies are evil.

Some specific further instances of the problem arising in the philosophical literature are as follows, interlaced with some observations of my own. Schiffer [37, pp.35–37] highlights an esoteric imputation problem with Fregean accounts of PA reports, in that belief reporters are unaware of the detailed nature of concepts, and notably of Fregean ones. Hornstein [27] characterizes many PA theories as requiring the belief reporter to have some grasp of theories of sense and reference, and he implies that this is mysterious. Edelberg [20] says that an approach by Kaplan to PA reports seems implausibly to require ordinary people to know and understand Kaplan's theory. Braun [16] suggests that the hypothesized speaker thoughts about modes of presentation in the above approach cannot be made explicit by speakers, casting doubt on the existence of those thoughts. Berg [14, pp.26–27] worries that an explanation of what it is to believe a proposition *under* a given mode of presentation is (what I would call) esoteric. Clapp [19] makes claims about major PA report accounts requiring speakers to know esoteric things about ordinary believers' thoughts, and he claims that attempts to mitigate this problem don't fully work and/or make the accounts fall into other problems. Clapp implies that even the authors who are aware of such [esoteric imputation] problems have failed to solve the problem.

To get some of the flavour of current discussion about the topic, we can consider Richard's [35, Ch.13] response to a complaint by Soames [38, p.170] against his account. Soames questions whether speakers really intend to commit themselves to complex claims (that he takes Richard's theory to involve) about the languages or internal mental representations used by believers to which they typically ascribe beliefs. Richard counters that the thoughts he is imputing to speakers are in fact not implausibly complex; and I also take him to argue that the thoughts are not esoteric. He says "it is uncontroversial that conversants routinely make presuppositions about how others represent the world[.]" This may be true but the question really is whether conversants have the particular sorts of thoughts about the particular sorts of representations that Richard proposes. I am made

[10] This section draws from [10].

[11] At least, it's unimaginable that they can consciously do so, and only with a theory that radically dislocates unconscious from conscious thought would allow them to unconsciously think in terms of such esoteric notions even though they cannot do so consciously. (My impression is that the tension here between unconscious and conscious thought is not commonly enough considered in the philosophical area in question.)

nervous by the following statement by Richard [35, Introduction, p.22], concerning a report of form "Boswell thinks that S." According to Richard's theory, this has a logical form that can be glossed in English along the following lines, where "annotated proposition" is a technical, rather esoteric notion that Richard has defined:

> There's an acceptable translation manual ... such that one of Boswell's beliefs (i.e. an annotated proposition determined by one of his belief states) is translated, under that translation manual, by the annotated proposition that S.

So, suppose we consider Yolanda believing that Boswell thinks that S. Does she then have something like the concept of a mental translation manual or of an annotated proposition? Perhaps it is plausible that she has such thoughts, via suitable modes of presentation perhaps, but it is up to Richard to convince us of it.

Also, the book by King, Soames and Speaks [28] contains several comments relevant to esoteric imputation. For instance, Soames's and Speaks's articles in the book complain that King's account there requires ordinary language users to have esoteric thoughts. But Soames's account in the book has, itself, an esoteric imputation problem. It is central to his proposal that people become familiar with their own cognitive acts and then abstract from these to become familiar with more general, agent-independent cognitive-act types (constituting propositions etc.). But I suspect that individual act types as portrayed by Soames are esoteric: certainly, discussions in the literature about them are highly esoteric. Also, if people's categories are generally based on prototypes and/or exemplars, then this may apply just as much to cognitive-act types as to other types of things; but then it becomes difficult to isolate objectively existing act types of Soames's sort.

Thus, we have evidence that it is extremely difficult to come up with theories that avoid esoteric imputation problems using current philosophical resources. While it may yet be possible to do so, it would appear to involve theoretical contortions of great agility and knottedness. In response, I suggest a different strategy, inspired by the claim in cognitive linguistics and elsewhere that people often conceive of mental states, along with many other abstract matters, with the help of metaphor. I suggest that PA theory should positively impute to ordinary agents thoughts about each other's mental states and processes that are *framed in terms of commonsensical metaphor.* The basic idea is that a hearer of, say, "John believes that spies are evil" will (typically unconsciously) think of John's mental state in a metaphorical way, e.g. by thinking of John saying something to himself (silently) in English, or as John having having a mental image of spies being evil, or some combination of these. Equally, in an iterated case such as "Mary thinks that John believes that spies are evil," the hearer imputes to Mary a metaphorical view of John's mental state. Of course, there is an important question here about what particular view or views Mary might impute to John. I discuss this in [10].

In short, the advocated approach *deliberately* imputes to ordinary people *commonsensical, metaphorical* thoughts about mental states, rather than *non-deliberately* imputing to them *non-commonsensical, esoteric* thoughts about mental states. Particular effects of this approach. apart from avoidance of esoteric imputation, include (a) a new range of ways in which believing (or hoping, wanting, ...) in general may be viewed in acts of attitude report understanding, and (b) metaphor-relativity in the distinctions between different styles of interpretation such as transparent and opaque, which have been much discussed in the philosophical and AI literatures as if they were objectively characterizable.

Naturally also, insofar as the metaphorical framing of a situation affects one's behaviour in/towards it, the approach has practical consequences for AI systems that are meant to be interacting with human beings who are having thoughts about other people's thoughts.

6 COGNITIVE ADDITION OF METAPHOR IN LANGUAGE UNDERSTANDING

The approach to propositional attitude reports advocated in the previous section rests on an assumption that metaphor can be *cognitively added* during understanding. The hearer's understanding of the sentence "John believes ..." is *metaphorically* couched in the hearer's mind, even though the sentence itself contains nothing that would typically be called metaphorical by metaphor researchers. Thus, metaphoricity has been added by the hearer. But this isn't a special assumption just to make that approach work. It arises very naturally out of much more general considerations.

Recall the view in cognitive linguistics that metaphor is a conceptual matter, not primarily a matter of language or other modes of external expression. For instance, it is supposed that people think about time using any of a variety of metaphorical views (see, e.g., [34]). Under one, the person is moving along a spatial axis towards events, and in a dual of this, events are moving toward the person. There has been much discussion of the use of such views in interpreting metaphorical sentences such as "The meeting was moved forward/back." However, my claim is that the interpretation even of a *literal* sentence such as "The meeting time was changed to noon on the next day" can be accompanied by metaphorical couching of what the sentence says. If the hearer's concept of and general private thoughts about time include metaphorical aspects (even if not irreducibly so) it is only natural to suppose that those aspects are activated even by literal utterances about time. Thus, for the sentence "The meeting time was changed to noon on the next day" the hearer may mentally construct a metaphorically couched thought that paints the meeting as having been moved along a spatial axis.

Recent work in empirical psycholinguistics such as in [23, 24] suggests that people do often activate concepts in the source domain of a metaphorical view when understanding a *metaphorical* utterance based on it. This can even happen when the metaphorical language is highly conventional or even supposedly "dead." It is not a big step from here to the idea that people also do cognitive addition of metaphor when understanding some literal language (which is often "dead" metaphor anyway).

But it appears that all work on metaphor within language in philosophy and AI is confined to the question of how to account for the meaning of sentences that are, so to speak, already metaphorical. There appears to be an uncritically adopted, tacit assumption that the understanding of an ostensibly literal sentence only ever involves semantic representations that are themselves directly about the subject matter at hand, rather than bearing a metaphorical or other indirect relationship to that subject matter. But in reality we must countenance the possibility that the figurativeness or otherwise of utterances is only weakly related to the figurativeness or otherwise of the mental representations arising from or giving rise to the utterances.

7 CONCLUSION

I commend the issues covered in this paper as possible discussion points for Computing & Philosophy researchers who are interested in metaphor or foundational issues concerning the meaning of language.

The different sections above depend on each other to a considerable extent, although there are islands of independence. The anti-analogy-extension thesis is facilitated by a fiction-based account, and perhaps requires such an account. Thus the particular points made about metaphor within thought, which exploit that thesis, also depend on a fiction-based approach (but other approaches could also embrace metaphor in thought in other ways). However, the general notion of cognitive addition of metaphor does not presuppose a fiction-based approach. The use of metaphor to address the esoteric imputation problem for propositional attitude theory assumes that thought can be metaphorical and that cognitive addition happens. In fact it assumes, though this was not explicitly stated above, that a person's X's thoughts about other people's thoughts are often irreducibly metaphorical, and this does amount to viewing X's thoughts as defining fictions that are not cashed out in non-fictional target scenarios in X's mind.

ACKNOWLEDGEMENTS

The research in this article was supported in part by a Research Project Grant F/00 094/BE from the Leverhulme Trust in the UK. I am grateful to an anonymous reviewer for several thought-provoking suggestions that led to significant improvements to the paper.

REFERENCES

[1] Bach, K. (2000). Do belief reports report beliefs? In K.M. Jaszczolt (Ed.), *The Pragmatics of Propositional Attitude Reports,* pp.111–136. Oxford: Elsevier.

[2] Barnden, J.A. (2001a). Uncertainty and conflict handling in the ATT-Meta context-based system for metaphorical reasoning. In V. Akman, P. Bouquet, R. Thomason & R.A. Young (Eds), *Modeling and Using Context: Third International and Interdisciplinary Conference* (CONTEXT 2001). Lecture Notes in Artificial Intelligence, Vol. 2116, pp.15–29. Berlin: Springer.

[3] Barnden, J.A. (2001b). Application of the ATT-Meta metaphor-understanding approach to various examples in the ATT-Meta project databank. Technical Report CSRP–01–02, School of Computer Science, The University of Birmingham, U.K.

[4] Barnden, J.A. (2006). Consequences for language learning of an AI approach to metaphor. In J. Salazar, M. Amengual & M. Juan (Eds), *Usos Sociales del Lenguaje y Aspectos Psicolingüísticos: Perspectivas Aplicadas,* pp.15–57. Palma, Mallorca: Universitat de les Illes Baleares.

[5] Barnden, J.A. (2008). Metaphor and artificial intelligence: Why they matter to each other. In R.W. Gibbs, Jr. (Ed.), *The Cambridge Handbook of Metaphor and Thought*, 311–338. Cambridge, U.K.: Cambridge University Press.

[6] Barnden, J.A. (2009). Metaphor and context: A perspective from artificial intelligence. In A. Musolff & J. Zinken (Eds), *Metaphor and Discourse*, pp.79–94. Basingstoke, UK: Palgrave Macmillan.

[7] Barnden, J.A. (2010). Metaphor and metonymy: Making their connections more slippery. *Cognitive Linguistics, 21*(1), pp.1–34.

[8] Barnden, J.A. (2015). Open-ended elaborations in creative metaphor. Invited chapter for Besold, T.R., Schorlemmer, M. & Smaill, A. (Eds,) *Computational Creativity Research: Towards Creative Machines,* pp.217–242. Atlantis Press (Springer). (Series: Atlantis Thinking Machines, Vol. 7.)

[9] Barnden, J.A. (in press). Mixed metaphor: Its depth, its breadth, and a pretence-based approach. In R.W. Gibbs, Jr. (Ed.), *Mixed Metaphor*. Amsterdam: John Benjamins.

[10] Barnden, J.A. (under review). Propositional attitudes and commonsense psychology: Using metaphor to slay an esoteric nettle. Submitted to a philosophy journal.

[11] Barnden, J.A., Glasbey, S.R., Lee, M.G. & Wallington, A.M. (2004). Varieties and directions of inter-domain influence in metaphor. *Metaphor and Symbol, 19*(1), pp.1–30.

[12] Barnden, J.A. & Lee, M.G. (2002). An artificial intelligence approach to metaphor understanding. In Tomasz Komendzinski (Ed.), "Metaphor: A Multidisciplinary Approach," a special issue, *Theoria et Historia Scientiarum, 6* (1), pp.399–412.

[13] Barnden, J.A. & Wallington, A.M. (2010). Metaphor and its unparalleled meaning and truth. In A. Burkhardt & B. Nerlich (Eds), *Tropical Truth(s): The Epistemology of Metaphor and Other Tropes,* pp.85–121. Berlin / New York: De Gruyter.

[14] Berg, J. (2012). *Direct belief: An essay on the semantics, pragmatics and metaphysics of belief*. Berlin/Boston: De Gruyter Mouton.

[15] Bowdle, B.F. & Gentner, D. (2005). The career of metaphor. *Psychological Review, 112*(1), pp.193–216.

[16] Braun, D. (1998). Understanding belief reports. *The Philosophical Review, 107* (4), pp. 555–595.

[17] Brians, P. (2003). *Common errors in English usage*. Franklin, Beedle & Associates, Inc.

[18] Carston, R. & Wearing, C. (2011). Metaphor, hyperbole and simile: A pragmatic approach. *Language and Cognition, 3*(2): pp.283–312.

[19] Clapp, L. (2000). Beyond sense and reference: An alternative response to the problem of opacity. In K.M. Jaszczolt (Ed.), *The Pragmatics of Propositional Attitude Reports,* pp.43–75. Oxford: Elsevier.

[20] Edelberg, W. (1992). Intentional identity and the attitudes. *Linguistics and Philosophy, 15*, pp.561–596.

[21] Fauconnier, G. & Turner, M. (2008). Rethinking metaphor. In R.W. Gibbs, Jr. (Ed.), *The Cambridge Handbook of Metaphor and Thought*, pp.53–66. Cambridge, U.K.: Cambridge University Press.

[22] Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science, 7* (2), 95–119.

[23] Gibbs, R.W., Jr & Matlock, T. (2008). Metaphor, imagination, and simulation: Psycholinguistic evidence. In R.W. Gibbs, Jr. (Ed.), *The Cambridge Handbook of Metaphor and Thought*, pp.161–176. Cambridge, U.K.: Cambridge University Press.

[24] Gibbs, R.W., Jr & Santa Cruz, M.J. (2012). Temporal unfolding of conceptual metaphoric experience. *Metaphor and Symbol, 27*(4), pp.299–311.

[25] Goatly, A. (1997). *The language of metaphors*. London and New York: Routledge.

[26] Grady, J.E. (1997). THEORIES ARE BUILDINGS revisited. *Cognitive Linguistics, 8*(4), pp.267–290.

[27] Hornstein, N. (1984). *Logic as grammar.* Cambridge, MA: MIT Press.

[28] King, J.C., Soames, S. & Speaks, J. (2014). *New thinking about propositions*. Oxford: Oxofrd University Press.

[29] Lakoff, G. & Johnson, M. (2003). *Metaphors we live by.* 2nd Ed. Chicago: University of Chicago Press.

[30] Langlotz, A. (2006). *Idiom creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English*. Amsterdam/Philadelphia: John Benjamins.

[31] Levin, S.R. (1988). *Metaphoric worlds*. New Haven, CT and London, U.K.: Yale University Press.

[32] Low, G. (1999)."This paper thinks...": Investigating the acceptability of the metaphor AN ESSAY IS A PERSON. In L. Cameron & G. Low (Eds), *Researching and Applying Metaphor,* pp.221–248. Cambridge, U.K.: Cambridge University Press.

[33] Murphy, G.L. (1996). On metaphoric representation. *Cognition, 60*, pp.173–204.

[34] Moore, K.E. (2006). Space-to-time mappings and temporal concepts. *Cognitive Linguistics, 17*(2), pp.199–244.

[35] Richard, M. (2013). *Context and the attitudes: Meaning in Context, Vol. 1*. Oxford: Oxford University Press.

[36] Schiffer, S. (1992). Belief ascription. *J. Philosophy, 89* (10), pp.499–521.

[37] Schiffer, S. (2003). *The things we mean*. Oxford: Oxford University Press.

[38] Soames, S. (2002). *Beyond rigidity*. Oxford: Oxford University Press.

[39] Stern, J. (2000). *Metaphor in context*. Cambridge, MA and London, UK: Bradford Books, MIT Press.

[40] Vervaeke, J. & Kennedy, J.M. (2004). Conceptual metaphor and abstract thought. *Metaphor and Symbol, 19*(3), pp.213–231.

[41] Waggoner, J.E. (1990). Interaction theories of metaphor: psychological perspectives. *Metaphor and Symbolic Activity, 5* (2), pp.91–108.

[42] Walton K. (2004/1990). Fiction and non-fiction. In E. John & D.M. Lopes (Eds), *Philosophy of Literature—Contemporary and Classic Readings: An Anthology*, pp.136–143. Oxford: Blackwell. Reprinted from *Mimesis and Make-Believe: On the Foundations of the Representational Arts,* pp.36–41, 70–73, 77–78, 81–82, 84–89, Cam., Mass: Harvard UP, 1990.

Metaphor and understanding *me*

Yasemin J. Erden [1]

Abstract. This paper explores the role of the metaphor-maker in the construction of meaningful metaphor construction. More specifically, the paper defends the claim that the semantic-language-user is key for the possibility of both meaning and the understanding of metaphor. This takes into account the seemingly contradictory status of two claims: (1) that words can be meaningful without context, intentionality or the presence of, or origin in a language-user, while (2) the expectation of a context, intention or speaker is central to finding meaning in words and particularly metaphors. The apparent contradiction can be resolved if we see that the possibility of meaningful metaphor says as much about our expectation and need for meaning as it does about the language itself. Understanding words is thus as much about understanding the utterer of the words, as about the words themselves. Through exploring Wittgenstein's ideas about metaphor, this idea should become clearer. The paper will then explore what the limitations of computational metaphor might be as a result.

1 INTRODUCTION

What does it mean to understand a person through their words? And what do words mean separate from a speaker? These are questions that this paper explores in order to understand the central question: how are metaphors *meaningful?* In this, the aim is not to discuss the meanings of individual words, but rather to explore the very possibility of meaning and to point to the central roles played by context, expectation, experience and embodiment. To do this we begin by looking at a short quotation from Ludwig Wittgenstein, which has puzzled commentators because of its self-referential turn of phrase. The claim is made (or rather, defended, since the claim is not new even if it remains controversial) that to understand the phrase requires that we understand the person, Wittgenstein, as well as the words he uttered in that sentence [2].

Building on this, I argue that the possibility of a meaningful metaphor relies on context within which language is embedded, such as described by Wittgenstein [3] in terms of *language-games*. This does not lead to a strong claim that computational-metaphor is impossible though it does suggest a weaker claim that to be *successful* (which includes indicators such as 'appropriateness' or even 'acceptable') in this area may be tricky. This is partly because what is considered either appropriate or acceptable in ordinary language is already tricky (including where highly creative language-use can muddy the waters of ordinary language substantially). It is also partly because of the role that *expectation of meaning* creates. As I discuss elsewhere [4] [5], meaningful language-games require not only a successful meeting of *rules*, but also a willing on the part of participants to *recognise* other speakers as meaningful language-users. In the case of the words uttered by Wittgenstein, it is precisely because scholars expect meaning to be found, that the search for a meaning is considered worthwhile.

To explore this further, we will also discuss the possibility of non-human (or computational) metaphor construction, interpretation and use, and discuss the likely limitations that may occur where such construction is disembodied and decontextualized. The concept of the language-game will be employed in this discussion, since Wittgenstein offers this as a metaphor for meaningful language use. The metaphor of a *game* is particularly helpful for exploring ideas about participation and mimicry, and thereby how we view the relationship between computational and non-computational approaches to both metaphor understanding and production. Will we accept a metaphor as creative or even useful if we do not believe the person (or program) has any idea (understanding or experience) of the individual components, let alone the comparison being drawn?

Finally, discussion will explore the way that, on the one hand we might measure the *success* of a program (in constructing or interpreting metaphorical language) according to a set of pre-determined rules (even if these can be later amended or more fully altered), while on the other hand, the idea that we can accept or reject metaphors based on issues aside from content, including context and expectation of meaning. An unusual or bizarre comparison might make sense where we look for (or expect) sense, for example from a person who I know uses and *understands* the same language as me, and not where we expect little sense to be found, such as in the babbling of a small infant. The expectation of meaning is an important element in drawing these sorts of comparisons, and can sometimes be unfair in the expectation (or not) of meaning and importantly for this discussion, in what is then accepted as either meaningful or indeed successful.

1 UNDERSTANDING *ME*

In proposition 6.54 of the *Tractatus Logico-Philosophicus* [1] (first published in 1921), Ludwig Wittgenstein states of his project: "My propositions serve as elucidations in the following way: anyone who understands me eventually recognises them as nonsensical." Understanding what Wittgenstein meant by these simple yet enigmatic words has dominated certain sub-sections of Wittgenstein scholarship. In one particular strand of scholarship, discussion centres on that little word "me" and why Wittgenstein did not instead write, "understands *them*" in reference the propositions of the text, as per the second half of his statement. Understanding why this might be important will have an impact on the arguments of this paper.

This paper picks up this discussion in order (in the first instance) to lend support to the interpretation offered by Cora Diamond [2, p. 151] whereby to understand this statement requires that we understand both Wittgenstein as well as his words. She claims this is a clear indication that Wittgenstein wanted to "draw attention to a contrast between understanding a person and understanding what the person says." This, she says, is pivotal for our understanding of the instruction that Wittgenstein presents in these words, which is that we should recognise the propositions of his text as nonsensical. This seeming contradiction puzzles, delights and infuriates readers

often in equal measure. How can the propositions be taken as nonsense if we can in fact understand them?[1] In following Diamond's solution we dissolve the contradiction since we can accept (if we like) that the *content* of the *Tractatus* is nonsense, while simultaneously acknowledging that we have somehow understood this nonsense because we understand the person. Thus we come to 'understand not the propositions but the author' [2, p.155].

One objection to this view, such as is offered by Priest [6, p. 150], argues that the conclusion of the nonsense uttered (and so-called) in the Tractatus results only in a contradiction. Regardless of context, it is clear that we have at some point understood nonsense—it must have made sense to us—otherwise what did we understand? Yet Diamond's reply to such arguments is that although we have seemingly understood what is later termed *nonsense*—Priest is not wrong in this—this does not mean it is any the less nonsensical. In fact, she holds [2, p. 150] it is not that we understood the nonsense propositions in the first instance, thus generating a contradiction, but rather that "in recognising that they are nonsense, [we] are giving up the idea that there is such a thing as understanding them". She concludes, "What Wittgenstein means by calling his propositions nonsense is not that they do not fit into some official category of his of intelligible propositions but that there is at most the illusion of understanding them". The reason for this approach, she claims, hinges on seeing Wittgenstein's request that we understand *him* as indicative of his personal engagement with those who talk nonsense, something she later describes [2, pp. 157-58] as requiring imagination:

> My point then is that the Tractatus, in its understanding of itself as addressed to those who are in the grip of philosophical nonsense, and in its understanding of the kind of demands it makes on its readers, supposes a kind of imaginative activity, an exercise of the capacity to enter into the taking of nonsense for sense, of the capacity to share imaginatively the inclination to think that one is thinking something in it. If I could not as it were see your nonsense as sense, imaginatively let myself feel its attractiveness, I could not understand you. And that is a very particular use of imagination.

This recourse to imagination is perhaps surprising (and is not itself uncontroversial or indisputable), but it is helpful for when we consider ideas about analogy, and more specifically metaphor, to which we now turn.

2 AN EXPECTATION OF MEANING

The discussion above offers a way in which to begin to see that the possibility of meaningful language and understanding relies on such words having been uttered by a semantic language-user (in the above example, Wittgenstein). In fact, the crux of this paper, where metaphor is concerned, is that people (lay- and scholars alike) would not have been so interested in the enigmatic aphorism noted above if the speaker had not been a person. If Wittgenstein had instead been the name of a complex computational program that uttered such words, it is unlikely the discussion about them would have lasted nearly a hundred years. More simply: if Wittgenstein had been a machine, we'd likely have ignored the odd turn of phrase, or perhaps described as a superficial error.

This approach to understanding an author over (or at least as well as) her/his words may seem in contrast to Barthes [7] and related post-structuralist ideas about the independence of text from an author (commonly referred to, in reference to Barthes, as *the death of the author*). However, the *death* of an author does not thereby presume *no* author. Instead the argument is a complex response to some traditional notion of the *individual*—the author—as the final locus of meaning. In other words, the authorial voice as judge, authority, "always finally the voice of one and the same person, the author, which delivered his 'confidence'" [7]. As he notes elsewhere, the crux is to do with culture, which is akin to context that I describe above:

> We know that a text does not consist of a line of words, releasing a single "theological" meaning (the "message" of the Author-God), but is a space of many dimensions, in which are wedded and contested various kinds of writing, no one of which is original: the text is a tissue of citations, resulting from the thousand sources of culture. [7]

The text and the author exist *simultaneously* on this account, and in this way, the text has as much authority as the author, the reader, and any other voice in dialogue about the text. "In this way is revealed the whole being of writing: a text consists of multiple writings, issuing from several cultures and entering into dialogue with each other" [7].

While this would seem to stand in tension to the discussion about Wittgenstein's text above—where we should understand Wittgenstein in order to understand the text—in fact we can see the same impetus of the centrality of the reader's voice in Wittgenstein's work also. In the Preface to the *Tractatus*, Wittgenstein says, "This book will perhaps only be understood by those who have themselves already thought the thoughts which are expressed in it—or similar thoughts. It is therefore not a text-book. Its object would be attained if it afforded pleasure to one who read it with understanding" [1]. Furthermore, the claim to a singular authorial voice is never made. As he explains a little further along, "How far my efforts agree with those of other philosophers I will not decide. Indeed what I have here written makes no claim to novelty in points of detail; and therefore I give no sources, because it is indifferent to me whether what I have thought has already been thought before my by another" [1]. Similar to Barthes, the authorial voice is not to be considered that of an individual in any absolute sense, or a decontextualised authority. Instead we can take Wittgenstein's words, his contribution to the dialogue, as direct engagement with, and an imploring to, the reader to understand. His request at the end of the text that we understand *him* specifically, is as much a part of this collective, contextual engagement, as Barthes' claims that,

> the unity of a text is not in its origin, it is in its destination; but this destination can no longer be personal: the reader is a man without history, without biography, without psychology; he is only that someone who holds gathered into a single field all the paths of which the text is constituted. [7]

This is not to say that there are no differences between their respective views however, and indeed I will return to this in Section 3 below.

From this we arrive back at the discussion above regarding context, and to this we can add shared experience, culture, history and meaning. For these reasons I offer the claim that the

[1] It is important to clarify that the author does not in fact take at face value the nonsensicality of the propositions in Wittgenstein's text, but this argument is outside the scope of this paper.

possibility of a meaningful metaphor relies on a context within which the language is embedded, such as described by Wittgenstein in a later work [3] in terms of *language-games*. A language-game on Wittgenstein's account brings "into prominence the fact that the speaking of language is part of an activity, or of a life-form" [3, §23]. As Monk [8, p. 330] explains, the purpose of language-games is "to free ourselves from the philosophical confusions that result from considering language in isolation from its place in the 'stream of life'".

These descriptions of Wittgenstein's approach reflect a broader polemic against a position that assumes we can somehow view things *sub specie aeterni*. Wittgenstein viewed such perspectives as negligent of one's own, necessarily earth-bound, position. In a note written to Sraffa in 1935, he describes the irritation caused by the thinking of "Cambridge people", that he formulates as follows: "Here are people who try to speak in a queer way 'impartially' about things, they pretend to be able to slip out of their own skins and they speak as though they could understand everybody's feelings, wishes, tendencies etc." [9, p. 235 n. 7]. In line with this perspective, the notion of a language-game evokes a sense in which, understanding language requires some sort of involvement in it. It is the connection with a game that draws this out, for we can only understand a game (how it is played, what its rules are, what significance it has) through engaging with it in someway. We can no more view our language from an objective perspective than we can *slip out of our skin*. This argument provides some basis to the centrality of context for metaphor, because the last metaphor includes translatable qualities (in terms of seeing things from the point of view of another), but it also has other qualities that make sense from the perspective of an embodied person. Simply: if you've never had skin, can you really understand the *ick* factor that comes when you think in more detail about what it would be like to slip out of it. Let alone to slip *into* the skin of another.

Let us consider another example (which formed part of the title for the first incarnation of this paper): *to find your feet*. In a very general sense the metaphor points to the sense of finding ones way around, or getting to know how things work, where things are, or to familiarise yourself with something in either general or specific terms. The literal meaning makes little sense, since someone with feet and legs will find their feet at the end of their legs where they always have been.[2] In this metaphor, I suggest that this *your* is embedded, meaning*ful*; and ineliminable. This does not mean that context is limited to a singular subjective experience. As Barthes and Wittgenstein both describe, our (linguistic) experiences are shared. Even in vastly different experiences there can be found many sorts of overlap. For instance, one person's experience of a rare or unusual illness does not preclude another person (who has not experienced that same illness) from understanding *something* about what it is to be ill. Illness is not unique, though of course each illness may engender a different kind of experience. Nevertheless the experience of illness *per se* is important to understanding the qualitative experience of illness, just as the experience of skin adds a particular quality of understanding the metaphor offered above. This is not to say that *all* understanding is impossible without it, but rather that the understanding will be qualitatively different, as well as more difficult.

From this we arrive at the crux of the argument, which is that the capacity for understanding arises from experience, and more specifically the very possibility of that experience. As Kant explains [10, B137/138],

> The synthetic unity of consciousness is, therefore, an objective condition of all knowledge. It is not merely a condition that I myself require in knowing an object, but is a condition under which every intuition must stand in order *to become an object for me*. For otherwise, in the absence of this synthesis, the manifold would *not* be united in one consciousness.

To put this another way, the very possibility of experience is wedded to the possibility of my ability *to* experience. For the purposes of this argument, consciousness here can be replaced by understanding, since the possibility for understanding *metaphor* on this account relies on the condition, or capacity *for understanding*. And understanding, wedded as it is to context, and more broadly experience, is poorer if not embodied and embedded. What this means for computational metaphor is our next concern.

3 COMPUTATIONAL METAPHOR

This argument has been offered in defence of a contextual, experiential, semantic understanding of metaphors. We have not, so far, given consideration to the possibility of computational metaphor, and indeed in stating these words my position is already (partly) declared. I do not doubt that there can be such a thing as computational metaphor (just as in [4] I did not doubt the possibility of computational creativity), but once again I offer the caveat that what it would mean to be *successful* in a computational metaphor (hereafter *c-metaphor*) is not going to be simple, and includes indicators such as appropriateness or even what is acceptable, but more than this it includes the issue of judgement.

Returning to Barthes, we have the question of whether a metaphor stands in judgement on its own, or whether we also judge its origin and what we think it represents. For instance, if I write here about the experience I had this morning drinking coffee, and I want to do this because I want you to know that the coffee I drank improved my mood and my experience of writing this paper, then I would do this because I wanted you to know something(s) about me. This includes things about my mood, my preference for coffee in the morning(s), my experience of writing this paper, and of all the combinations that these elements produce. In so doing my primary motive would not be that you should know something about coffee separate to me and to my experience, especially as I as author chose this example purposefully. Instead, I would want you to know something about *me*. This is no different to conversations that happen about coffee outside of an academic paper. Of course, not all use of words either inspires, requires, or expects this sort of meaning (which is why I think that Barthes is right to be suspicious of the individuality of the author-god), but in this case, as in many other cases, the individual here (me) wants the reader (you) to know something about my experience of the world. If I use a metaphor to illustrate this, say, *this morning's shot of coffee,* then I would highlight both the literal size of the coffee (espresso sized, akin to a shot-sized measure of alcohol), as well as the medicinal quality of having my shot of caffeine. In this way I am pointing to my experience of coffee more generally and in a way that I hope would be familiar to you the reader. Nevertheless I would not want to divorce this metaphor, nor the description that came before it, from my own personal experience this morning. Not because I am an egomaniacal author with god delusions, but

[2] In exceptional circumstances, for instance because of a neurological disorder, or an impairment of proprioception, we can imagine someone experiencing a sense of not knowing where their feet are (or even that their feet are their own, rather like in *alien hand syndrome*).

because in the use of a personal experience I quite liked the idea you might understand *me* as a result. Which brings us back to Wittgenstein.

The aphorism at 6.54 [1] does in fact end with an analogy about a ladder, and it's worth a little more consideration:

> My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them—as steps—to climb up beyond them. (He must, so to speak, throw away the ladder after he has climbed up it.). He must transcend these propositions, and then he will see the world aright. What we cannot speak about we must pass over in silence.

What is particularly interesting about this metaphor is that it is preceded by that word *me*. My argument on this is that, rather like my description of the coffee, the metaphor offered by Wittgenstein cannot be divorced from the author. This is not to say that the interpretations that arise from the text must therefore be ordained by the Wittgenstein-god (since this is both unlikely as well as unnecessary), but rather that the experience that Wittgenstein had with the text, and with the ideas and metaphors he offers, should instead be part of the rich interpretative experience that comes from reading those words. This includes the image of the ladder and all that it might represent. Especially if you've read a lot of Wittgenstein.

Which brings us to c-metaphors. While these can of course satisfy some requirements of metaphor, including claims to novelty, utility, new aspects on the familiar, these descriptors are judged according to a context external to the computer's own capacity, and do in fact follow our own values. Added to which, these values (e.g. of novelty) and utility can contrast with other features of success, for instance, understanding what is trying to be communicated. What, for instance, would a program want to communicate and why? What would a program know of coffee, of skin, of ladders?

In simple terms, do we value a novel metaphor if we do not believe a person (or, in this case, a program) has any idea—including understanding or experience—of the individual components, let alone the comparison being drawn? If, for instance, I had offered the metaphor about coffee to you over lunch, and you happened to know me well enough to know I do not in fact drink coffee,[3] then some value of the metaphor may be lost or at least compromised. We expect that metaphors that reflect an experience have at least some basis in the user's experience otherwise they lose their potency as a basis for communication (as opposed to just literary word play).

This follows especially for unusual or bizarre comparisons that make sense where we look for or expect sense, but not where we might expect little sense (for instance in the babbling of a very young infant). Kingsey Amis' description of a hangover in *Lucky Jim* is one such example, and (to my mind) one of the finest:

> Dixon was alive again. Consciousness was upon him before he could get out of the way; not for him the slow, gracious wandering from the halls of sleep, but a summary, forcible ejection. He lay sprawled, too wicked to move, spewed up like a broken spider-crab on the tarry shingle of the morning. The light did him harm, but not as much as looking at things did; he resolved, having done it once, never to move his eye-balls again. A dusty thudding in his head made the scene before him beat like a pulse. His mouth had been used as a latrine by some small creature of the night, and then as its mausoleum. During the night, too, he'd somehow been on a cross-country run and then been expertly beaten up by secret police. He felt bad. [11]

My faith in this description of a hangover is partly borne out by my own experiences, yet had I not had those, then it would be based in a judgement of the author's, or at least the character's own knowledge, and here it requires not only that we understand the words, but that we understand them meaningfully. The above description by Jim is what it is to have a hangover in his view, as perhaps for Amis, and in terms of the rest of the novel, the description is in kilter. We can of course measure the success of a metaphor based on content, or according to any number of rules, whether these are pre or post hoc, amendable, or alterable, but we can also accept as well as reject metaphors based on context and expectation of meaning, which includes both judgement and bias. If the description of the hangover above had come from someone that you knew to be teetotal, you might still accept its accuracy as a measure of success, but again, the value of the metaphor might be compromised.

If this seems arbitrary or even unfair, I would be inclined to agree. But it's no more arbitrary or unfair than the decisions or processes by which terms either become or cease to be colloquial, slang or popular. What is considered either appropriate or acceptable in ordinary language is also tricky, including where highly creative language-use can muddy the waters of ordinary language substantially (not least where profanities are concerned). It is also partly because of the role that *expectation of meaning* creates. As I discuss above and elsewhere [4] [5], meaningful language-games in Wittgenstein's terms require not only a successful meeting of *rules*, but also a willing on the part of participants to *recognise* other speakers as meaningful language-users. In the case of proposition 6.54 above, it is precisely because scholars expect meaning to be found, that the search for a meaning is considered worthwhile.

C-metaphor construction, interpretation, use, and so on, is not impossible or even unlikely. Whether these metaphors are accepted, adopted or even considered worth paying attention to, however, remains to be seen. Even if the c-metaphor is interesting or impressive, this does not strike me as any more meaningful than when a very small child stumbles across a successful metaphor without really understanding the words or the implications of the word order. This is not to say that they absolutely did not understand, but then again, this is easier to resolve with a program than with a small child, since children do become meaningful language-users.

Where language-use is disembodied and decontextualized, the concept of the language-game makes little meaningful sense. Indeed the metaphor of the *game* is particularly helpful, since it points to the ideas of participation and mimicry. Both are key in the learning and using of language in a meaningful way. As a result, we may not accept a metaphor as creative or even useful if we do not believe the person (or program) has any idea (whether meaningful understanding or experience) of the individual components, let alone the comparison being drawn. Just as we might have doubts about the non-coffee drinker's, or teetotaller's use of certain metaphors about either tea or alcohol. This is not to say we'd necessarily reject the metaphor, but only that we may doubt the success of the utterer or even of the uttered as a result.

[3] In fact I do drink coffee, but in a thought experiment anything is possible.

4 CONCLUSION

This paper has sketched out an argument about metaphor, which remains in its infancy but which contains a number of propositions. The first is that for metaphor to be meaningful both context and embodied experience is required. These add colour (experience, meaning) to words, through which we come to understand and interpret the words themselves as well as those who utter them. Where this is missing, a crucial element of communication is thereby also missing. The question thus becomes: if you've not experienced colour, then can you really *understand* the metaphor I've offered above?

The author has not sought to suggest that words cannot have a meaning without context. Indeed there are many examples of this in all kinds of places (including on walls). Nor is it the argument that all words that are spoken or written must have an individual intention towards a particular meaning. There is sufficient evidence against such a claim, and Barthes' discussion of the author-god provides some sense of this. The author also finds it acceptable to say that language, at least in terms of signs, can be manipulated without a language-user, though I rather agree with Searle on this point that this can be described in terms of syntax rather than semantics [12].

Instead the author has sought to show that the expectation of a context, an intention, or a speaker is central to finding meaning in words, and particularly in a metaphor or other creative language. I bet you imagined the author as someone who drinks coffee at least once during the reading of this paper, and if you did then you have begun to understand *me*, or at least me as coffee-drinker. Of course this assumes you know about coffee, and have imagination, but I'm happy to assume this about the reader, and to imagine what it might be to be *you*.

ACKNOWLEDGEMENTS

I am grateful to the organisers, reviewers, and participants of the IAS/AISB Joint Workshop 2014: "Figurative language: its patterns and meanings in domain-specific discourse", where I presented the first incarnation of this paper. The constructive feedback both before and during that presentation helped to shape its direction, and it is this revised account that I wish to present at AISB 2015. All mistakes and obsessions with coffee are the author's own.

REFERENCES

[1] L. Wittgenstein, *Tractatus Logico-Philosophicus* (Trans. D. Pears & B. McGuinness). London: Routledge, 1974.

[2] C. Diamond, Ethics, Imagination and the Method of Wittgenstein's Tractatus. In A. M. Crary & R. J. Read (Eds.), *The New Wittgenstein* (pp. 149-173). London: Routledge, 2000.

[3] L. Wittgenstein, *Philosophical Investigations*. Oxford: Blackwell (Trans. G. E. M. Anscombe), 2001.

[4] Y. J. Erden, Could a created being ever be creative? Some philosophical remarks on creativity and AI development, *Minds and Machines*, 20(3): 349-362, 2010.

[5] Y. J. Erden, The 'simple-minded' metaphor: Why the brain is *not* a computer, via a defence of Searle, AISB Proceedings, 2013.

[6] G. Priest, *Beyond the Limits of Thought*. Oxford: Clarendon Press, 2002.

[7] R. Barthes, The Death of the Author, (Trans. Richard Howard), New York: Hill and Wang, 1977. Available at: http://ww2.valdosta.edu/~tthompson/ppts/3060/spring2013/death_authorbarthes.pdf (Accessed 20/03/15)

[8] R. Monk, *Ludwig Wittgenstein: The Duty of Genius*. London: Vintage, 1990.

[9] B. McGuinness, What Wittgenstein got from Sraffa. In G. Chiodi & L. Ditta (Eds.), *Sraffa or an Alternative Economics*. Basingstoke: Palgrave Macmillan, 2007.

[10] I. Kant, *Critique of Pure Reason* (trans. N. K. Smith), London: Macmillan press, 1933 (2nd ed.).

[11] K. Amis, *Lucky Jim*. London: Penguin, 1954.

[12] J. Searle, *The Rediscovery of the Mind.* Cambridge, Massachusetts: M.I.T. Press, 1992.

Automatic Metaphor-Interpretation in the Framework of Structural Semantics

Christian J. Feldbacher[1]

Abstract. Given that metaphors can be important parts of arguments and that the common methods for evaluating literal claims and arguments are not (directly) applicable to metaphorical ones, several questions arise: In which way are metaphors important? How do metaphorical premises of an argument support its conclusion? What is an adequate evaluation procedure for metaphorical claims and arguments? In this paper we will give answers especially to the first and second question and indicate how an answer to the third question might look like. Metaphors in arguments—so our analysis—introduce some very general assumptions about the domain of investigation and these general assumptions—spelled out explicitly—are in support of the conclusion of the argument. To render our analysis more precisely we will outline an implementation of automatic metaphor recognition and interpretation with the help of structural semantics. By applying such an implementation it is aimed at reducing the question of evaluation to that one of evaluating by logical or probabilistic means literal arguments.

Keywords: metaphorical argumentation, automatic metaphor recognition, automatic metaphor interpretation, structural semantics

1 Objective

"Religious beliefs are viruses of the mind."—this is a popular metaphor used to argue against religious belief. Metaphors often play an important role in such arguments. They are not only used to attack, e.g., opposing claims, but also to explain why a phenomenon as, e.g., religion has a specific property—here: is so wide spread and firmly established in society as well as significantly involved in cultural processes. In order to analyse such arguments properly, one is in need of an evaluation method for metaphorical arguments. In this paper we are going to sketch a first approach by assuming a reductive stance towards the evaluation of metaphorical arguments. As a reductive stance we propose to first translate metaphorical arguments to literal ones and then analyse them by the ordinary means of logic and probability theory. In especially we are going to sketch our intermediate results on:

- Metaphor recognition
- Metaphor interpretation
- Automation of metaphor recognition and interpretation

2 Analyzing Metaphorical Claims and Arguments

Metaphorical claims and arguments are used quite frequently, even in scientific contexts. The common methods for evaluating literal claims and arguments are not (directly) applicable to metaphorical ones. So one needs an evaluation procedure for metaphorical claims and arguments. Such a procedure may be reduced to classical evaluation procedures for arguments with expressions in literal meaning as follows:

1. Analyze the metaphorical expressions. Outcome of this process is a list of expressions possibly used as metaphors.
2. Find out implicit claims (hidden assumptions). Here we get as outcome a reduced list of such expressions and a list of claims using this expressions.
3. Reconstruct the metaphorical claim or argument. The Outcome of this process is a list of claims containing expressions in literal use only.
4. Evaluate the reconstructed claim or argument using common methods. This is just the standard procedure of evaluating arguments with literally used expressions only.

What is needed for evaluation of metaphorical arguments in the first place, is a method of analyzing and interpreting metaphors which is the main objective of this paper. With 'literal' we mean here the possibly manyfold meaning of an expression that is listed in natural language dictionaries. We intend here only a very rudimentary treatment and incorporation of such meanings, as is present, e.g., in word clouds.

2.1 Simple Accounts of Analyzing Metaphors

Traditional accounts of analyzing metaphors are, e.g., the so-called *substitutional view* (cf. [6] and [3]):

- Metaphors of the form 'X is Y' can be reduced to literal statements of the form 'X is Z, where 'Z' is a literal substitute of 'Y'.
- The metaphor is primarily about X.

and, e.g., the so-called *comparison view* (cf. [4]):

- Metaphors of the form 'X is Y' can be reduced to literal statements of the form 'X is like Y (in being Z)'.
- The metaphor is just as well about X as about Y.

Problems of the substitutional view are to be found in an adequate characterisation of synonymity as is needed in order to figure out adequate substitutivity. Problems of the comparison view lie in the question of how to interpret the likeness-relation between the relata. For this reason more sophisticated accounts were introduced.

2.2 More Sophisticated Accounts of Analyzing Metaphors

A little bit more sophisticated is the so-called *interaction view* of [1]. According to this view, metaphorical usage of language makes some

[1] Duesseldorf Center for Logic and Philosophy of Science, Germany, email: christian.feldbacher@uni-duesseldorf.de

implications expressing interactions between the relata. A heuristics to figure out the literal meaning of an expression is as follows:

1. A metaphor of the form 'X is Y' is given.
2. Construct a list of associated commonplaces w.r.t. the secondary subject:
   - C('Y') = $\langle$'Y is Y_1',. . ., 'Y is Y_m'$\rangle$
3. Construct from CP('Y') a list of implications by transferring the commonplaces of the secondary subject 'Y' to the primary subject 'X' by help of an interpretation function I.
   - I('X','Y') = $\langle$'X is Y_1',. . ., 'X is Y_m'$\rangle$
4. Select a list of relevant implications from I('X','Y') by means of an appropriate strategy:
   - RI('X','Y') = $\langle$'X is Y_{i_1}',. . ., 'X is Y_{i_k}'$\rangle$
5. Then 'X is Y_{i_1} and . . . and Y_{i_k}' is a possible interpretation (paraphrase) of the metaphor 'X is Y'.

A problem of the interaction view is this: It is not clear how to figure out the commonplaces w.r.t. a subject and then figure out a set of relevant implications. Also the heuristics presented here starts from a situation where metaphors are already identified. So we would like to offer a new account for metaphor recognition and interpretation that makes Black's presupposed concepts more explicit.

To sum up: Problems of the traditional accounts are:

- The substitutional and the comparison view are too vague and non-constructive.
- Black's interaction account is more adequate. But: If automated, it requires a large amount of manual intervention. There is no general method of determining commonplaces and selecting relevant implications.

Our account aims at the following task:

- To develop an adaption of the interaction account that can be automated so that it does only little or not at all require manual intervention.

For this purpose we want to use structural semantics.

3 Automatic Metaphor Interpretation

Automatic metaphor interpretation is a field of linguistics and computer science, concerned with software based analysis of metaphors. There are two main tasks of automatic metaphor interpretation ([cf. 8, p.1029]):

1. Automatic metaphor recognition
2. Automatic metaphor interpretation

Both tasks are closely connected: Simplified speaking, a metaphorical expression in a context is an expression used not in its literal meaning in the context. To give an interpretation of a metaphorical expression is to paraphrase it with expressions used in their literal meanings ([cf. 8]).

3.1 Metaphor Recognition

What does it mean that an expression in a context is not used in its literal meaning?

Definition 1 (very general criterion) *An expression is a metaphorical expression in a context iff*

1. *the context is assumed to be semantically perfect and*
2. *if the expression is used in its literal meaning, then the context is obviously semantically imperfect.*

E.g.: 'Achilles was a lion in the battle.'. If we take 'Achilles' to be understood in its literal meaning, i.e. talking about a human, and also 'lion' in its literal meaning, i.e. talking about a non-human animal, then the sentence (context) is obviously wrong (semantically imperfect). Hence, at least one of the expressions is a metaphorical one.

There are three very central notions used in the criterion:

- 'context'
- 'semantical perfectness'
- 'obviousness'

The context in our example was a sentence. But there are many more other types of contexts possible:

- bottom-up, e.g.: arguments, argument hierarchies
- top-down, e.g.: term-forming expressions (e.g. definite descriptions, functors), predicate-forming expressions (e.g. lambda-expressions) etc.

Depending on the context there are different types of semantical perfectness/imperfectness:

- arguments: valid/invalid, strong/weak
- sentences: true/false, adequate/inadequate, etc.
- term-forming expressions: referential/non-referential

With the help of our general characterization we can provide a systematic formal categorization of metaphors:

1. Propositional metaphors. With sub-species, e.g.:
   (a) Identity metaphors: $t_1 = t_2$ ('Juliet is the sun.')
   (b) Monadic predicative metaphors: $P^1(t)$ ('Juliet is brilliant.')
   (c) Polyadic predicative metaphors: $P^n(t_1, \ldots, t_n)$ ('Juliet is Romeos manna.')
   (d) General subjunctive metaphors: $\forall x(Px \rightarrow Qx)$ ('Religions are viruses.')
2. Term-forming metaphors. With sub-species, e.g.:
   (a) Metaphorical names: c ('Romeo' for a charming man)
   (b) Metaphorical functors: $f^n(t_1, \ldots, t_n)$ ('the heart of his beliefs')

One notion still has to be clarified: 'obviousness'. 'Obviousness' seems to be necessary in order to distinguish semantical imperfectness through metaphors from semantical imperfectness in general. E.g., to claim 'All birds can fly.' is just false, not speaking metaphorically. There are different degrees of the obviousness of semantical imperfectness:

D1 Semantical imperfectness through mixing up categories (sometimes also expressed as stating something which is neither true nor false). E.g. 'Colorless green ideas sleep furiously.'
D2 Semantical imperfectness through logical or definitional falsity. E.g. 'Sophia Loren is a star and not a star.' or 'Soldiers are machines.'
D3 Semantical imperfectness through contradicting commonplaces. E.g. 'Achilles was a lion in the battle.'

⋮

We assume that obviousness of semantical imperfectness up to the degree D3 is characteristic for metaphors. I.e.: An expression that is not recognizable in a context as a metaphorical expression up to the knowledge of commonplaces counts as being literally used in the context. To illustrate this assumption, let's take our example 'All birds can fly.'!

- '. . . flies' is defined on a set containing also birds, so there is no mixing up of categories. D1: passed. . .
- The claim is neither logically nor definitionally false (the dictionary just states: 'Birds can fly in general.' which doesn't contradict the claim.) D2: passed. . .
- The claim also doesn't contradict commonplaces since 'to fly' is even a connotation of 'being a bird'. D3: passed. . .

If we consider our example 'Achilles was a lion in the battle.', it turns out that at least one expression is used metaphorically:

- '. . . is a lion' is defined on a set containing animals (including humans), so there is no mixing up of categories. D1: passed. . .
- The claim is not logically false, but definitionally (the dictionary states two opposing characteristics for 'lion' and 'man' (as genus of 'Achilles'), namely 'non-human' and 'human') D2: not passed. . .

Our choice of semantical imperfectness up to the degree D3 is motivated by the intended automation which is based on dictionaries and semantical networks and not on "world knowledge" in general. Whether this choice suffices to identify adequately a huge set of metaphorical claims remains an empirical question settled by investigations of performances of our heuristics.

The criterion provided here does not allow us to figure out which expression is the metaphorical one. Someone could speak, e.g., about the Achilles of Homer's *Iliad*, fighting bravely the Trojans. But someone could, e.g., according to our analysis speak also about a lion fighting against a rival as bravely as Achilles did. But this kind of ambiguity, as is mentioned, e.g., also in [cf. 2, p.483,p.485], can be resolved by a non-compositional analysis of the statement in question. The question of identifying the target and the source can be decided only with respect to a broader context.

In order to decide this question, we expand our framework and use some important parts of the semiotical theory *structural semantics*, which was invented in 1966 by Algirdas Julien Greimas ([cf. 7, part.V, section on Greimas]). This is no unconventional choice since the framework of structural semantics is commonly used in literary theory for interpreting literature and importantly also for interpreting metaphors in literature.

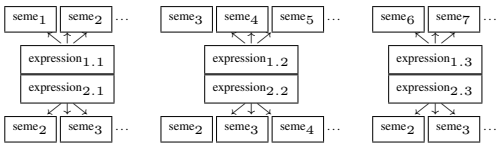
There are two important notions of structural semantics needed for our automatized metaphor recognition (and later on: interpretation):

- Seme: "The seme is the minimal unit of semantics, whose function is to differentiate significations." ([7, p.317])
- Isotopy: "Greimas defines isotopy as the principle that allows the semantic concatenation of utterances" where the "iterativity (recurrence) of contextual semes, which connect the semantic elements of discourse (sememes), assures its textual homogeneity and coherence." ([7, p.317])

Very simplified speaking one can say that:

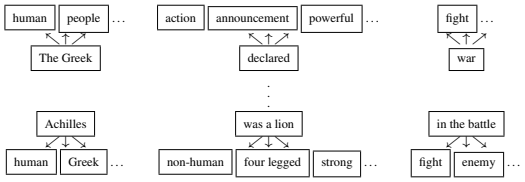
- Semes are the minimal semantical units that are mapped to expressions.
- If an expression is used in a text, then the semes of the expression are set.
- The more a seme is set within a text, the more dominant it is in the text (iteration increases dominance).
- The most dominant semes within a text are the isotopes of the text.

Example:



Isotopies: $seme_2$ and $seme_3$

Let's take 'Achilles was a lion in the battle.' with some more context:



Isotopies: 'human', 'fight'

As can be seen, one seme of 'Achilles' is an isotopy, whereas no seme of 'lion' is an isotopy. Since expressions are used normally literally (default), it is likely that metaphorical expressions do not contain isotopies.

We therefore expand the conditions of the criterion for metaphor recognition within the framework of structural semantics:

Definition 2 (more detailed criterion) *An expression is a metaphorical expression in a context iff 1, 2 (of definition 1 above) and:*

3. *No seme of the expression is an isotopy with respect to the overall context. (In comparing expressions one may take the degree of dominance of the expressions' semes for a comparison.)*

The framework of structural semantics is not only useful for the identification of metaphors, but also for their interpretation. In the following we will provide a short sketch of metaphor interpretation in this framework.

3.2 Metaphor Interpretation

Once we have identified metaphors, the question arises of how to paraphrase them in a way such that the paraphrase is non-metaphorical. Just to replace the metaphorical expression by all its semes is inadequate, since this would just make the semantical imperfectness still more obvious (D3⇒D2⇒D1⇒). E.g.:

- If we replace the metaphorical expression 'lion' . . .
- . . . in the sentence 'Achilles was a lion in the battle.' . . .
- . . . by its semes 'non-human', 'four legged', 'strong', 'animal' etc. . . .
- . . . then we end up indeed with a purely literal paraphrase, . . .
- . . . but on cost of inadequacy:
- 'Achilles was a non-human four legged strong animal in the battle.'

What is needed is some kind of relevance filter, dropping out 'non-human', 'four legged', 'animal' and keeping 'strong'. Here again the *iteration increases dominance* principle of structural semantics is of

some use: The more dominant a seme of a metaphorical expression is within the overall context, the more likely it is to be of relevance.

If the overall context does not increase a seme's degree of dominance, then the seme is less likely to be recognised as a relevant part of a metaphor. And also the other way round: The more dominant a seme is, the easier it is to be recognised as a relevant part of a metaphor. So, for the interpretation of a metaphor one just has to replace the metaphorical expression by the dominant semes to get a literal paraphrase.

3.3 A Fundamental Proviso

Quite common is the point of view that a reductive stance as ours is fundamentally wrong since linguistically and psychologically seen a relation of reduction should be assumed at most the other way round: It is not the literal meaning of an expression we should start of, but a metaphorical one (cf., e.g., [5]). Also Cohen and Margalit claim, e.g., that "it is psychogenetically more illuminating to view literal patterns of word-use as the result of imposing certain restrictions on metaphorical ones, than to view metaphorical patterns as the results of removing certain restrictions from literal ones" ([2, p.470]). Heading into this direction by arguing against the possibility of reducing metaphorical expressions to literal ones, Cohen and Margalit argue as follows—[cf. 2, p.471] (simplified and slightly changed):

1. The meaning of a complex expression is determined by the meaning of its components alone, where the meanings of the basic components are described in dictionaries. (Principle of compositional semantics)
2. Hence: The meaning of a metaphorical expression is either described in a dictionary directly or is determined by meanings of its components described in a dictionary. (1)
3. Dictionaries usually record the current use of expressions whereas metaphors are usually innovative, i.e. an expression's metaphorical usage is new. (general assumption)
4. Hence: The meaning of a metaphorical expression is neither described in a dictionary directly, nor is it determined by—in such a way described—components (otherwise it wouldn't be innovative). (3)
5. Hence, metaphors cannot be analysed compositionally. (1, 2–4)

This argument may be seen as counterargument to a reductive stance of metaphors to literal expressions by identifying compositionality with reducibility. Again simplified speaking, Cohen and Margalit propose instead of such a reduction the following analysis—[cf. 2, pp.476ff]: The meaning of an expression is learned inductively by uttering combinations of expressions and taking into account the affirmative or negative responses of trained language users. In doing so one may figure out that, e.g., generally 'shout at me' may go together with 'Peter', but not, e.g., with 'car'. So, we end up with a semantical hypothesis like 'shout' names or describes an action involving as variables a loud tone etc. and is affected, e.g., by the live/non-living variable (according to general usage non-living entities don't shout). Metaphorical usage of 'shout', as, e.g., in 'The car shouted at me.' consists then just in "removing any restrictions in relation to certain variables from the appropriate section or sections of its semantical hypothesis" ([cf. 2, p.482]). So, the psychological relation seems to be as follows:

- Expressions are learned by such combinations and taking into account affirmative or negative feedback.
- Learning of an expression consists in figuring out the relevant variables and putting restrictions on them.
- By this we end up with literal meaning(s) of an expression.
- Speaking in metaphors consists just in relaxing such restrictions again, i.e. in going some steps back in the whole process.

We think that our account is not in contrast to this point of view. Regarding Cohen and Margalit's argument above our approach also denies compositionality, but we still stick to reducibility: According to our theory the correct interpretation of a metaphorical statement is not only based on the meaning of its components alone. Rather it is based on the meaning of its components and the contextually dominant-set semes. By this Cohen and Margalit's claim about the fundamental ambiguity of statements like 'That old man is a baby.' also remains for our approach: "Either its subject is literal and its predicate metaphorical, or vice versa" ([cf. 2, p.483]). Considering the statement alone, 'That old man is a baby.' may be paraphrased adequately by 'That old man behaves like a baby.' or 'That small little thing with this face wrinkled like an old man is a baby.'. But considering it with respect to a context with dominant-set semes as, e.g., the semes of 'experienced', 'wise' etc. in the former and that of 'tiny', 'newborn' etc. in the latter case allows for a disambiguation.

So, to sum up the proviso one may say that our approach also denies the adequacy of compositional reduction, but not that of context-dependent reduction.

3.4 Heuristics for an Automatic Analysis

For automatic metaphor recognition and interpretation in a similar line as described in [10], [9] we used syntactic and semantic databases—at this time only for a text corpus in German (Canoo, Duden, in the future: GermaNet). The flow diagram can be summarized as follows:

- *Basic analysis*
  1. Get the syntactical information of the expressions! (Canoo)
  2. Transform the expressions into their normal form: Nom.Sg/Inf! (Canoo)
  3. Extract the semes of the expressions! (GermaNet)
  4. Extract the connotations of the expressions! (Duden)
- *Metaphor recognition*
  1. Check whether there are any opposing semes or connotations! (Synonym- and Antonym-Databases)
  2. If so, check which semes are more dominant! (Preceding Analysis)
- *Metaphor interpretation*
  1. Extract the most dominant semes! (Preceding Analysis)
  2. Transform them into the syntactical form of the metaphorical expression! (Canoo)
  3. Replace the metaphorical expression by a concatenation of these transformations!

4 Conclusion

In this paper we indicated how two main tasks of theories on metaphors, namely metaphor recognition and metaphor interpretation, may be approached by an automatized analysis. For this purpose the so-called *interaction account* of metaphors served as rough model; we suggested to explicate the key-concepts of this model, i.e.

the concept of 'commonplace' and 'implication', by help of structural semantics: Commonplaces are connections between the semes of an expression and implications are figured out by a dominance operation of the context acting on the metaphorical statement under investigation. Furthermore dominance is operationalized via counting the iteration of semes. The theory is currently implemented into *Perl* for an application on a German text corpus. The implementation is still carried out and it is tried to be expanded on English text corpora too.

References

[1] Max Black. *Models and Metaphors. Studies in Language and Philosophy*. Ithaca: Cornell University Press, 1962.

[2] L. Jonathan Cohen and Avishai Margalit. "The Role of Inductive Reasoning in the Interpretation of Metaphor". In: *Synthese* 21.3–4 (1970), pp.469–487. DOI: 10.1007/BF00484812. URL: http://dx.doi.org/10.1007/BF00484812.

[3] Raymond W. Gibbs, ed. *The Cambridge Handbook of Metaphor and Thought*. Cambridge: Cambridge University Press, 2008.

[4] Paul Henle. "Metaphor". In: *Language, Thought, and Culture*. Ed. by Paul Henle. Michigan: University of Michigan Press, 1958, pp.173–195.

[5] George Lakoff and Mark Johnson. "Conceptual Metaphor in Everyday Language". English. In: *The Journal of Philosophy* 77.8 (1980), pp. 453–486. ISSN: 0022362X. URL: http://www.jstor.org/stable/2025464.

[6] Heinrich Lausberg. *Handbuch der literarischen Rhetorik*. München: Steiner, 1973.

[7] Winfried Nöth. *Handbook of Semiotics*. Bloomington: Indiana University Press, 1995.

[8] Ekaterina Shutova. "Automatic Metaphor Interpretation as a Paraphrasing Task". In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*. Ed. by The Association for Computational Linguistics. Los Angeles: The Association for Computational Linguistics, 2010, pp. 1029–1037. ISBN: 978-1-932432-65-7.

[9] Eric Steinhart. *The Logic of Metaphor: Analogous Parts of Possible Worlds*. Dordrecht: Kluwer Academic Publishers, 2001.

[10] Eric Steinhart and Eva Kittay. "Generating Metaphors from Networs". In: *Approaches to Metaphor*. Ed. by Jaakko Hintikka. Dordrecht: Kluwer Academic Publishers, 1994, pp.41–94.

Metaphorical Minds, Illusory Introspection, and Two Kinds of Analogical Reasoning

Eugen Fischer[1]

Abstract Introspective conceptions of the mind are inconsistent with recent findings from cognitive and social psychology, but remain intuitive and culturally influential. This paper builds up to a debunking explanation of intuitions which, historically, are at the root of introspective conceptions. The explanation exposes these intuitions as cognitive illusions. It shows that they are devoid of determinate meaning and traces them back to seductive mistakes at the mapping stage of analogical reasoning. The argument employs key principles of the ATT-Meta model of metaphor comprehension and a structure-mapping account of analogical reasoning. The paper argues that, as a default strategy, the comprehension of extended metaphors involves only a very restricted form of analogical inference. It shows how 'full blooded' analogical reasoning with metaphor-transcendent mappings leads to conclusions incapable of metaphorical interpretation through that default strategy. It explains why those transcendent mappings are made, and identifies a previously unrecognised fallacy at the mapping stage of analogical reasoning, the 'metaphor-overextension fallacy'.

1 INTRODUCTION

Intuitive conceptions of the mind, which frequently pass for common sense, credit us with introspective access to, and hence direct knowledge of, a wide range of mental states and processes. These intuitive conceptions have been called into question by several strands of now famous work in social psychology [1, 2], cognitive psychology [3, 4], and cognitive neuroscience [5]. These contributions have forged a new picture of everyday action, decision-making, judgment, and belief-formation: In the absence of determinate prior attitudes or information, people typically perform actions, take decisions and form beliefs due to processes of automatic cognition into which they have little, if any, insight. In many such cases, they then rationalize their actions and beliefs with reasons that do not reflect the factors that moved them. These reasons are hence of little explanatory or predictive value. Instead, rationalisations take up one of several readily available, socially accepted patterns of justification, apparently arbitrarily. The stated reasons might then as well have taken up another pattern, justifying different actions or beliefs. Where this happens, these reasons have only limited justificatory value. It is therefore scarcely an exaggeration to say that, as often as not, when people make up their minds, everything important happens at the level of automatic cognition of which we are largely unaware, and subsequently stated reasons explain nothing and justify little [cp. 6, 7].

To help assess and resolve the manifest tension between this new picture and intuitive introspective conceptions, this paper will prepare the ground for a debunking explanation of relevant "introspective intuitions", of the kind sometimes sought by one strand of current experimental philosophy [8], known as the 'sources project' [9] or 'cognitive epistemology' [10]. Students of metaphor have prominently suggested these intuitive introspective conceptions are due to unwitting use of visual metaphors [11]. Proceeding from a case-study on four key intuitions from the early modern philosophy of mind, this paper will argue that only a fallacy in analogical reasoning with these conceptual metaphors leads to the intuitions targeted and leads us to give introspection a wider scope than is consistent with the new scientific picture.

The heuristics and biases programme in the psychology of judgment has sought to explain intuitive judgments as outcomes of automatic inferences with heuristic rules which are generally reliable but generate cognitive illusions under specific circumstances [12, 13, 14]. The overtly heuristic character of the rules of analogical reasoning opens up the prospect of an in some ways analogous argument.

To set the stage, we will contrast a default reasoning strategy and a default comprehension strategy: We will consider a default strategy of analogical reasoning, as commonly conceived in cognitive psychology (review: [15]) and build up to a default strategy for motivating and interpreting fresh metaphorical language. We will build up to the latter strategy by integrating notions from the cognitive psychology of analogy and metaphor with psycholinguistic findings about the role of stereotypes in verb comprehension [16, 17], and building on key insights from the ATT-Meta model of metaphor processing [18, 19]. We will argue that a very restricted form of analogical reasoning suffices to build up, e.g., from stereotypical implications of verbs to conceptual metaphors of (roughly) the sort posited in cognitive linguistics (review: [20]) (Section 2).

By reconstructing how the default reasoning strategy can generate four key tenets of an early modern introspective conception of the mind (Section 3), we will then see how, and when, the default reasoning strategy can lead us to cognitive illusions, namely, to illusions of sense: to conclusions which cannot be interpreted with the default comprehension strategy and are therefore liable to lack determinate meaning (Section 4). We will see that this happens the moment more complex analogical inferences employ extensions that 'transcend' the extended mappings properly constitutive of conceptual metaphors. Finally, we will propose an explanation of why these extensions are made (Section 5), i.e., of why perfectly competent speakers come to overextend the conceptual metaphors at issue, namely, in non-intentional analogical inferences [21] which have been found to be involved in problem-solving [22, 23, 24; but cp. 25].

2 TWO STRATEGIES: FULL-BLOODED AND RESTRICTED ANALOGICAL REASONING

Our argument will rely on the distinction between two strategies:

[1] School of Politics, Philosophy, Language and Communication Studies, University of East Anglia, Norwich NR4 7TJ, UK. E.Fischer@uea.ac.uk

We now briefly sketch a default strategy for 'full-blooded' analogical reasoning, and then build up to a strategy for motivating and comprehending metaphorical talk, which makes use of a more restricted form of analogical inferencing.

As standardly conceived in cognitive psychology (review: [15]), analogical reasoning about a target domain TD (say, atoms) involves at least three steps: First, a model or source-domain SD (e.g. the solar system) is identified, and knowledge about it is retrieved from memory. Second, model and target are aligned, and elements of the source-model (planets, sun, relations between them: x revolves around y, y attracts x, etc.) are mapped onto elements of the target domain (electrons, nucleus, etc.), subject to semantic and structural constraints: According to influential models of analogical inference (including SME: [26, cp. 27]), we first correlate source- and target-domain elements which are *semantically similar* (which we believe to share properties or stand in the same relations), and then prune these correlations and add new ones by enforcing structural constraints including *1-to-1 mapping* and *parallel connectivity* (when mapping a relation or property onto another, also map their relata or bearers onto each other). Third, the actual inferences are made through *copying with substitution and generation (CWSG)* from a (partial) representation of the source domain SD.

Within the philosophically familiar format of inferences from a set of premises, such *standard analogical (CWSG) inferences* are governed by these three rules: Wherever the premises invoke a SD element which has been mapped onto a TD element,

1. *copy* the representations of relations and relata attached to the SD element, into a set of candidate conclusions about the TD.
2. In the candidates, *substitute* representations of SD relations and relata by representations of TD elements onto which they are mapped.
3. If no such mapping exists, copy the representation of the SD element unchanged into the conclusions ('*generation*').

This default strategy for analogical reasoning contrasts with what I will suggest is a default strategy for motivating and interpreting fresh metaphorical language (*pace* [28]).

According to the ATT-Meta model of metaphor processing, only some of the resources involved in the default strategy for 'full-blooded' analogical (CWSG) reasoning are employed in facilitating metaphorical talk [18, 29]. Two of the model's principles are particularly pertinent for our purposes:

a) Coherent mappings from a source- to a target-domain (conceptual metaphors CM) are built up from single core mappings by a few generic default processes ('vehicle-neutral mapping adjuncts').
b) The mappings obtained with these slender resources are deployed to maximum effect, namely, in interpreting metaphorical uses of expressions which literally stand for 'CM-transcendent' source-domain elements, i.e. for elements which are not mapped by CMs that are built up in this way. Such uses are typically interpreted not by adding further mappings to the conceptual metaphor but by relating the elements 'transcending' it to elements mapped by it.

Many core mappings can be obtained from stereotypical inferences we routinely execute in language comprehension: When interpreting nouns [30] and verbs [16], competent speaker/hearers automatically infer stereotypically associated attributes and consequences, in line with the neo-Gricean I-heuristic: 'Find interpretations that are stereotypical and specific!' [31]. E.g., when people see something happening, they typically know it is happening. Speakers can therefore extend the use of words (e.g., "see") to stand for the stereotypically associated consequence (the subject knows) that hearers will automatically infer, in the absence of explicit indications to the contrary. Such use turns stereotypical into necessary consequences, and defeasible pragmatic into non-defeasible semantic inferences. (You can 'see a kidnapping' without realising what it is, but cannot 'see my point' without knowing what it is.) Such 'pragmatic strengthening' [32] is one of several processes that can endow expressions with metaphorical senses in which they apply in fresh (here: non-visual) contexts [33].

Very elementary automatic analogical inferences [21] can then treat these extensions as cross-domain mappings (here: from the SD of vision to the TD of knowledge) and build up to further, related mappings, which can, in turn, motivate the metaphorical extension of further, related expressions. This happens through generic default operations which unfold, e.g., the conceptual metaphor 'Knowing as Seeing' from the core mapping (here and below, "→" represents mapping, not implication or entailment):

(1) S sees x → S knows x

These default operations can be conceptualised as the very simplest analogical inferences, namely, analogical inferences which invoke only such a core mapping and generic (e.g. logical) functions and relations which obtain across domains, and hence get mapped onto themselves. These *elementary CWS inferences (ECWS inferences)* involve

(i) only copying with substitution (CWS),
(ii) no generation, and
(iii) employ only core mappings like (1) and 'mappings onto self', which are the first mappings to be made in analogical reasoning (cp. Forbus et al. 1995).

Such elementary inferences can proceed from closed and open sentences. In the latter case, we obtain fresh mappings of relations onto relations. Table 1 gives a particularly simple example, resulting in the fresh mapping

(2) S does not see x → S does not know x

Table 1. An elementary CWS inference

	SD premise	**Operation**	**TD conclusion**
1	¬	Substitution (identical)	¬
2	S sees X	Substitution with (1)	S knows X

Other ECWS inferences yield, e.g.:

(3) It is possible for S to see x → It is possible for S to know x
(4) It is not possible for S to see x → It is not possible for S to know x
(5) X makes it possible for S to see y → X makes it possible for S to know y

(6) X makes it impossible for S to see y → X makes it impossible for S to know y

According to ATT-Meta, not only logical and modal but also temporal, causal, enabling, and disenabling relations are invoked in generic expansion of core mappings [18, 29]. A core mapping and the further mappings obtainable through ECWS inferences are jointly *'constitutive' of a conceptual metaphor* (here: 'Knowing as Seeing').

Elementary automatic inferences can follow equally automatic stereotypical or semantic inferences. Such brief *inference chains* allow hearers to spontaneously give metaphorical interpretations to further expressions. This motivates the metaphorical extension of these expressions. Consider, e.g., the extension of "beyond my ken" from its literal meaning, 'beyond my range of vision'. When something is beyond someone's ken, he typically cannot get to see it. A *stereotypical inference* hence has it that

(SI) If X is beyond the ken of S, then S cannot get to see X.

An elementary *analogical inference* (with mapping 4 above) then takes us from the consequent to:

(AI) S cannot get to know X.

Speakers can extend the use of expressions (here: "X is beyond the ken of S") to stand for the conclusions of such chained inferences ('S cannot get to know X'). A variant of pragmatic strengthening can then make these inferences indefeasible, and the new metaphorical sense conventional. Let's say that the meaning or interpretation derivable through this *two-step default interpretation strategy* is 'induced by the conceptual metaphor CM' that is used for the final analogical inference (*'CM-induced'*).

Where the strategy draws on stereotypical, rather than semantic inferences about the SD, complex expressions will thus acquire as a whole a meaning that is *non-compositional*, i.e., not a function of the meaning, literal or metaphorical, of the expression's constituent parts (here: "beyond", "ken"). Where the strategy employs semantic inferences about the SD, the fresh metaphorical meaning of a complex expression can be regarded as a function (also) of the literal meanings of its constituent parts. In neither case will the former be a function of metaphorical meanings of the latter. These constituents (e.g., "beyond" and "ken") need not have any metaphorical meanings.

In line with the second of our two principles (from ATT-Meta), the metaphorical interpretation of the expression "x is beyond my ken" does not involve reliance on a fresh mapping of the source-domain element 'ken' to the target-domain but rather a chained inference that invokes only a mapping constitutive of the conceptual metaphor. As a default, the kind of analogical reasoning involved in the use and comprehension of metaphors involves only a very restricted range of mappings: the mappings that can be obtained from core mappings through ECWS inference.

3 METAPHORICAL MINDS

As we will now see, introspective conceptions of the mind essentially rely on rather more 'full-blooded' analogical reasoning that (a) involves copying with substitution *and generation* (full CWSG) and (b) invokes both mappings constitutive of visual metaphors and further mappings that 'transcend' these metaphors. While the terminology varies slightly, seminal early modern texts work with the twin mappings (see, e.g., Fischer [34] on Locke [35]):

Mapping M: visual field → mind
Mapping N: eyes → understanding

These mappings cannot be obtained through ECWS inferences from the core mappings of visual cognition metaphors. Nor are they constitutive of other familiar conceptual metaphors that are linguistically realised in pre-philosophical English. To see this, consider the spatial-inclusion metaphor of remembering and thinking-of which is the home of many uses of "the mind": It unfolds from the core

Mapping R: X is inside a space belonging to S → S remembers / thinks of X

This personal space is typically called 'the mind'. The conceptual metaphor thus motivates saying that we 'keep' or 'have' something 'in mind' when we can think of or remember it, that things 'come to mind' when we actually think of them, and that they 'slip' or (archaically) 'go from our mind' when we forget, temporarily or permanently, etc. [34, pp.41-45]. Where mind-talk is motivated by this metaphor or visual cognition metaphors, "the mind" is used only as part of complex expressions (like "S keeps X in mind", "S's mind was empty" = "S had an empty mind", etc.) whose meanings are not a function of any target-domain meanings of their constituent parts (Section 5). In these contexts "the mind" does not refer to any distinct element of the TD. But mapping M treats the mind as such an element. Hence none of these familiar metaphors include M.

We will now show that analogical reasoning with visual cognition metaphors can take us to the key tenets of classical introspective conceptions of the mind when – and only when – it employs these further mappings which 'transcend' these familiar cognition metaphors [10, 36]. Relevant visual cognition metaphors include the metaphor 'Knowing as Seeing' discussed above (Section 2) and the metaphor 'Thinking-about as Looking-at' which motivates metaphorical talk of 'looking hard at the problem', 'looking at the issue from different angles', or 'looking at the options available'. These conceptual metaphors were extended by adding mappings M and N to them.

Relevant analogical (CWSG) inferences then proceed from source-domain truisms, as in Table 2:

Table 2. A CWSG inference with transcendent mapping

	SD premise	Operation	TD conclusion
1	S looks at X	Substitution: mapping Looking at → Thinking about	S thinks about X
2	(1) Implies (3-4)	Substitution: identical	(1) Implies (3-4)
3	X before Y	Generation	X before Y
4	Y=eyes(S)	Substitution: mapping N	Y=understanding(S)

We thus obtain (non-identical substitutions underlined, generated elements in italics):

P_1 When we look at things, things are before our eyes.
C_1 When we think about things, things *are before* our understanding.

P_2 When we look at things, things are in our visual field.
C_2 When we think about things, things *are in* our mind.

P_3 Things before our eyes are in our visual field.
C_3 Things *before* our understanding *are in* our mind.

P_4 When we look at things, we perceive things with our eyes, in our visual field.
C_4 When we think about things, we perceive things with our understanding, *in* our mind.

These intuitions generate the spatial relations 'X is before Y' and 'X is in Y' in the TD and jointly transform 'the mind' into a personal space of perception, turn 'the understanding' from a 'faculty [!] of reason, intellect, or understanding' (*Oxford English Dictionary*), into an organ of sense that peers into that space, and grant us quasi-perceptual access to the objects of our own thought – but not others'. (Sometimes, 'the understanding' gets replaced by 'the mind' which then doubles as both a space and an organ of 'inner' perception, in violation of the 1-on-1 mapping constraint.)

Crucially, *only* the new mappings N and M take us through familiar visual metaphors to these intuitions and an introspective conception of the mind. To see this, consider what conclusions we obtain through analogical inferences from the present premises when we do not employ the new fare but make do with mappings constitutive of visual metaphors for knowledge or understanding. We then get different conclusions; these conclusions do not generate any spatial relations in the TD; and when interpreted in line with the default comprehension strategy (Section 2) they do not even faintly suggest that thinking involves the use of any organ or space of 'inner' perception.

Relevant inference from P_1 yields

C_1* When we think about things, things are before our eyes.

This has a literal interpretation (which is true: when I think – or do anything else, for that matter – something or other will be in front of my eyes, and sometimes I even think about the very things then in front of me). Crucially, it also has a metaphorical interpretation motivated by the visual metaphor: When something is before my eyes, it is typically easy for me to notice (get to see). Stereotypical inference therefore furnishes the premise for an ECWS inference to the conclusion that it is easy for me to get to know or understand. This yields this interpretation of C_1*:

> 'When we think about things, things are easy to understand'

– perhaps unduly optimistic and not idiomatic, but intelligible.

Similarly, analogical inference without M leads from P_2 to

C_2* When we think about things, things are within our ken.

When something is within our ken, it is typically possible for us to see. Again, therefore, stereotypical inference furnishes the premise for an ECWS inference (with mapping 3 above) to a straightforward conclusion:

> 'When we think about things, we can understand things.'

Since none of the elements P_3 refers to are mapped by the conceptual metaphors at issue, analogical inferences with these metaphors cannot be directly made from this premise. However, P_3 itself employs phrases which have stereotypical implications in the source domain of vision: When things are before our eyes, it is easy to see them, and when things are in our visual field, it is at any rate possible for us to see them. ECWS inferences lead from the conclusions of the corresponding stereotypical inferences to an undeniable conclusion:

> 'When things are easy to understand, we can understand things'.

Finally, analogical inference with visual metaphors but without M and N does not take us much beyond P_4: Since "perceive", explained by the *OED* as 'to apprehend with the mind or senses', stands for an epistemic relation that can obtain in both the SD of seeing and the TD of cognition, it initially gets mapped onto, and substituted by, itself. We thus obtain:

C_4* When we think about things, we perceive things with our eyes, in our visual field.

But when we perceive something with our eyes, we see it. This semantic implication provides the basis for analogical inferences with core mappings of visual cognition metaphors, e.g., to the conclusion:

> 'When we think about things, we understand things.'

("…in our visual field" may be disregarded as redundant: how or where else could we possibly see things?) As in the three previous cases, we obtain a conclusion that, interpreted in line with our default comprehension strategy, does not speak of organs or spaces of inner perception.

To sum up: Analogical reasoning with visual cognition metaphors only gets us from SD truisms (like P_1 to P_4) to the conclusions (C_1 to C_4) constitutive of the introspective conception of the mind, if we make use of further mapping (like M and N) which 'transcend' those metaphors.

4 ILLUSIONS OF SENSE

We will now outline how and when the use of these further mappings M and N, which 'transcend' visual and other familiar cognition metaphors, can give rise to a particular kind of cognitive illusion: The moment it employs such 'transcending' mappings, the default strategy for analogical reasoning can systematically take us to conclusions which cannot be interpreted either literally or in line with the default strategy for motivating and interpreting fresh metaphorical talk. Barring semantic rescue through fortuitous other conceptual metaphors or metonymies, etc. these conclusions lack determinate meaning. Where they strike us as perfectly intelligible, we are subject to illusions of sense.

Our first set of conclusions, C_1 to C_4, is a case in point. In contrast with their starred counterparts, they lack metaphorical interpretations motivated by visual metaphors. They all employ at least one of two phrases we obtain when applying N and M to source-domain truisms: "before our understanding" and "in our mind". Neither has a metaphorical interpretation motivated by visual cognition metaphors: In contrast with the source-domain expression "x is before our eyes" from which it is obtained, "x is before our understanding" has no stereotypical or semantic implications in the visual SD. Hence there is nothing for visual cognition metaphors to map, and our default comprehension strategy of making ECWS inferences with mappings constitutive

of the relevant – here: visual – metaphor, from source-domain implications, gets no grip. The same holds true of "in my mind": In contrast to, say, "within my ken", it has no stereotypical or semantic implications in the source domain of vision that could furnish a premise for subsequent ECWS inference with a mapping constitutive of a visual metaphor. The two key phrases lack metaphorical interpretations motivated by visual metaphors.

They also lack literal interpretations: Today as four hundred years ago, "the understanding" ordinarily refers to a faculty. Faculties cannot be literally placed in spatial relations (like the generated relation 'x is before y'). Hence "before our understanding" cannot be interpreted literally. Below (Section 5), we will consider peculiarities of mind-talk and see that, where it is motivated by spatial or visual metaphors, "the mind" always forms part of complex expressions which have no application in the metaphors' SD and possess non-compositional meanings in TD talk. Where a constituent expression (say, "x is in y") takes "the mind" as an argument, it hence cannot be given a literal interpretation. Since C_1 to C_4 all use at least one of the phrases "before the understanding" and "in the mind", these conclusions lack both a literal interpretation and a metaphorical interpretation motivated by visual metaphors.

Other conceptual metaphors, or metonymies, may come to the semantic rescue: E.g., the core mapping R of the spatial memory metaphor (above) lets us interpret the conclusion C_2 as expressing the truism 'When we think about things, we think of things', even if thinkers will have difficulties coming up with this interpretation as long as they are using mapping M. Alternatively, we can exploit semantic entailments ('perceiving' entails 'knowing') and interpret the first part of C_4, 'When we think about things, we perceive things with our understanding' as saying, 'When we think about things, we get to know things by employing our power of reasoning', though thinkers will be unlikely to come up with this interpretation when they are using mapping N. In the absence of such fortunate coincidences (and prior to exploiting them), thinkers are unable to give determinate meaning and content to conclusions like C_1 to C_4. Subsequent *ad hoc* explications were applied inconsistently, frequently disregarded by their own authors, and fail to provide determinate meanings [34, pp.35-41].

The resulting lack of determinate meaning may be obscured by subjective plausibility: C_1 to C_4 have us posit higher-order relations between mapped and generated relations:

(C_1) *When* we *think about* X, it *is before* our understanding.
(C_2) When we think about X, it *is in* our mind.
(C_3) *When* X *is before* the understanding, it *is in* the mind.
(C_4) *When* an object of thought X *is perceived with* the understanding, it *is before* the understanding and *in* the mind.

Such deeply integrated mappings endow analogical conclusions with high subjective plausibility [37, 38]. Furthermore, the posited framework of higher-order relations facilitates inferences from and to constituent and related claims, despite their lack of determinate meaning. E.g.: If something 'is before our understanding' (whatever that might mean exactly), it 'is in our mind' (whatever that might mean here), and 'we perceive it there with our understanding' (ditto). Thinkers may thus be subject to *illusions of sense*: Since they can make various inferences from and to sentences employing these phrases, they may think that these have a determinate meaning, and that they know it, even though they cannot satisfactorily explain the meaning, or apply the phrases consistently to concrete situations.

In our examples, the lack of determinate meaning is due to the use of 'transcendent' mappings M and N. These mappings have us make substitutions within complex expressions (like "before S's eyes" or "within S's ken") that, as a whole, have stereotypical or semantic implications in the SD (e.g. 'It is possible for S to see x') that are mapped onto the TD ('It is possible for S to understand x') by a mapping constitutive of a conceptual metaphor CM. They have us, e.g., replace 'ken' or 'visual field' by 'mind', and 'eyes' by 'understanding'. These substitutions deprive the overall expression E (say, "x is within the ken of S") of the SD implications that facilitate its CM-induced interpretation in line with our default comprehension strategy (Section 2). In this sense, those mappings are *inconsistent with the CM-induced interpretation of E*.

Once metaphorical uses have become familiar or conventional, their interpretation no longer requires analogical inference [39]. The present inconsistency hence does not prevent the philosophers at issue from correctly interpreting familiar metaphorical uses of, say, "beyond my ken" or any other expression E with a conventionalised metaphorical use. The problem arises rather when we use our default strategy for analogical reasoning, in reasoning from SD premises employing a complex expression E: When we then make simultaneous use of a conceptual metaphor CM and mappings inconsistent with CM-induced interpretation of E that has a non-compositional metaphorical meaning, we will obtain a fresh conclusion that cannot be interpreted in line with our default comprehension strategy for metaphorical talk. I.e., our fresh conclusion will lack a default metaphorical interpretation. By forcing substitutions in the complex expression E, those mappings will simultaneously force generation of relations from the remaining frame, in our case the spatial relations 'x is before y' and 'x is in y'. Where such concrete relations are generated in otherwise more abstract talk (like here), literal interpretation of the resulting conclusions is likely to involve category mistakes precluding it ('idea spatially before the understanding', etc.). Failing 'accidental' semantic rescue, such a fresh conclusion will lack determinate meaning.

We have thus built up to a potentially hard-to-spot fallacy committed at the mapping-stage of analogical reasoning. Let's call it the *'metaphor-overextension fallacy'*. It consists in extending a conceptual metaphor CM (such as, e.g., Knowing-as-Seeing) by adding mappings inconsistent with CM-induced interpretations (like mappings M and N). The rules of analogical (CWSG) inference are then liable to take us from true premises to semantically deficient conclusions. Absent semantic rescue through other conceptual metaphors (or fortuitous metonymy, etc.), they will lead to such conclusions whenever CWSG inferences simultaneously employ mappings constitutive of a conceptual metaphor CM and mappings that are inconsistent with the CM-induced interpretation of a complex expression employed in the premises.

5 EXPLAING THE TRANSCENDENT MAPPINGS

But why should competent thinkers commit this fallacy? At the outset (Section 1), we took note of the basic principles of analogical reasoning, as conceived by the influential structure-

mapping theory [40, 37, 26]. We will now identify some factors due to which these principles have us make these mappings even where they lead us from truisms to nonsense.

In some cases, mapping N is straightforward. The structure-mapping account stipulates that in analogical reasoning, with or without metaphor, we routinely add new mappings, where (i) some relations have already been mapped, (ii) the requirement of parallel connectivity demands that we map their relata, and (iii) the target domain contains suitably related elements [41, 42]. This general mapping-rule leads to mapping N, in inferences from premises such as:

P_5 When we look at something, we use our eyes.

The first verb is mapped by the basic mapping of the metaphor *Thinking-about as looking-at*. The next verb, "x uses y", stands for a generic relation that obtains in both the visual SD and the intellectual TD. This relation is hence immediately mapped onto itself [27]. This leaves us looking for an element of the intellectual TD that corresponds to our eyes. The latter are introduced here as a relatum of the *use*-relation, temporally linked to the *looking-at* relation that gets mapped onto *thinking-about*. The requirement of parallel connectivity hence has us look for something we use when we think. Since we then use our wits, reason, intellect, or understanding – different labels for the same faculty – we thus obtain

Mapping N: eyes → understanding

Mutatis mutandis, the same applies to inferences employing other visual metaphors, say, from 'When we see something, we use our eyes' to 'When we understand something, we use our intellect'.

Where mappings are *ad hoc*, i.e. involved only in analogical inferences from specific premises, they are easily disregarded in different contexts where they would lead to semantically deficient conclusions. The persistence of N in inferences to such deficient conclusions as the crucial claims C_1 to C_4 therefore requires further explanation.

Parallel connectivity yields N in analogical reasoning from premises like P_5, with the core mappings of different related conceptual metaphors: 'Thinking-about as Looking-at', 'Understanding as Seeing', etc. Like many action- and event-nouns [30], all these verbs are associated with quite complex stereotypes known as 'generalised situation schemas' [16, 17]. These are made up of typical features of the action or event that the verb refers to, of the agents performing the action, and of the 'patients' on which it is performed. These features crucially include instruments typically used in performing the action [43]. The strength of stereotypical association is commonly measured through the 'cloze probability' or frequency with which the relevant concept is used to complete sentences such as:

(1) She was sewing the socks with a _____
(2) The man was arrested by _____
(3) When we look at things, we use our _____
(4) When we think about things, we use our ____

The most frequent responses are (1) 'needle' and (2) 'the police' or 'cops' [17]. And while the cloze frequencies for (3) and (4) have not yet been systematically elicited, readers will have little trouble completing them with (3) 'eyes' and (4) 'brains' or 'minds', 'wits', 'reason', 'intelligence' – early moderns would have said our 'intellect' or 'understanding'. Arguably, just as 'sewing' is associated with the subject-property 'uses a needle', 'looking at' is associated with 'uses his eyes', and 'thinking about' with 'uses his brain / mind/ reason / understanding'.

When we encounter or use a verb, all the concepts belonging to the associated generalised situation schema are activated – irrespective of contextual relevance, and the more swiftly and strongly, the stronger the association is [44]. The more strongly a concept is activated, the more likely it will be used in various cognitive processes. If the subject is engaged in analogical reasoning, the concept is hence more likely to be mapped or generated. Where an action or event designated by a source-domain verb gets mapped onto a target-domain concept, all key elements of the situation schema associated with the verb are hence likely to be mapped or generated. Where the schema associated with the TD verb contains an element that stands in the same relation (say, the instrument-relation) to the TD action as the SD associate to the SD action, the SD associate will be mapped onto the TD associate – regardless of whether that relation actually figures in the premise. Thus 'eyes' get mapped onto 'mind' or 'understanding' even in inferences from premises in which the instrument-relation does not figure, like (P_1) 'When we look at things, things are before our eyes.' Enforcing the constraint of 1-on-1 mapping in reasoning that also employs mapping M, of 'visual field' onto 'mind', then leads to the preference of 'understanding' over 'mind' we can observe in early modern texts (cp. [34]).

The case of this second mapping M, is more complex. While the patient property 'x is in the visual field of S' presumably is part of the generalised situation schemas associated with vision verbs including "S sees x" and "S looks at x", the mapping onto 'the mind' can never be obtained simply by enforcing parallel connectivity in mapping from SD to TD of a visual cognition metaphor. It cannot, because 'the mind' does not belong to the target domain of such metaphors. In talk motivated by such metaphors, "the mind" is what I propose to call a 'non-member target term'. In first approximation: While it is used only in talk about the target domain, it does not, in any sense, 'stand for' a distinct element of that domain.

To develop this notion, consider how semantic or stereotypical inferences about the SD followed by elementary analogical inferences from their conclusions (Section 2) can motivate common metaphorical expressions. Take, for instance, "S keeps X in mind", as motivated by the spatial memory metaphor unfolding from Mapping R that is the home of English mind-talk. Here, we begin with a semantic inference in the spatial SD:

(SI_1) When S keeps something x in a space (belonging to him), then X continues to be in the space belonging to S.

A mapping of this temporal relation onto the TD relation 'S continues to think of X' can be generated from the core Mapping R through ECWS inferences (what ATT-Meta calls 'vehicle neutral mapping adjuncts'). Analogical inference with this further mapping takes us from the consequent of (SI_1) to

(AI) S continues to think of X.

According to our default strategy, this would motivate a fresh metaphorical use of the SD expression "S keeps X in his space"; instead, we say "S keeps X in mind". Once the chained inference has motivated metaphorical uses of complex expressions including the words "space belonging to S", the latter get

replaced by "mind", as the new lexical item, e.g., "to keep in mind" is formed.

Mutatis mutandis, the same holds true of mind-talk motivated by visual cognition metaphors. Consider how stereotypical followed by analogical inferences could motivate metaphorical uses of complex expressions containing the expression "visual field": Typically,

(SI_2) When something is at the forefront of my visual field, I cannot help looking at it.
(SI_3) When something is at the back of my visual field, I don't look at it but am aware of it.

Analogical inference with the mapping 'Thinking-about as Looking-at' leads from the stereotypical conclusion (e.g., 'I cannot help looking at it') to a further conclusion (e.g., 'I cannot help thinking about it'). Inference chaining would motivate saying that something is 'at the forefront of my visual field' when I cannot help thinking about it, or 'at the back of my visual field' when I don't think about it, but am aware of it. ('aware of' is a generic epistemic relation that obtains in both source and target domain, hence gets mapped onto itself, and therefore can figure in ECWS inferences of the sort yielding CM-induced interpretations.) But of course we say, instead, that things are 'at the forefront' or 'back of' our 'mind'. Once the chained inference has motivated metaphorical uses of complex expressions including the words "visual field", the latter get replaced by "mind", as the new lexical item, e.g., "at the forefront of the mind" is formed.

"The mind" thus is a *non-member target term* in this more precise sense: On the one hand, it is used only in talk about the target domain, and is not used in talk or reasoning about the source domain. Within the default strategy for motivating and interpreting metaphorical talk, it is not used in reasoning about the SD but replaces source-domain words only once reasoning about the SD has motivated fresh uses of complex expressions containing those words. (In terms of the ATT-Meta approach, the term can figure in reasoning within the pretence cocoon, and its conclusions about the target domain, but not in statements about the source domain.) Hence "the mind" is a 'target term'.

On the other hand, in the cases at issue it merely replaces source-domain terms ("space", "visual field") in more complex expressions. The resulting expressions (e.g. "S keeps X in mind") can be said to refer to elements of the TD, mainly to relations between subjects and objects of thought or knowledge (e.g. 'S continues to think of X'). When the word "mind" is used as synonym of "intellect", etc. it can be said to be individually used to refer to a further TD element, namely, the faculty of reasoning thinkers may employ in thinking. When it is used in metonymies building on this use ("Two great minds [i.e. people with great cognitive abilities] debated the issue"), "the mind" is used to refer to subjects who stand in the relevant relations. But in the present cases, "the mind" merely figures in expressions that, as a whole, have target-domain meanings that are not a function of any target-domain meanings of their constituents. (Indeed, these constituents need not have any such meanings.) In these cases, the constituent expression "mind" cannot be said to refer to any distinct element of the TD: It then forms part of a complex expression that stands for a relationship between a subject and an object of thought or knowledge (e.g. 'S continues to think of X') but not for any further element distinct from such relations and their relata. Hence "the mind" is here used as a 'non-member term': It is here used in talk about the TD but not to stand for any member or element of that domain.

So why does 'the mind' get treated as a TD element, in analogical reasoning which employs mapping M alongside visual metaphors? An as yet speculative answer points out that this may be facilitated by three factors. First, "the mind" replaces words that stand for source-domain elements and whose literal meaning does influence the literal meaning of the complex expressions they enter in. It is therefore tempting to think that the complex expressions into which "the mind" enters must also have a meaning that is a function of the meaning of their constituent parts, and to look for a referent for the constituent "the mind". Since the word is used only in talk about the intellectual target domains, it is natural to look for this referent in them. And, third, the spatial memory metaphor that is its home and anchor has what we may call a *'generic source domain'*: The 'personal space' figuring in core mapping R can be instantiated by an actual physical space belonging to me, e.g., by the space enclosed by my cranium. Hence with R we can motivate saying that I 'cannot keep everything in the head' (when we cannot remember everything) or that we should try to keep certain things 'out of our head' (when we should not think of them). But the conceptual metaphor is not tied to this or any other specific physical instantiation, and the expression "the mind" is used precisely when no such specific instantiation is invoked. This may have us spontaneously rate the term as more abstract and group it with the more abstract concepts from the intellectual TD, rather than the more concrete concepts from spatial or visual source domains invoked.

Once the crucial mistake of treating 'the mind' as a TD element has been made, standard mapping principles have us map 'visual field' onto it: In a first step, SD elements get mapped onto the TD elements deemed most similar to them (Section 2). Through post-inference replacements in antecedents of inferences like (SI_1) to (SI_3), the 'mind' appears to be credited will all the abstract features (properties and relations) of delineated spaces (in which things can be kept, etc.) and, more specifically, visual fields (which have forefronts and backs, i.e. depth). Through such apparent attributions, 'visual fields' and 'minds' come to be deemed similar enough to get mapped in the first stage of mapping. The presently relevant premises P_1 to P_4 do not provide any other relata for 'x is in y', so the mapping does not fall foul of structural constraints, in the second stage.

6 CONCLUSIONS

This paper has distinguished two strategies (Section 2): In line with the ATT-Meta model, it has assumed a default strategy for motivating and interpreting (fresh) metaphorical expressions, which makes do with a very restricted form of analogical reasoning, viz., ECWS inferences from core mappings of conceptual metaphors. In line with structure-mapping accounts of analogy, it assumed a default strategy of analogical reasoning that involves a wider range of mappings and full CWSG inference. We then explored how the latter reasoning strategy can lead us from truisms about the visual SD to conclusions about the intellectual TD that cannot be understood through the former interpretation strategy. In the absence of fortunate coincidences, they lack determinate meaning; embedded in inferential links, they strike us as intelligible, even so (Section 4). These illusions of sense are due to mistakes at the mapping

stage of analogical reasoning, namely to an overextension of conceptual metaphors. We explained their extension through problematic mappings by reference to the psychology of schema activation (mapping N) and the peculiar use of "the mind" as a non-member target term (mapping M) (Section 5). The intuitions traced back to these seductive mistakes at the level of mapping are constitutive of early modern conceptions of the mind as a realm of inner perception (Section 3). We have thus obtained a debunking explanation of intuitions at the root of introspective conceptions of the mind. To the extent to which it goes beyond application of key principles of structure mapping theory, on the one hand, and ATT-Meta, on the other, it remains to be computationally developed and experimentally tested.[2]

REFERENCES

[1] R.E. Nisbett and T.D. Wilson. Telling more than we can know: verbal reports on mental processes, *Psychological Review,* 84: 231-59 (1977).

[2] J.A. Bargh, M. Chen, and L. Burrows. Automaticity of social behaviour: Direct effects of trait constructs and stereotype activation on action, *Journal of Personality and Social Psychology,* 71: 230-244 (1996).

[3] A. Tversky and D. Kahneman. Judgment under uncertainty: heuristics and biases. *Science,* 185: 1124-1131 (1974).

[4] P. Slovic, M. Finucane, E: Peters, and D: MacGregor. The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (eds.), *Intuitive Judgement: Heuristics and Biases*. Cambridge: CUP (2002).

[5] M.S. Gazzaniga. *The Social Brain*. New York: Basic Books (1985).

[6] D.M. Wegner. *The Illusion of Conscious Will*. Cambridge, Mass.: MIT Press (2002).

[7] T.D. Wilson. *Strangers to Ourselves. Discovering the adaptive unconscious*, Cambridge, Mass.: Harvard UP (2002).

[8] J. Knobe, J. and S. Nichols. An experimental philosophy manifesto. In their (eds.), *Experimental Philosophy* (pp. 3-14). Oxford: OUP (2008).

[9] J. Pust. Intuition. In E.N. Zalta (ed.), The Stanford Encyclopedia of Philosophy (Winter 2012 Edition). URL = http://plato.stanford.edu/archives/win2012/entries/intuition/ (2012).

[10] E. Fischer. Philosophical intuitions, heuristics, and metaphors. *Synthese,* 191: 569-606 (2014).

[11] G. Lakoff and M. Johnson. *Philosophy in the Flesh*. New York: Basic Books (1999).

[12] A. Tversky and D. Kahneman. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, 90: 293-315 (1983).

[13] D: Kahneman and S. Frederick. A model of heuristic judgment. In K.J. Holyoak and R. Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267-293). Cambridge: CUP (2005).

[14] D. Kahneman. *Thinking fast and slow*. London: Allen Lane (2011).

[15] K.J. Holyoak. Analogy and relational reasoning. In K.J. Holyoak and R.G. Morrison (eds.), *Oxford Handbook of Thinking and Reasoning* (234-59). New York: OUP (2012).

[16] T.R. Ferretti, K. McRae, and A. Hatherell. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language,* 44: 516-547 (2001).

[17] K. McRae, T.R. Ferretti, and I. Amyote. Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, 12: 137-176 (1997).

[18] J.A. Barnden. Metaphor and context: A perspective from artificial intelligence. In A. Musolff & J. Zinken (eds.), *Metaphor and Discourse* (pp.79-94). Palgrave Macmillan (2009).

[19] J.A. Barnden. Creative metaphor and metaphorical creativity. In T.R. Besold, M. Schorlemmer & A. Smaill (eds.): *Computational Creativity Research: Towards Creative Machines* (pp. 217-242). New York: Springer (2015).

[20] R.W. Gibbs. Evaluating conceptual metaphor theory. *Discourse Processes, 48*: 529-562 (2011).

[21] S.B. Day and D. Gentner. Non-intentional analogical inference in text-comprehension. *Memory and Cognition,* 35: 39-49 (2007).

[22] L.A. Keefer, M.J. Landau, D. Sullivan and Z.K. Rothschild. Embodied metaphor and abstract problem solving: testing a metaphoric fit hypothesis in the health domain. *Journal of Experimental Social Psychology*, 53: 12-20 (2014).

[23] P.H. Thibodeau and L. Boroditsky. Metaphors we think with: the role of metaphor in reasoning. *PLoS One 6(2)*, e16782, doi:10.1371/journal.pone.0016782 (2011).

[24] P.H. Thibodeau, and L. Boroditsky. Natural language metaphors covertly influence reasoning. *PLoS One 8(1):* e52961. doi:10.1371/journal.pone.0052961 (2013).

[25] G.J. Steen, W.G. Reijnierse, C. Burgers. When do natural language metaphors influence reasoning? A follow-Up study to Thibodeau and Boroditsky (2013). *PLoS One 9(12):* e113536. doi:10.1371/journal.pone.0113536 (2014).

[26] B. Falkenhainer, K. D. Forbus, and D. Gentner: The structure-mapping engine: Algorithm and examples. *Artificial Intelligence,* 41: 1–63 (1989).

[27] K.D. Forbus, D. Gentner, and K. Law. MAC/FAC: a model of similarity-based retrieval. *Cognitive Science, 19*: 141-205 (1995).

[28] P. Wolff and D. Gentner. Structure-mapping in metaphor comprehension. *Cognitive Science*, 35: 1456–1488 (2011).

[29] A.M. Wallington. Systematicity in metaphor and the use of invariant mappings. In: G. Low et al. (eds.), *Researching and Applying Metaphor in the Real World* (pp. 209-244). Amsterdam: John Benjamins (2010).

[30] M. Hare, M. Jones, C. Thomson, S. Kelly, and K. McRae. Activating event knowledge. *Cognition,* 111: 151-67 (2009).

[31] S.C. Levinson. *Presumptive Meanings. The Theory of Generalized Conversational Implicature*. Cambridge, Mass.: MIT Press (2000).

[32] E.C. Traugott. The rise of epistemic meanings in English: example of subjectification in semantic change. *Language,* 65: 31-55 (1989).

[33] E.C. Traugott and R.B. Dasher. *Regularity in Semantic Change*, Cambridge: CUP (2005).

[34] E. Fischer. *Philosophical Delusion and its Therapy*. New York: Routledge (2011).

[35] J. Locke. *An Essay Concerning Human Understanding*, 4th ed. Ed. P. Nidditch. Oxford: Clarendon Press (1700/1975).

[36] E. Fischer. Mind the metaphor! A systematic fallacy in analogical reasoning. *Analysis,* 75: 67-77 (2015).

[37] D. Gentner, M. Ratterman, and K. Forbus. The roles of similarity in transfer: separating retrievability from inferential soundness. *Cognitive Psychology,* 25: 527-75 (1993).

[38] M.E. Lassaline. Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*: 754-70 (1996).

[39] Bowdle, B. and Gentner, D. The career of metaphor. *Psychological Review,* 112: 193-216 (2005).

[40] D. Gentner. Structure mapping: a theoretical framework for analogy. *Cognitive Science,* 7: 155-70 (1983).

[41] D. Gentner and A. Markman. Structure mapping in analogy and similarity. *American Psychologist,* 52: 45-56 (1997).

[42] D. Gentner and A. Markman. Defining structural similarity. *Journal of Cognitive Science,* 6: 1-20 (2005).

[43] M. Harmon-Vukić, S. Guéraud, K.A. Lassonde, and E.J. O'Brien. The activation and instantiation of instrumental inferences. *Discourse Processes,* 46: 467-90 (2009).

[44] R. Giora. *On our Mind. Salience, context, and figurative meaning*. Oxford: OUP (2003).

[2] For helpful comments on a previous draft, I am indebted to John Barnden and an anonymous referee.

Metaphors in Theory of Information. Why They Capture Our Concepts and Undertakings

Marek Hetmański

Abstract. Metaphors are common in various types of discourse; even natural sciences are engaged with the figurative way of expression mostly characteristic of the humanities. They are also suited, to an astonishing extent, to the exact, strict and formal theories of information, as has been presented in the first part of the paper, on the example of the Shannon & Weaver's Mathematical Theory of Communication. The metaphoric entanglement of the information category shows that its commonsensical and figurative conceptualization is unavoidable. Nevertheless, it also opens certain crucial questions concerning the ways of conceptualizing the probable, uncertain events which happen in the course of communication and deciding.

1 COGNITIVE TOOL

Metaphors are both linguistic and rhetoric means for making analogies between different domains of things. They facilitate the understanding of a complex, obscure, or unfamiliar domain of things, processes, and events through reference to another – one that is more concrete, familiar and comprehensible. Metaphors traditionally function as verbal expressions and utterances of particularly suggestive and pervasive power. They mainly operate as linguistic tools useful in conceiving and describing the world not only in literature but also in science, where they have been manifest and useful throughout the history of science.

But metaphors are not merely verbal in their nature, they are not limited to engaging only the linguistic or communicative competences and faculties of their users. They express deep and complex human mental states and ways of thinking, which are the crucial backdrop for these figurative expressions. Specifically, the nature of metaphors is *conceptual* rather than exclusively verbal - as it is commonly but misleadingly conceived and as is widely investigated and advocated in the theories of *cognitive metaphor* (see [1, 4, 5, 6, 7]). By comparing two different things, processes or events (the subject domains – source and target) with regard to one important aspect, i.e. saying that X *is* (is *like*) Y, metaphor helps to perceive, imagine, and understand one thing (target) in terms of another (source). Although it is expressed in an expressive, concise way, it is in fact a product of *image schemas* (*conceptual* frameworks) underlying said verbal expression. The frameworks which constitute the agent's mind are sensory-motor in their nature, encompassing such abstract and universal elements as: (1) time and space correlations, (2) before-after things sequences, (3) top-down and/or bottom-up directions, (4) horizontal and/or vertical orientations, as well as the agent's (5) behavioral patterns of movement, manipulation and control. These frameworks organize the agent's experience, be it of his/her immediate environment or the furthest expanses of the universe. Notably, image schemas are especially helpful in trying to envisage the possible, probable or entirely random situations, when planning and predicting the agent's future activities becomes crucial. This has important consequences both in terms of mental and practical aspects of metaphoric discourse. As metaphors shape and guide the agent's behavior in specific directions, they not only explain (as one can obviously expect) that which is is metaphorically expressed, but also unexpectedly hide or obscure is the actual content of the metaphoric thinking. "[A] metaphorical concept can keep us from focusing on other aspects of the concept that are inconsistent with that metaphor" [2]. These somewhat paradoxical consequences will be more closely examined when we consider the metaphoric nature of probable states (Section 3).

[1] Marie Curie-Sklodowska University, Lublin, Poland, email: marek.hetmanski@poczta.umcs.lublin.pl

2 METAPHORS OF INFORMATION

Claude Shannon and Warren Weaver's Mathematical Theory of Communication [8] is an example of metaphoric thinking engaged in the abstract domain of communication. The linguistic aspect merely implicitly accompanies that which is explicitly (formally, quantitatively) stated. The authors admit using the word *communication* in "a very broad sense to include all of the procedures by which one mind affects another", or as they later specify, "in fact all human behavior (…) one which would include the procedures by means of which one mechanism (…) affects another mechanism" [8]. It is a very broad and general depiction of communication. The examples of communicational mechanisms include not only oral and written speech but also music, theater, pictorial art,

television, and ballet as well as a guided missile weapons system; all of the above employ procedures of sending, transmitting and processing signals that change the states of the communication processes. However "the language of this memorandum," as Shannon and Weaver relate to their paper, "will often appear to refer to the special, but still broad and important, field of the communication of speech" [8] whereby it aspires to account for all of the above examples of communication. The authors' intention has had certain consequences affecting both their own and other researchers' understanding of information.

The subject of communication as such is considered at three levels: (1) technical – consisting in matching specific signals and symbols while transmitting them during the communication process; (2) semantic – consisting in finding "how precisely do the transmitted symbols convey the desired meaning?"; and (3) pragmatic – "how effectively does the received meaning affect conduct in the desired way?". The last two levels are crucial in that they concern *changes* which communication may bring about, namely, "the success with which the meaning conveyed to the receiver leads to the desired conduct on his part". The essence of communication, including the transfer of signals, lies in changes experienced by the agents involved; information is a function of these changes. The meaning of those signals, analyzed at the semantic and pragmatic levels, is not their main characteristics. It is a relative feature of the transmitted signals and depends on particular sender/receiver intentions. But it is only in ordinary thinking that meaning is identified with information and a particular message having a content. From the point of view of mathematical theory of communication the above statement is misleading. "In particular," say the authors, "*information* must not be confused with meaning" [8]. Two messages – one of which is meaningful and the other completely nonsense – can be formally equivalent and regarded as carrying the same amount of information, no matter the things and situations they refer to.

By adopting the cognitive theory of metaphor as the theoretical background, it is possible to identify in the authors' paper certain crucial elements constituting the structure of each conceptual metaphor. Firstly, there is the *target domain* consisting of the following elements: (1) probable states of events which constitute signals (called "source"); (2) an abstract place/space where signals are transmitted ("channel"); (3) random disturbances of signals as well as interferences between the same and other elements of the channel ("noise"); (4) a way in which signals are organized into a message ("code"); (5) an effective (despite entropy) way of transmitting signals ("redundancy"); and finally, (6) transmission of signals with minimal dispersion to prevent loss of information.

To explain what the above abstract elements (characteristics of any communication) are, Shannon and Weaver provide many *analogies* with empirical and concrete phenomena and situations derived from instances of human communication. They compose a "story" explaining in detail what the subject matter of their paper/report is. In doing so, they constitute a *source domain* consisting of the following, plainly described, consequtive elements: (1) physical signals constituting the message (the news); (2) voice, writing, signals of the nervous system, all of which are constituents of the medium in which transmission takes place; (3) audible sounds or visible seen (e.g. in analog telephony or television) which disturb the process of communication; (4) language and alphabetic coding ; (5) linguistic and literary styles which help to organize a system of signs into a message; and finally, (6) the actual act of communication. By using self-explanatory and simple analogies to everyday events and situations, the authors try to grasp the essence of information. They do it in a metaphorical – indirect rather than strict or formal – way, which helps them to home in on the general nature of information. But metaphoric understanding of information is neither exclusive nor even dominant over the formal conception of the same. It takes place, so to speak, spontaneously, in accord with ordinary language rules; it shapes the theory in a specific way, leaving on it a remarkable mark. Summarizing their theory of information/communication, Shannon and Weaver write in a tellingly metaphoric way: "An engineering communication theory is just like a very proper an discreet girl accepting your telegram. She pays no attention to the meaning whether it be sad, or joyous, or embarrassing. But she must be prepared to deal with all that come to her desk" [8]. They suggest, in other words, that their conception of information has universal meaning what they express nevertheless through the metaphoric words. Presenting information in this phrase as merely a physical thing (telegram coming to desk) by analogy to the *meaning* of message which is always a concrete thing (which they recommend rather to separate from information as such), they unintentionally but inevitably deprive it of its abstract sense, which depends on *probabilistic* nature of information. In that way mathematical theory of communication due to its metaphoric confinement has been involved in methodological situation. The empirical and vivid elements from the source domain affected, if not dominated, characteristics of the target domain

The mentioned metaphorical aspect of the information theory, generally speaking, stems from the model of a communication act in which the speaker puts ideas (as objects) into words (as containers) and sends them (along a conduit, in a channel) to the listener, who then takes the idea/object out of the words/containers, performing all these activities automatically and without difficulty. This simplified model – to which Michael Reddy refers to as the "conduit metaphor" [6] – is very suggestive and effective in explaining both interpersonal and mass communication. We come across its realizations in ordinary thinking as well as in different conceptions and theories attempting to define communication as such. Mathematical theory of communication is partially tailored to the idea which it reciprocally reinforces. The conduit metaphor generally suggests that communication is reasonable, almost effortless, and does not bring about any interpretational problems. But Reddy argues that this reduced and simplified model fails to represent the actual complexity and richness of human communication; it is presumed that only simple examples of transmissions in mass communication can be reduced to the same. Human communication depends on changing the interlocutors' states of mind but not transmitting the thoughts alongside ideal channel. It occurs and takes place in human minds and acts, rather than in language alone. As it is never perfect, aberrations and disturbances are unavoidable, they are not obstacles but rather circumstances of its development and progress. "They are tendencies inherent in the system, which can only be counteracted by continuous effort and by large amounts of verbal interaction" [6]. The real and rich (informative)

model of such communication must consider dynamic changes rather than static and one-way mechanisms.

3 CHOICE OF PROBABLE STATES

How does the metaphoric confinement of information change our understanding of this category? To what extent does it reveal, or obscure, the essence of the same? Shannon and Weaver seem to be aware of all of these problems and consequences, however, they are not overly focused on the figurative aspect of their discourse. Their main proposition is a purely objective, not subjective (i.e. not agent-oriented), conception of communication and communication. Their basic thesis holds that information is *selection* and *choice* made among the probable states caused and demanded by communication. Transmission of signals involves selecting from a set of alternative states at the source and announcing it at the destination. It concerns not so much what really happens (the fact) as what would happen (possibility) during communicating. "[T]his word information in communication theory relates not so much to what you *do* say, as what you *could* say. That is, information is a measure of one's freedom of choice when one selects a message" [8]. During the process of communication, no messages are simply sent, instead signals are chosen, transmitted and selected. Communicating *per se* is altering both the initial and final states of this process, the result of which yields information. It is therefore in line neither with the common (intuitive) understanding of communication, nor with the model of information as the message. The natural conceptual schemas – linear, before-after sequences of things, as well as time-after sequences of events – underlying the mathematical theory of communication are used by their authors unconsciously. The metaphorical effect is caused without any prior intention.

As they mention that "the unit information indicating that in this situation [i.e. transmitting the signals] one has an amount of freedom of choice, in selecting a message" [6], Shannon and Weaver concentrate on the formal nature of the key concept. Grasping its complex, partially counter-intuitive nature demands a specific cognitive ability. They hold that the abstract "amount of freedom of choice" appeals to any type of communication when the agent's choice – no matter who or what it is, a human being or a machine – results in receiving information. To be more specific and understandable, they turn to figurative modes of expression, which ultimately makes the quantitative problem rather complicated, open to metaphoric discourse. Mathematical (probable) interpretation of information conceives it as an act of choice between possibilities with which the agent is confronted. The agent should distinguish among all probable things, events and processes and then act effectively by selecting one of the same. There is no information without choice, if the agent had no choice at all, information would not appear. Selection and choice among the possible states result increased *uncertainty,* which formally characterizes this situation. "Information is, we must steadily remember, a measure of one's freedom of choice, and hence the greater the information, the greater is the uncertainty that the message actually selected is some particular one. Thus greater freedom of choice, greater uncertainty, greater information go hand in hand" [8]. The authors explain that in order not to fall into "the semantic trap" (when one should remember that the word "information" is used in a special, narrowed meaning), one ought to conceive information as the concept which "measures freedom of choice and hence uncertainty as to what choice has been made".

4 DECISION MAKING

Coping with the probable states of things and situations is a complex task, both cognitively and practically. It demands proper, prior comprehension of what is probability as such and then a subsequent realization of some general intuitions as well as elementary rules. The ambiguous, somehow counter-intuitive (qualitative) and at the same time exact and strict (quantitative) nature of the *concept* of probability is a challenging issue of science and common experience alike. Its scientific and commonsensical meanings are different in some regards and convergent in others. They are all in principle connected with an act of *making decisions* – a situation in which the agent pursues one direction and steers clear of others on the basis of signals/information he or she receives. For this reason, decision making is a communicational act with an informational aspect; on the other hand, any communication is at the same time intrinsically burdened with choice and decision making.

The decision-making mechanism engaged in communication is commonly compared to tossing up (flipping a coin) or betting on randomized games. This evident metaphorical aspect of conceiving what making a choice/decision when faced with a number of probable states is, brings about certain serious interpretational difficulties. Namely, it demands selecting and choosing the proper picture or model from among all the available alternatives (each with its own metaphorical power) of such a situation. And then the chosen model moulds the comprehension of the nature of probability. In such a situation people perceive and define all types of decision making as concrete games such as dice, roulette wheels or other gambling devices, and also in receiving the news – unexpected and astonishing. Empirical examples derived from everyday life dominate people's imagination and understanding of the choices they are obliged to make. At such times, the probability of scientifically-investigated events (e.g. statistics) is important and decisive.

But the very concept of probability has, in principle, two different meanings – statistical (formal, quantitative) and epistemological (psychological, qualitative) – both of which are constantly misread and used interchangeably thus leading to many problems. "Statistical probability was the sole legitimate form of probability, the sole basis for knowledge. Consequently, »statistical probability« − and the associated world of »randomizing devices« − has become a metaphor for epistemological probability" [5]. The mathematical concept is what gave the idea of probability its content and epistemic aspect. Conversely, *epistemological* probability, secondary and derivative to the statistical one, is the result of preferred theoretical interpretation rather than correlations between actual events. In this sense, the formal (mathematical/statistical) aspects serves as the basis for presenting the target – the agent's imagination of probability as well as his/her experience of uncertainty (mental states). In other words, the abstract serves as a metaphor for the concrete.

Regardless of these ambiguities and reciprocal relations (recognizable at the theoretical level), people commonly conceive, and subsequently cope with, probability as a state of their own *beliefs* rather than events or affairs. It so

happens that statistical probability becomes a definition – a convincing metaphor of people's thoughts and actions – affecting the experience of the world and any knowledge one might claim to have about it. Such a metaphor serves the descriptive function of supplying explanation for unstable, unpredictable, unfamiliar cognitive phenomena such as making choices, predictions or decisions under conditions of uncertainty. Besides, to a certain extent, it also plays a rhetoric function of encouraging people to perform particular socio-cognitive acts with the expectation of securing some profits, especially in the context of randomized events and situations. But in either case metaphoric thinking obscures that which it actually aims to reveal and explain. That is why Raymond W. Gibbs recognizes a specific "»paradox of metaphor« in which metaphor is creative, novel, culturally sensitive, and allows us to transcend the mundane while also being rooted in pervasive patterns of bodily experience common to all people" [1]. It is not particularly rare for this simple figurative manner of thinking to change ways in which more complex phenomena such as the probability of events are conceived.

This seemingly contradictory nature of metaphoric thinking would mean that people engaged in the same are really unable to exceed their physically, mentally and culturally entrenched limits, their conceptual schemas. In transcending what is empirically evident (source domain) and consequently entering cognitively into new, more complex intellectual domains (target), agents are confronted with many empirical constrains – gestures, mental and linguistic schemas, and/or social customs and values. They conceptualize complex and abstract phenomena by means of material, practical devices and instruments, which is especially evident in the context of probability. This specific conceptual-instrumental equipment is of particular use when coping with randomness.

Empirical studies on the mentioned problems of probability and information [2] have lead to interesting conclusions which shed some light on the metaphoric confinement of communication and information. Gerd Gigerenzer holds that all types of decision-making, ranging from simple and intuitive to more complex and rational, are based on limited information. It means they all such choices are far from rational where agents would be equipped with complete and reliable knowledge. Indeed, situations of complete information – where an agent would be able to compute all available courses of action and thus make a fully informed choice – are unattainable. Considering possibilities and selecting probabilities is not algorithmic but mostly heuristic. People tend to make *correct choices* (when buying, investing or communicating) more easily and more often when they are faced with relatively few alternatives, otherwise they would be overwhelmed with the extent of analysis necessary during decision-making. This is a strategy which relies on gut feelings, the so called rule of thumb, in other words intuition. "The quality of intuition lies in the intelligence of the unconscious, the ability to know without thinking which rule to rely on in which situation" [2]. Intuition might give the agent a chance to use more discretional ways of expression, which he/she conceives as similar as well as more (or less) probable. In this way metaphoric thinking combines with intuition and helps us to understand complex situations.

The same correlation has been observed and empirically studied by Daniel Kahneman and Amos Tversky [9] in their theory of making decision under uncertainty. They hold that while making a decision or solving practical and cognitive problems, the agent utilizes relatively constant *cognitive biases* which reflect his/her specific, unavoidable cognitive faults and errors. They include intuitive judgements and beliefs which play a particular role in the assessment of random events and their probability. "[P]eople rely on a limited number of heuristic principles which reduce the complex tasks of assessing the probabilities and predicting values to simpler judgmental operations" [9]. In particular, biases such as: (1) not properly identifying representativeness in a sequence of events, (2) excessive ease in evaluating such sequences, and (3) incorrectly settling statistical problems based on an erroneous evaluation of input data, are decisive for the agent's cognitive faculties. There are also others that result from the agent's cognitive inability to conceive probabilities of events. Namely, the agent assumes erroneous representativeness relative to the transfer of qualities or probability from one class of events to another. It is due to his/her incessant *search for similarities* between facts and events, despite their evident dissimilarity. In conditions of such cognitively biased thinking, the agent becomes especially susceptive to any suggestive expressions that strengthen this tendency, which is when the role of metaphors becomes particularly crucial.

5 PRACTICAL CONSEQUENCES

It is worth mentioning that the problem of developing proper metaphoric concepts and models of information and communication, apart from the strictly methodological aspects of the same, has certain practical consequences. Shannon and Weaver did not consider these consequences to be relevant to only the explanatory aspect of metaphoric phrases they have themselves used on occasion. But if the conduit metaphor, implied in their conception, might confuse people, be it experts, theorists and laymen conceiving what information is and how it is communicated, the issue of the metaphoric confinement of the very concept of information acquires significance. It may influence the way people communicate and decisions while selecting and processing signals and information. Indeed, it may induce or even compel them to make wrong choices while sending and receiving various types of messages such as orders**,** inquiries, requests, the news, pictures, texts etc. Such instances occur in the context of education, public affairs, political domains or mass culture, wherein communication is fundamental. In these sociocultural domains – in their institutions and organizations such as schools, colleges, universities, libraries, cultural, scientific and research centers – metaphoric phrases, definitions and conceptions of information and knowledge are of particular importance. Only metaphors possessing dynamic and probabilistic, rather than static or linear connotations in their source domain can describe processes of knowledge acquisition and communication whose quantitative aspect is information. By appealing to astonishing phenomena, they can adequately anticipate new and unforeseen informational processes and events; their rhetorical impact would thus change the previous, conservative conceptualization. Only such enriched figurative thinking is able to evoke human *creativity* in cognitive, intellectual, social and cultural areas.

The cognitive, or more precisely descriptive role of informational metaphors is largely realised within the

discipline of information and knowledge organisation, which commonly employs the definition of information formulated by the mathematical theory of communication. As was already discussed in [3], metaphors pertaining to various data bases utilised by libraries, offices or governmental bodies, as well as any open (internet) repositories of knowledge, play a significant role in defining ways in which these can be organised and used. Rather than merely describe and model, they also provide opportunities for creation and administration, as well as, most importantly, effective utilisation of the same by various users. Many of the existing metaphors of knowledge organisation employ metaphorical descriptions, many dating back as far as antiquity or the middle ages, which compare accumulated and available knowledge to buildings (towers, libraries), labyrinths, vast open spaces (on land or sea), trees, maps, networks, or rootstalks. Each of the above emphasises the physical and spatial (geometric, linear and finite) characterisation of knowledge which is typically depicted as a complete and perfect source of information. Consequently, any attempts to acquire knowledge, expand it, discover new content, or establish new connections, will be described using metaphors such as juggling, wandering, exploring, leafing through, deciphering, enquiring, responding, etc. Such metaphors will normally emphasise a rather passive and unproblematic use of information gathered in static and invariable deposits and data bases. If such metaphors are to serve the function of directives or recommendations, rather than merely descriptions or models, they are likely to be addressed to persons involved in the creation and management of such resources, and not so much to regular users of knowledge systems. The latter have in recent decades been approached with ever more plentiful metaphorical expressions pertaining to the internet, which predominantly carry either clearly positive or negative cognitive and emotional connotations and relate to repositories of information and processes of researching for the same. If inclined positively, such metaphors employ phrases whose source domain includes such positively charged expressions as surfing, exploring, richness, surprise, enrichment, etc. Otherwise, information and the internet may likely be metaphorically described as junk, smog, excess, boundlessness, impoverishment, threat, etc.

All informational metaphors (regardless of their axiological associations) become significant only if used in such a way that, aside from their obvious function of describing (modelling) the existing knowledge and information resources, they also encourage their addressees to engage in a particular course of cognitive action. Shannon's conclusion that the gist of information refers not to what is, but to what *can be communicated*, constitutes an important methodological directive in constructing metaphorical references to knowledge and information. Namely, they should refer to the cognitive expectations of particular agents and the realistic possibility of their fulfilment, rather than merely ready-made realisations and factuality. To accomplish this, however, it is necessary to have a criterion allowing for a distinction between: (1) real (realised, own) cognitive *needs* of internet users and (2) apparent (imposed, unrealised) cognitive *demands* encountered when using software tools and applications. Such metaphorical expressions – suggestive but free of obtrusive marketing and advertising tricks – should take the form of directives and guidelines, commands and, most importantly, warnings addressed to internet users.

Any metaphors but particularly those functioning as suggestive linguistic expressions have (as dictated by their rhetorical and eristic origin) a considerably persuasive force which is manifested through inspiring specific behaviours. If an informational and communicational metaphor comprises in its source domain expressions and phrases relating to the expected, possible, and likely, rather than exclusively actual and unambiguous cognitive situations, it will be successful in performing its persuasive function. It can then become an instrument shaping the attitudes of the cognitively wealthy rather than just the informatively impoverished. Moreover, a properly structured metaphor of knowledge organisation will facilitate internet users in making decisions and tackling cognitive problems, wherein access to suitable information is the necessary condition of success. By indicating possibilities and likelihoods – hidden behind apparent information, unavailable to software users overly preoccupied with the operation of these instruments – such a metaphor may reveal the full informative value of a cognitive situation and allow its due recognition.

6 CONCLUSION

It has been shown that metaphoric phrases used by Claude Shannon and Warren Weaver in their Mathematical Theory of Communication are only complementary, and not main in describing what information is. Owing to the theory of conceptual metaphor, one can recognize the implicit mental structures underlying such way of conceptualizing. It has also been suggested that informational metaphors might constitute useful instruments in coping with probable states while making decisions.

REFERENCES

[1] R.W. Gibbs, (Ed.), *The Cambridge Handbook of Metaphor and Thought*, Cambridge University Press, Cambridge, 2008.

[2] G. Gigerenzer, *Gut Feelings: The Intelligence of Unconscious*, Penguin Group, London, 2007.

[3] M. Hetmański, "The actual role of metaphors in knowledge organization", in W. Babik (Ed.), *Knowledge Organization in the 21st Century*, Ergon Verlag, vol. 14, 73-79, 2014.

[4] J. Lakoff, and M. Johnson, *Metaphors We Live By*, The University of Chicago Press, Chicago-London, 1980.

[5] A. Ortony, A. (Ed.), *Metaphor and Thought* Cambridge University Press, Cambridge, 2nd ed., 1993.

[6] M. Reddy, "The conduit metaphor: A case of frame conflict in our language about language", in A. Ortony, (Ed.), *Metaphor and Thought*, Cambridge University Press, Cambridge, 164-201, 1993.

[7] D. Ritchie, "Statistical Probability as a Metaphor for Epistemological Probability", *Metaphor and Symbol*, Vol.18, No. 1, 1-12, 2003.

[8] C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication* The University of Illinois Press, Urbana, 1948/1964.

[9] A. Tversky and D. Kahneman, "Judgment under uncertainty", in A. Tversky and D. Kahneman, (Ed.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, 3-28, 1974/1998.

From Metaphor to Hypertext: an Interplay of Organic and Mechanical Metaphorics in the Context of New Media Discovering

Zuzana Kobíková[1], **Jakub Mácha**[1]

Abstract. Hypertextual linking of information is one of the basic principles of digital media. We suppose this principle to be discovered in metaphorical thinking with the help of the so-called absolute metaphors. We derive the notion of an absolute metaphor from Hans Blumenberg's metaphorology, and we interpret metaphors according to Max Black's interaction theory. Our aim is to interpret these absolute metaphors as being open to new implications, just as they are open to a pragmatically determined dialectical interaction of organic and mechanical metaphorics. We follow the direction of interactions within these metaphorics in a philosophical attempt to explain the nature of mechanical and organic systems. In particular we will analyse the metaphors 'association is trail' (Bush), 'computer is a clerk' (Engelbart) and 'hypertext is a Xanadu' (Nelson). All these metaphors are both organic and mechanical. That is why we can say that hypertext is both an organic and mechanical system.

'It is reality that awakens possibilities, and nothing would be more perverse than to deny it. Even so, it will always be the same possibilities, in sum or on the average, that go on repeating themselves until a man comes along who does not value the, actuality above the idea. It is he who first gives the new possibilities their meaning: their direction, and he awakens them. But such a man is far from being a simple proposition. Since his ideas, to the extent that they are not idle fantasies, are nothing but realities as yet unborn, he, too, naturally has a sense of reality; but it is a sense of possible reality, and arrives at its goal much more slowly than most people's sense of their real possibilities.'

Robert Musil, *The Man Without Qualities*, ([1], p. 12)

1 INTRODUCTION

It has been convincingly argued (e.g., [2,3,4]) that a metaphor appears often at an outset of scientific discoveries. We can illustrate this statement in the case of the discovery of hypertext. As a nonlinear text with links containing references to other pieces of information, hypertext presents a new form of media, formed through the remediation of a prior, analogue medium of a text. We present how is the outset of this discovery articulated in a figurative way of metaphor and model.

Etymologically speaking, metaphor means a transfer. According to Arendt [5], we need to use a metaphor, when we need to transcend the borders of the real (given) world and then lead into speculation, (in our case speculation about the as yet non-existing hypertext, which we have no words for yet). A metaphor means, in this sense, a transfer from something imagined into something existing, thus into a material, functional medium. This is possible with the help of so called predicative metaphors based on analogies. We thus interpret the process of the invention of hypertext as a metaphor in the sense of transfer, which bridges the gap between a possible and an existing reality, as suggested in our epigraph from Musil.

We want to show why that figurative thinking is constructive and worthwhile in the discovery of hypertext and its explanation, and for which roles metaphors and models play in the scientific conceptualising of hypertext.

We will suggest that all inventors of hypertext concepts, mentioned in this paper, make up their concepts of hypertext by employing so-called absolute metaphors. This term, from Blumenberg ([6], pp. 62–69), means a background metaphorical complex, or some leading idea, that systematically informs the thinking of individuals and entire epochs by reference to an implicit model, such as a mechanism or an organism.

We will present the concepts of hypertext as systems based on metaphors, which connect organic and mechanical metaphors together. We will show this in detail with the help of the analysis of the following predicative metaphors, which we suppose to be absolute:

- 'association is trail' (Vannevar Bush, 1945) [7],
- 'computer is a clerk' (Douglas Carl Engelbart, 1962) [8],
- 'hypertext is a Xanadu' (Theodor Holm Nelson 1974) [9].

We want to show that above mentioned metaphors of hypertext are not mutually independent. They have evolved from the first one to the third one, as we will show below.

Vannevar Bush (1890–1974) is our first prototype of Musil's man with a sense for a possible reality. Bush started the transfer between a real (unsatisfactory) and a possible (better) feature of a new form of text. We will interpret his memex as a theoretical model developed from the 'association is a trail' metaphor. Bush wanted to improve the way how scientists deal with information. His memex (imagined as a mechanical machine) would archive all the data that a scientist has collected. The memex would link all this information by means of metadata indexing. Bush 'mobilized' his idea by the means of scientific communication. He described himself as a man of the mechanical age [10] and who wanted to address the scientists of the digital age and to encourage them to transfer his theoretical model onto a functional medium.

The Engelbart and Nelson concepts of hypertext are built upon Bush's metaphor.

[1] Dept. of Philosophy, Masaryk University, Arna Nováka 1, 602 00 Brno, Czech Republic. Email: {kobikova, macha}@mail.muni.cz.

We can start with a presentation of this organic and mechanical interplay of metaphorics from a methodological and historical viewpoint.

2 METHODOLOGY

We draw on Hans Blumenberg's metaphorology [6], combined with Max Black's interaction theory of metaphor [11], and his view of how metaphors and models are used in scientific discoveries [2]. Why have we chosen these authors? Black's sematic theory has now been superseded by pragmatic accounts, which rightly accentuate the pragmatic dimension of the metaphor. But we do not need to focus on the actual speech situation. In this paper we focus on metaphors and their implications mentioned in the scientific texts about hypertext.

Blumenberg's metaphorology resembles Lakoff's and Johnson's theory of the conceptual metaphor [12] which has received much more attention in the past few years. But Blumenberg's account is arguably more complex in its historical point of view, which is also our main focus.

Metaphorology is not just another theory of metaphor in our modern sense, i.e. an analysis of the concept of metaphor, but it is an investigation into some prominent instances of this concept. The first aim of metaphorology is to substantiate the existence of the so-called *absolute metaphor* which, hypothetically for the time being, can be considered as a foundational element of philosophical language. According to Blumenberg, absolute metaphors cannot be translated into unambiguous literal language,[2] they are, so to speak, 'resistant' ([6], pp. 3–5). Blumenberg does not however explain why this or that metaphor is absolute. In his historical perspective, a metaphor is absolute if it has resisted being fully translated thus far. This does not exclude the fact that such a metaphor could be fully translated in the future. We suppose that, in a nutshell, a metaphor is absolute (for a given period), if every attempt at its explanation results in another metaphor or analogy.

The fact that an absolute metaphor cannot be translated into literal language – and this is the second step in Blumenberg's project – does not prevent it from replacing or correcting another absolute metaphor. Such transformations take place in history and they are important subjects of metaphorology ([6], p. 3).

For instance, there are a lot of metaphors about the world: 'the world (order) is (like) a machine' (*machina mundi*) or 'the world is clockwork'.

These two metaphors are not mutually independent, as the latter is a certain specification of the former. In numerous quotations from philosophers and scientists, Blumenberg tried to show how the *machina mundi* metaphor has been transformed into the clockwork-metaphor with the dawning of the Enlightenment ([6], pp. 62–69).

In this paper we will focus on two particular metaphors or rather metaphorical themes (which we call 'metaphorics') – on mechanical and organic metaphors, their dialectical interplay and blending when explaining the nature of associative memory, text and hypertext. In order to do so, we follow Blumenberg's need to examine the *consequences* of this or that particular metaphor by various thinkers. A set of non-contradictory consequences of a metaphor is what we call, following Black's interactions theory [11], its *interpretation*.

Max Black provides a complex method of interpreting vital, predicative metaphors of the form 'A is B'. The basic idea is that if such an utterance is intended or/and recognized as a metaphor then the literal meaning of 'A' *interacts* with the literal meaning of 'B' resulting into a *metaphorical* meaning 'B' which is hereby being predicated of 'A' The core of this method consists of explaining how these two meanings interact. They do indirectly through so-called implication-complexes or associated implications. An implication-complex is a set of implications predicable to a term. An implication complex A is a set of implications in the form of 'A implies A_i' and an implication-complex B is a set of implications in the form 'B implies B_i'. These implications do not need to be true; they only have to be considered to be true in a given context. The very interaction consists of pairing members of these complexes $f([A_i,B_i])$. The meaning B_i is transformed by f so that it is predicable of A instead of A_i. The function f may stand for an '(a) identity, (b) extension, typically ad hoc, (c) similarity, (d) analogy, or (e) what might be called a metaphorical coupling', (where, as often happenes, the original metaphor implicates subordinated metaphors). ([11], p. 31) Black does not further explicate these terms. For our purposes, we will take *identity, extension, similarity* to be nonfigurative transfers based on a surface similarity. *Analogy* based on a structural similarity and *metaphorical coupling*, based on a subordinate metaphor are, on the other hand, figurative connections of two implications. They are nested metaphors.

Let us illustrate this method with an example of Thomas Hobbes' mechanical metaphor 'Consequence is a train of thoughts'.[3] The implication-complexes, which depend on the context of utterance or reception, might be:

Thomas Hobbes: Consequence is train of thoughts			
Primary subject: consequence	Secondary subject: train of thoughts		
Implications	**Implications**	**Pairing**	**Way of pairing**
consequence is a succession	train implies movement	[succession, movement]	extension
consequence is a link connecting thoughts	train is a link connecting parts	[link, link]	identity
consequence is a causal connection	train connection is mechanic	[causal, mechanic]	extension
consequence is difficult to avoid	train is difficult to stop	[difficult to avoid, difficult to stop]	analogy
Additional implications			
consequence follow logical laws	trains follow timetables	[follows logical laws, follows timetable]	metaphoric al coupling

[2] By 'literal language' we mean the unambiguous language of modern science.

[3] 'BY "consequence," or "train," of thoughts I understand that succession of one thought to another which is called, to distinguish it from discourse in words, "mental discourse."
When a man thinketh on anything whatever, his next thought after is not altogether so casual as it seems to be. Not every thought to every thought succeeds indifferently.' ([13], Ch. III, p. 11.) Hobbes' emphasis on a causal connection between thoughts gives us the reason for taking this metaphor to be mechanical.

Table 1. Interpretation of Thomas Hobbes' mechanical metaphor 'Consequence is a train of thoughts'

The first pair is a case of an extension. The concept of a train's movement is extended so that it covers a succession of thoughts. The second pair is a plain identity. The third pair may be a case of an extension as well. The mechanical way of a train's moving is extended to a broadly causal way of our logical thinking functions (or at least, that is what Hobbes believed). The fourth pair seems to involve an analogy, where the difficulty of bringing a train to standstill is analogous with the difficulty of avoiding a derivation of a consequence. The last pair is a case of an analogy, or a metaphorical coupling. Logical laws are analogous to timetables.[4] However, in which respects? They both express regularities – in a train's movement and in our thinking. Or they both have a normative force, i.e. they both prescribe how things ought to be. There are many aspects in which logical laws are like timetables. Here it is a case of a nested metaphor whose interpretation is open-ended. If this is so, then the interpretation of the original metaphor 'Consequence is a train of thoughts' is open-ended as well.

This example shows that (interpretations of) some metaphors are open-ended or unbounded. This means that such metaphors cannot be easily captured by literal paraphrases. They are absolute metaphors in Blumenberg's sense. Black's interaction theory is, thus, rich enough to be used for analysing absolute metaphors. Black's terminology enables us to recursively qualify metaphors as absolute. A metaphor is absolute if its implication-complexes are connected by analogy or a nested metaphor that is absolute too, because organic and mechanical metaphorics interact here. [5]

Black sees every implication-complex supported by a metaphor's secondary subject as a model of the ascription imputed to the primary subject ([11], p. 31) He develops this theory into the so-called theoretical model. (We describe the memex in terms of a theoretical model in Section 4.) Theoretical models resemble the use of metaphors in requiring analogical transfer of a vocabulary. Metaphor and model creating reveal new relationships. But a metaphor operates largely with *commonplace* implications, says Black, but the author of a scientific model must have prior control of a well-knit scientific theory. Systematic complexity of the source of the model and a capacity for analogical development are essential qualities of models. Black cites another philosopher of science, Stephen Toulmin:

'It is in fact a great virtue of a good model that it does suggest further questions, taking us beyond the phenomena from which we began, and tempts us to formulate hypotheses which turn out to be experimentally fertile… Certainly it is this suggestiveness, and systematic deployability, that makes a good model something more than a simple metaphor.' ([14], pp. 38–39)

[4] To be sure, Hobbes couldn't have had in mind trains as we have today. But wooden railways were common in England in the 17th century. They were used for transporting coal from mines. The fifth implication most probably wasn't intended by Hobbes. However, this need not stop us interpreting the metaphor beyond its author's intention.

[5] There can be other reasons of unparaphrasability as the impossibility to spell out all the implications in practice (because they are too subtle, or there are infinitely many implications, or the metaphorical theme is too abstract). These reasons are not our concern.

A successful model must be isomorphic with its domain of application. In stretching the language, by which the model is described in such a way as to fit the new domain, we pin our hopes upon the existence of a common structure in both fields. If the hope is fulfilled, there will have been established objective ground for the analogical transfer. We can determine the validity of a given model by checking the extent of its isomorphism with its intended application. In appraising models as good or bad, we can, in principle at least, determine the 'goodness' of their 'fit'.

In the next section we move to some deeper characterizations of mechanical and organic metaphorics from a historical perspective. We introduce the dialectical relationship between these two metaphorics on examples from Plato's, Kant's and Alberti's absolute metaphors.

3 MECHANICAL AND ORGANIC METAPHORICS FROM A HISTORICAL POINT OF VIEW

The mechanical, as well as the organic metaphorics has a long history. Mechanical metaphors are usually expressed in terms like 'mechanism', 'mechanics', 'machine', but also by 'construction'. Organic metaphors are connected with 'organism', 'life', 'vitality', 'generative' and its cognates. Mechanical metaphorics mean often-detached elements, atoms, driven by abstract forces that exhibit certain regularities or laws. Mechanisms are constructed or discovered by a *bottom-up* approach where pieces, elements, atoms are *composed* together to give rise to a complex system. Elements are prior to the whole. Organic metaphors, on the other hand, highlight the priority of the whole over its parts or the priority of a principle over its instantiations. Parts are here only because of the whole, which is more than a composition of its parts. Organic systems are recognized by a *top-down* approach where the whole is *decomposed* into its functional subsystems.

The main idea, which drives our investigation, is that of a dialectical relationship between organic and mechanical metaphorics. They are interconnected or even entangled into each other. A mechanical explanation is usually insufficient at a certain point or to a certain extent – an absolute metaphor cannot be fully explained. This gap can be filled by an organic explanation. And this is true also the other way around.

Kant sought in his first *Critique* that nature can be explained by mechanical laws which are derived from the forms of our understanding. This explanation turned out to be insufficient in explaining actions of humans as free beings, but even in explaining some objects occurring in nature like living organisms. They have to be explained teleologically by their inner purposiveness. We can better understand a living organism by asking what its purpose is in nature, not by tracing back its mechanism, which defies any mechanical explanation. Teleological (organic) explanations, however, have for Kant only a heuristic, so to say provisional, role by showing us the directions where to look for mechanical explanations.

The opposite direction is also conceivable. Machines are imitations of organic bodies. This is the traditional Aristotelian view of technology as mimesis. Machines are, in some respect, enhanced bodies (e.g. they are stronger or less prone to malfunctioning), they are, in some other respect, deficient (e.g. they lack

intelligence or they are single-purpose). Here is an illustrative passage from Leon Battista Alberti ([15], p. 175):

'Here we need only consider the machine as a form of extremely strong animal with hands, an animal that can move weights in almost the same way as we do ourselves. These machines must therefore have the same extensions of member and muscle that we use when pressing, pushing, pulling, and carrying.'[6]

Machines are conceived here as extensions of human powers, which is something that will be important in the theories of hypertext. Only (human) organisms as opposed to machines can initiate causal claims.

It is typical that mechanical metaphors aim to explain organic systems and *vice versa*. To use Black's terms, mechanical metaphors are nested in the implication-complexes of organic metaphors. We can, thus, use a mechanic explanation within an overall organic system (and *vice versa*). The decision whether one takes or prefers an organic or mechanical vocabulary depends on the communicative intentions of particular authors. Blumenberg calls this a 'pragmatics function of absolute metaphors.'

In the following three sections we will focus on mechanical and organic metaphorics, their dialectical interplay and blending when explaining the nature of memory, text and hypertext. In order to do so, we, following Blumenberg, need to examine the *consequences* of this or that particular metaphor by hypertext thinkers.

4 MEMEX: MECHANISATION OF ORGANIC MEMORY

We begin this section with an analysis and interpretation of the metaphor 'association is a trail', abstracted from Bush's text. We have chosen it because it helps us to understand as the basic metaphor of hypertext. Engelbart and Nelson (subsequent hypertext investigators) further developed their hypertextual systems from the 'association is a trail' metaphor by developing its open implications. From a theoretical point of view, the 'association is a trail' metaphor fulfils our criteria of an absolute metaphor born from an organic and mechanical background metaphorics. In accordance with Bush, we consider an association as organic, connoted with complexity, unpredictability and intricacy. A trail seems to be more mechanical, systematic, better marked, and easier to follow – at least in Bush's overall aim to mechanize human memory.

Let us follow the directions in a dialectical interaction of organic and mechanic metaphorics in the 'association is a trail' metaphor. Bush describes the methods of mechanical, artificial indexing, which he finds inappropriate at first.

'[…] significant attainments become lost in the mass of the inconsequential […] Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path.' ([7] p. 1)

The mechanical way of linking content is insufficient. Therefore Bush finds a solution in the organic quality of an association:

'The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain. It has other characteristics, of course; trails that are not frequently followed are prone to fade. Items are not fully permanent and memory is transitory. Yet the speed of action, the intricacy of trails, the detail of mental pictures, is awe-inspiring beyond all else in nature.' ([7], p. 6)

Bush sees the mechanical, 'artificial indexing' as more organic, more in line with human associative memory. Bush does not want to explain an 'association' in terms of a 'trail', his aim is to transfer the organic and the mechanical characteristics of associations and trails from metaphor into a mechanical device. So he moves back to a mechanical idea (or the idea of mechanization, more precisely said):

'Selection by association, rather than indexing, may yet be mechanised. One cannot hope thus to equal the speed and flexibility with which the mind follows an associative trail, but it should be possible to beat the mind decisively in regard to the permanence and clarity of the items resurrected from storage.' ([7], p. 6)

Bush finds machine-transferable qualities in associations. The organic is extended by the mechanism of marking (indexing) associations as marked trails to prevent them fading.

Applying Black's method of interpreting predicative metaphors, we are able to find similar directions of the meaning interaction:

Vannevar Bush: Association is trail			
Primary subject: association	Secondary subject: trail		
Implications	**Implications**	**Pairing**	**Way of pairing**
association is a connection of thoughts	trail is connection of places	[connection of thoughts, connection of places]	analogy
association is called into mind by symbols, which are given by some convention	trail is equipped with marks	[association's symbol, trail marking]	metaphorical coupling
it is hard to remember associated items without remembering the convention, i. e. by mnemonic devices	it is hard to follow a trail without maps and marks	[mnemonic devices, maps and marks]	metaphorical coupling
a not followed association is prone to fade	a not used trail fades	[association's fading, trail's fading]	analogy

Table 2. Interpretation of Vannevar Bush's mechanic metaphor 'Association is trail'

[6] Quoted from Blumenberg [6], p. 67.

The first pair of implications is an example of an analogy. The implication 'trail is a connection of places' is analogous to 'an association is a connection of thoughts'. The primary, organic subject is seen in light of the secondary, mechanical subject. The better-known concept of the trail is extended so that it covers an association. The second pair is an example of metaphorical coupling. Trails are usually provided with marks. Such marks are metaphors for symbols by which associations are called into mind. The third pair may be a case of a metaphorical coupling again. We use marks, or more generally maps, in order to follow trails. In our metaphor we use mnemonic devices in order to follow our association, or to remember associated items. The fourth pair seems to be the case of a metaphorical coupling too: Disused trails fade. This is analogous to a not followed association. They are prone to fade.

Black says, the literal meaning of 'an association' interacts with the literal meaning of a 'trail' resulting in a metaphorical meaning of a 'trail' which is hereby being predicated by an 'association'. The very same metaphor says something about the secondary subject: Bush sees a 'trail' in the light of an 'association'.

'An association is a trail' is a case of absolute metaphors in Blumenberg's sense. It is the unifying representation, which help us to orient in the evolving concept of hypertext. In this stage of discovering hypertext, it is not possible to translate its idea into unambiguous, scientific language. There is no existing technology allowing us to run the memex. There is no scientific terminology yet and it would not be fruitful to establish it. The inventor is only able to show the first orientation of his ideas. In the next step he develops his metaphor into a theoretical model of the hypertextual linking of information, a memex. Nevertheless, a detailed analysis of the memex is a theme for a more detailed investigation. We can only confirm the memex as a fruitful theoretical model in this paper due to the following reasons: The memex resembles the use of metaphors in requiring an analogical transfer of vocabulary. Bush wants to mechanise an organic association trail in his memex. His aim is to improve an organic, transitory memory by means of a mechanical, permanent trail of an association. Bush speaks about the mechanical memex using the terminology of an organic, associative memory. In stretching the language by which the associative memory is described, in such a way, as to fit the new domain (memex), Bush pins his hopes upon the existence of a common structure in both fields. His hope is fulfilled, so there is objective ground for the analogical transfer.

We can describe the memex in the terms of Black's model as a 'system of imaginaries' ([2], p. 234). Bush concentrates on the principle of indexing associative trails. The memex allows the establishing, marking and following of associative trails to be permanent. The memex is supposed to add the organic factors of speed and convenience to the ordinary mechanical filing-system processes. Bush is aware that it cannot work at the same speed as an organic, human memory. But he believes it will be possible in the future, that new technologies will allow future machines to work at the same speed as humans can think. This example shows that Bush was not limited by considering only the real means that were available to him. He built a model, a system of the possible, system of imaginaries. We can consider such a system, pragmatically built as an equilibrium to be consisting of both the organic and mechanical qualities of a human and a machine.

According to Black ([2]), we appreciate the memex as a very vital model. The memex is based on implications rich enough to suggest novel hypotheses and speculations in the primary field of investigation. It suggests further questions, it takes us beyond the phenomena from which we began, and it tempts us to formulate hypotheses which turn out to be experimentally fertile in the future of hypertext development. Bush supposes that clever usage of an associative trail manipulation can augment human associative memory. As we will show in the next section, his concept of associative linking content was inspirational in the questions of human intellect augmentation, by means of a technological extension.

In this section we have analysed Bush's metaphorical thinking in detail. Seen in the broader context of hypertext inventing, the mechanisation of organic qualities of a human mind is essential for contemplating hypertext. In the following section, we will show how the direction of interaction changes. The new direction will lead us to the following question: how can a system of mechanised associations become more organic by means of human machine interaction and cooperation? Will this be fruitful to think about mechanical devices in terms of a text?

5 NLS: INTERACTION BETWEEN HUMAN AND MACHINE

We tried to find some innovative metaphors about content linking for our analysis of Engelbart's text. Nevertheless, Engelbart uses Bush's metaphor mentioned above. In this section we analyse and interpret the metaphor 'a computer is a clerk'[7], abstracted from Engelbart's text *Augmenting Human Intellect: a Conceptual Framework* [8]. We believe that it is helpful in our understanding of the next metaphorics turn and also in the context of hypertext development. We will complete our analysis with an interpretation of Engelbart's NLS system. As will become evident, Engelbart speaks about this machine in the same way as a text, which is an essential direction for hypertext development.

The 'a computer is a clerk' metaphor fulfils our criteria of an absolute metaphor, because it is created as an analogy of an organic and mechanical subject. Allegedly, a computer seems to connote mechanic qualities whereas a clerk is organic, connoted with human qualities. Based on the analysis following Black's interaction theory, we argue that Engelbart turns to see a machine being more organic: as a human being and, in the case of the NLS system, as a text.

Engelbart begins his paper with the task of augmenting the human capability to solve problems:

'By "augmenting human intellect" we mean increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems.' ([8], p. 1)

Engelbart's main aim is to invent a means that would make the individuals, intellectually more effective, by means of a human-computer interaction:

[7] 'Let us consider an augmented architect at work. He sits at a working station […]; this is his working surface, and it is controlled by a computer (his "clerk") with which he can communicate by means of a small keyboard and various other devices.' ([8], p. 70)

'We see the quickest gains emerging from (1) giving the human minute-by-minute services of a digital computer […], and (2) developing new methods of thinking and working that allow the human to capitalize upon the computer's help. By this same strategy, we recommend that an initial research effort develop a prototype system of this sort aimed at increasing human effectiveness in the task of computer programming.' ([8], p. 3)

Engelbart uses the analogy of a computer as a clerk, as a 'fast and agile helper'[8].

Douglas Engelbart: Computer is clerk			
Primary subject: computer	Secondary subject: clerk		
Implications	**Implications**	**Pairing**	**Way of pairing**
computers have users	clerks have supervisors	[user, supervisor]	analogy
computer is a fast and agile helper	clerk is an agile helper	[helper, helper]	metaphorical coupling
computer is programmed	clerk have to follow rules and laws	[following programs, following rules]	analogy
computers work mechanically	clerks do a lot of mechanical routines	[mechanic work, mechanic routines]	analogy
computers are without emotions and errors	clerks have to avoid emotions and errors	[mechanic, suppressing organic qualities]	analogy

Table 3. Interpretation of Douglas Engelbart's organic metaphor 'Computer is a clerk'

The first pair of implications suggests that computer users are analogous to clerks' supervisors. Engelbart imagines the computer of the future in terms of human collaboration, as a mechanic helper, which needs to be programmed and led by his organic supervisor. The idea of programming is essential in the concept of interaction. The second pair of implications shows that, for Engelbart, a computer is a fast and an agile helper. A clerk is also seen usually as an agile helper. Something mechanical (computer) is analogous to something organic the (clerk). Only mechanical features of clerks are transferred according to this metaphor. We select only the mechanical features of an organic secondary subject. This is going to be explicit in our third implication: Computers work mechanically whereas clerks perform a lot of mechanical routines. The direction of interaction (from 'A to B' or 'B to A') is evident in the last implication. Clerks should be free of emotions in order to avoid errors. They have to suppress their organic qualities and work mechanically. Their mechanised, programmed way of working is now transferred into computers.

Seen from a metaphorological perspective, Engelbart follows his contemporary influential thinkers. Licklider [16] speaks of 'man-computer symbiosis' and Ulam [17] uses the term 'synergesis'. Most comprehensive is Ramo's [18] term 'synnoetics', applicable generally to a cooperative interaction of people, mechanisms and automata into a system whose mental power is greater than that of its components. We find these organic and mechanical metaphorics to be leading at the beginning of the digital age. Engelbart's text reflects the difficulties with describing his images about the future and possible reality, in the way of literal and scientific terms, Reading between the lines here, he creates his vision in the figurative way of imaginations and he supposes this way to be more comprehensible to his readers.

'The picture of how one can view the possibilities for a systematic approach to increasing human intellectual effectiveness, as put forth in Section II in the sober and general terms of an initial basic analysis, does not seem to convey all of the richness and promise that was stimulated by the development of that picture. Consequently, Section III is intended to present some definite images that illustrate meaningful possibilities deriveable from the conceptual framework presented in Section II. The style of Section III seems to make for easier reading. […] Section III will provide a context within which the reader can go back and finish Section II with less effort.' ([8], p. 3)

However, let us return to the pragmatic reasons for hypertext discoveries. We have to mention Engelbart's account of linking. In the third section of his *Augmentation*, Engelbart comments on Bush's main ideas about a hypertextual content linking, derived from the 'association is a trail' metaphor. From a technical point of view, Engelbart continues in Bush's effort to mechanise linking information by indexing. He broadens this task, because he thinks about links and connections as about interactions. The literal meaning of interactions stresses the meaning of a two-way connection and communication, just like the meaning of feedback. Engelbart with his team was capable of creating a functional, collaborative knowledge environment system called the NLS (for oNLine System). (It was first demonstrated in 1968.) Engelbart's lab used NLS for all its own knowledge work, drafting, publishing, shared screen collaborative viewing and editing, document cataloguing, project management including a shared address book – all of these in an integrated hyper groupware environment. It was possible to edit the structure as well as the text.

While Bush saw the memex as a tangible, a mechanised, a personal library, Engelbart considered the NLS to be an editable text with rewritable links. He saw it as a sort of self-organizing retrieval system, which dealt with the symbolic structures by means of programming.

How does the direction of the organic and mechanical metaphorics interaction change with Engelbart? Engelbart sees mechanical devices in the light of organic, human qualities, interacting by means of symbolic communication. He tries to put the mechanic implications nested in organic terms (i.e. systematization, logic, routines) into machines and augment them. He suppresses (for his pragmatically determined aim) any undesirable organic characteristics in his machine, (i.e. a high error rate, forgetfulness, tiredness etc.). In the next step Engelbart tries to improve mechanical devices by means of suitable organic qualities (i.e. the ability of symbolic communication, ability of feedback, speed of associative processes etc.). In contrast to prior historical eras, he started to explain organic qualities as nested in mechanical metaphorics. Or we can say, the metaphorics of the mechanical is replaced by the metaphorics of programming.

With these thinkers considering pursuing this direction, the metaphor of the mechanical is now becoming corrected (or furthermore developed) by the metaphorics of the algorithmisation. In the next section, we will follow how the text becomes

[8] 'Such a fast and agile helper as a computer can run around between a number of masters and seldom keep any of them waiting […] ([8], p. 70)

hypertextual in Nelson's thinking, and the figurative conceptualising of the new information media.

6 XANADU: ORGANIC MACHINE AS MORTAL MACHINE

In this section we analyse and interpret the metaphor 'a hypertext is a Xanadu', abstracted from Nelson's hypertextual project [20]. Nelson coined the term 'hypertext' and defines its properties in 1965 ([21], p. 96). In *Literary Machines* ([9], p. 30) he describes his most famous hypertext project Xanadu as a 'magic place of literary memory'. His hypertext concept is supposed to be analogical to Coleridge's Xanadu [22]. We will concentrate on Nelson's implications from this metaphor.

Nelson wants to transcend the possibilities of textual form, determined by the qualities of mechanical printing machines. The metaphor, which he chooses, answers this purpose. We can see the connection with Engelbart's approach. Nelson and he sees a machine as a text. While Engelbart only notices this analogy, Nelson is able to develop it in a very detailed way with the help of figurative language, but also in unambiguous, scientific definitions of hypertext qualities. The word 'hypertext' we can consider as specific type of metaphor, catachresis, which, according to Black, fulfils the gap in the existing vocabulary. As with Musil's man from the epigraph, with a sense for the possible, he abstracts from the given (mechanical) reality which is insufficient for him:

'The sense of "hyper-" connotes extension and generality; […] The criterion for this prefix is the inability of these objects to be comprised sensibly into linear media […]. ' ([21], p. 98)

Hypertext is the presentation of information as a linked network of nodes which readers are free to navigate in a non-linear (organic, associative, creative) fashion. Nelson does not want to mechanise the organic, as Bush did. Most of all, he wants to create a new, more organic, more human media. Which organic qualities does he transfer into his literary machine, i.e. hypertext? He wants to teach machines human skills such as writing and reading. The Xanadu user is the reader and the writer of the text at first. And he is a programmer too. As Fuller and Goffey [23] show, programming is a new use of a language and the language has a very organic, human quality.

Ted Nelson: Hypertext is Xanadu			
Primary subject: hypertext	Secondary subject: Xanadu		
Implications	**Implications**	**Pairing**	**Way of pairing**
hypertext concept is rich	Xanadu offers a lot	[rich, offers a lot]	analogy
hypertext is a text with a new dimension	Xanadu is a magic place	[new dimension, magic place]	metaphorical coupling
hypertext is a text with references to other texts	Xanadu is a place of literary memory	[web of texts, literary memory]	analogy

Table 4. Interpretation of Ted Nelson's organic metaphor 'Hypertext is a Xanadu'

Nelson explains his hypertext as a Xanadu. The first pair of implications suggests that the concept of hypertext is as rich as a Xanadu. The second pair of implication-complexes is a case of metaphorical coupling: a Xanadu is a magic place in Coleridge's poem, while Nelson's hypertextual Xanadu adds a new dimension to the text. Coleridge's Xanadu transcends the materiality of our world, hypertext remediates materiality of 'paper' with its qualities. The third pair of implications defines Xanadu as a place of literary memory. This is analogous to hypertext being a text with references to other texts. Coleridge's Xanadu is a metaphor for the never-ending finding of a magical place. It is dedicated to active and creative users. It functions, after forty years of development in a limited version. It will stay in a dream as in Coleridge's Xanadu. It is too difficult to be the main principle of the contemporary leading hypertextual system, the more mechanical WWW. As Nelson says:

'Today's popular software simulates paper. The World Wide Web (another imitation of paper) trivializes our original hypertext model with one-way ever-breaking links and no management of version or contents.' ([20])

The reason is pragmatic: for general purposes we need an easier solution. In this aspect, the historical dialectical interplay of metaphorics, at the turn of the twentieth and the twenty-first century, shows us that a more mechanical medium is more vital than an organic one. But Xanadu has a chance to inspire a specialised, professional system for scientists and people who have to think in a more complex way. Or, we can change the direction of metaphorics, and go along with Rushkoff, to suit people, who do not want to be programmed, but want to programme [24].

7 CONCLUSIONS & FUTURE WORK

The common pattern of the analysed metaphors in Black's interactive view is that the interaction of the meanings in them goes in two ways. The implications of the mechanical and the organic metaphorics are nested one in the other and therefore these metaphors are absolute in the Blumenberg sense. The interpretations of our metaphors are open-ended and fruitful for new concepts of hypertext. We applied this idea in models and concepts of hypertext: All of our hypertext thinkers speak about the human-machine interaction in terms of finding the best equilibrium of the possible and the real, of organic and mechanical qualities. The direction of their investigations leads from the need of a mechanical machine, based on organic principle to a new medium, based on the transfer of many human organic qualities and skills into an interactive medium.

Bush mechanised the way of human, organic associative indexing and makes mechanical ways of indexing more organic, more in line with human thinking. On the other hand, he contemplates the mechanisation of associations.

Engelbart's hypertextual equilibrium stresses the interaction of human (organic) and mechanical (computerized) elements. He speaks mostly in terms of mechanic qualities nested in organic, human elements. He stresses the idea of seeing a machine as an (organic) text, as a medium.

Nelson builds upon his predecessors' idea, that the medium is more organic. He wants to transcend the possibilities of the textual form determined by the qualities of mechanical printing machines. He speaks about hypertext in more organic terms. His concept is very organic and therefore mortal, as we have shown.

In the period in question, the history of the concept of hypertext started with an organic metaphor of association. It continued through the idea of mechanisation and furthermore through the

idea of organic-mechanical interaction and was complemented by the organic metaphorics of reading, writing and programming. In the context of hypertext discovering, a mechanical solution became insufficient. This insufficiency is supposed to be filled by an organic solution. The next step consists in the mechanization of organic qualities, and the following one in their algoritmisation in the era of digital media.

Absolute metaphors, as metaphors in general, fulfil the function of stressing some aspect of the source domain. This function is pragmatically determined. In our case the pragmatic reasons are the following:

(1) to augment human intellect by mechanical means,

(2) to enable other people to understand such difficult thoughts, as Musil's 'unawakened realities', which are not translatable into the literal language of science. ([1], p. 12)

The history of media is the history of attempts at understanding human, organic qualities and to use them as extensions by transferring them into machines. After a successful transfer, the direction of this interaction then changes. Now we start to use media as a translation, as a metaphor for explaining human, organic qualities. It seems that in the era of algoritmisation[9], the metaphorics of mechanical machines have lost its importance. It has been corrected by the metaphorics of the digital media, just as the metaphorics of linear (mechanical) text has been corrected by the metaphorics of (organic) hypertext.

In our future work we will continue pursuing the history of this metaphorics in relation to the WWW. We expect to interpret it as a mechanised organic medium of Nelsonian hypertext. We see the importance in investigating more unique hypertexts such as scientific ontologies.

REFERENCES

[1] R. Musil. *The Man without Qualities*. Trans. by Sophie Wilkins. Vol. I. Vintage Books, London, UK (1996).

[2] M. Black. *Models and Metaphors: Studies in Language and Philosophy*. Cornell University Press, Ithaca, USA, pp. 219–243 (1962).

[3] T.S. Kuhn. Metaphor In Science. In: *Metaphor and Thought*. A. Ortony, (Ed.). Cambridge University Press, Cambridge, UK, pp. 409–419 (1979).

[4] R. Boyd. Metaphor and Theory Change. In: *Metaphor and Thought*. A. Ortony, (Ed.). Cambridge University Press, Cambridge, UK, pp. 481–532 (1979).

[5] H. Arendt. Metaphor and the Ineffable: Illumination on 'the Nobility of Sight'. In: *Organism, Medicine, and Metaphysics*. S.F. Spiecker, (Ed.) Springer, Dordrecht, Netherlands, pp. 303–316 (1978).

[6] H. Blumenberg. *Paradigms for a Metaphorology*, trans. R. Savage. Cornell University Press, Ithaca, USA (2010). For orig. see: *Paradigmen zu einer Metaphorologie*. Insel, Frankfurt am Main, Germany (1999).

[7] V. Bush. As We May Think. *Athlantic Monthly*, July 1945, 176 (1):101–108 (1945).

[8] D.C. Engelbart. *Augmentation Human Intellect: A Conceptual Framework*. SRI, Menlo Park, USA (1962).

[9] T.H. Nelson. *Literary Machines*. Mindful Press, Sausalito, USA (1981).

[10] V. Bush. Memex Revisited. In: J.M. Nyce and P. Kahn (Eds.): *From Memex to Hypertext*. Academic Press Professional, Inc. San Diego, CA, USA, pp. 197–216 (1967).

[11] M. Black. More about Metaphor. In: *Metaphor and Thought*. A. Ortony, (Ed.). Cambridge University Press, Cambridge, UK, pp. 19–43 (1979).

[12] G. Lakoff and M. Johnson. *Metaphors We Live by*. The University of Chicago Press, Chicago – London, USA – UK (1980).

[13] T. Hobbes. *Leviathan*. Critical edition by Noel Malcolm. Oxford University Press, UK (2012).

[14] S.E. Toulmin. *The Philosophy of Science*. The Mayflower Press, London, UK, https://archive.org/stream/philosophyofscie032167mbp#page/n43/mode/2up. (1953).

[15] L.B. Alberti. *On the Art of Building*, trans. J Rykwert, N. Leach and R.Tavernor. MIT Press, Cambridge, USA (1994).

[16] J.C.R Licklider. Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, (March 1960).

[17] S.M. Ulam. A *Collection of Mathematical Problems*. Interscience Publishers, Inc., New York, USA (1960).

[18] S. Ramo. The Scientific Extension of the Human Intellect. *Computers and Automation*, February, (1961).

[19] J.M. Lawler. *Metaphors we compute by*. http://www. ling. lsa. umich. edu/jlawler. (1987).

[20] T.H. Nelson. *Project Xanadu*. http://xanadu.com. (2014).

[21] T.H. Nelson. *A File Structure for The Complex, The Changing and the Indeterminate*, Proceeding of the ACM 20th National Conference, ACM, August. pp. 84–100 (1965).

[22] S.T. Coleridge. *Christabel: Kubla Khan, a Vision; The Pains of Sleep*. William Bulmer, London, UK (1816).

[23] M. Fuller and A. Goffey. *Evil Media*. MIT Press Cambridge, USA (2012).

[24] D. Rushkkoff. *Program or Be Programmed*. OR Books, New York, USA (2010).

[9] According to Fuller and Goffey, it is a process of a reality that occurs in the conversion of process, which we know from the physical world into sequence of writable and readable algorithms that drive our human-media interaction. ([23], p. 80)

Metaphor, Meaning, Computers and Consciousness

Stephen McGregor[1] and **Matthew Purver**[2] and **Geraint Wiggins**[3]

Abstract. This paper seeks to situate the computational modelling of metaphor within the context of questions about the relationship between the meaning and use of language. The results of this pragmatic assessment are used as the theoretical basis for a proposed computational implementation that seeks metaphor in the geometry of a vector space model of distributional semantics. This statistical approach to the analysis and generation of metaphor is taken as a platform for a consideration of the fraught relationship between computational models of cognitive processes and the study of consciousness.

1 Introduction

Aristotle is commonly credited as the earliest thinker to seriously consider metaphor as a linguistic device, lauding its use as an indication of the highest level of genius [1]. But while a historical account of scholarship about metaphor is a worthwhile topic, and one which will feature throughout this paper, the history of metaphor itself is as convoluted and unobtainable as the history of language. In fact, if it serves any purpose to think about such a remote event as the inception of language, it seems impossible to imagine a clever speaker not immediately taking the agreed definitions of the world's first words and doing something unexpected with them. If anything, a more accurate take on early academic discussion of metaphor might be to consider Aristotle as one of the first philosophers to ponder the question of the relationship between what words mean and what words do.

This paper will seek to evaluate metaphor from a pragmatic point of view, and to situate this evaluation in terms of a framework for the computational analysis and generation of metaphor. This marks a shift from what has become the standard computational approach to metaphor, which considers language in terms of formalisms that are intuitively compatible with symbol manipulating machines. Implicit in these standard approaches is the assumption that words and concepts exist on different levels of abstraction, and that metaphor is a product of a process of transference or mapping that occurs on the conceptual level, with words acting as a kind of index of this process. But the idea that words merely point to concepts runs into trouble in light of certain properties of metaphor that cannot be explained in terms of an abstract conceptual construct of the entities nominated by words. At the root of the approach proposed in this paper is a contention regarding the difficult topic of consciousness: metaphor is often based on the direct experience of perception, and the ease with which a cognitive agent can express the actual quality of one particular percept in terms of the idea of another general percept is rooted in the direct connection between phenomenology and language. The very relevance of the term "like" to figurative language, manifest when metaphor is translated into simile, suggests that the "likeness" of the conscious experience of qualia is intrinsic in the perpetually unfolding construction of metaphor.

One of the claims made in this paper is that consciousness is always understood metaphorically, and one of the most pervasive and at the same time disputed contemporary metaphors involving consciousness has been the trope that casts the mind as a computer. This particular construct is compelling, in that the mind can be conceived of as having input in the form of perceptual stimulation and output in terms of either conceptualisation of the world or directed action in the world. At the same time, the analogy is disreputable in its relegation of the richness of consciousness to the domain of a rule following, data processing apparatus that is subject to an arbitrary, observer relative interpretation. It seems that a good model of metaphor should explain the appeal of comparing the engine of its own operation – the mind – to a device that is arguably at best just an aid to thought. The model should also account for the perceptual, imagistic aspect of metaphor-making, evident in light of the necessity of comparing one experience to another when trying to describe what it is like to be conscious.

The solution offered here involves turning to high dimensional representations of meaning based on a statistical analysis of the distribution of words in large scale corpora, and, in so doing, embracing the modelling power of the computer, if not the explanatory power of the mind-computer metaphor. The theory behind the system that will be described is based on the idea that a statistical treatment of a large collection of words found in their natural habitat, so to speak, can simulate the construction of a space of meanings. This space, in turn, becomes the linguistic environment in which metaphors are discovered in the process of solving communicative problems: congruences in the geometries of these statistical word-objects suggest ways in which they can be combined in order to construct expressions. The metaphor-making procedure, modelled as a fundamental aspect of ongoing entanglement with a richly informative environment, is finally presented as a key component in the expression of consciousness, a characteristic that may shed some light on the evident propensity for qualia sensing agents to project their own consciousness onto everything else in the world.

2 Consciousness Is a Metaphor

The tension that metaphor has traditionally introduced to the study of language has arisen from the dynamic between words and truth: figurative statements that are clearly contrary to the facts of reality are nonetheless effective at conveying truthful information about the world. This aspect of metaphor poses at least a superficial problem for truth conditional approaches to semantics, which hold that there is either a correspondence between propositions and the world they

[1] Queen Mary University of London, email: s.e.mcgregor@qmul.ac.uk
[2] Queen Mary University of London, email: m.purver@qmul.ac.uk
[3] Queen Mary University of London, email: geraint.wiggins@qmul.ac.uk

portend to describe [36], or a coherence between the set of propositions that collectively constitute a truthful system of beliefs [13]. For Floridi, the imperative of truthfulness means that "semantic information" is necessarily defined in terms of data that remits "veridicality" in relation to the world that it models [19]. Dretske likewise distinguishes between information and the the semantic representation indicated by a correct interpretation of that information [16].

Taking Dretske's ideas about indication and interpretation as a point of departure, it is possible to formulate a theory whereby the truthfulness of figurative propositions lies in the correct interpretation of the intention behind a non-veridical statement. Here metaphor becomes a mechanism for encoding information, with the projection from source to target allowing for the transference of a set of intensions from a general case of the source to a specific instance of the target. If this is the case, then a metaphor can be deciphered into a more extensive array of literal propositions. The well studied metaphor "that surgeon is a butcher", for instance, takes the bloodiness and brutality stereotypically associated with the profession of a butcher and efficiently applies them to the behaviour of some disreputable surgeon. This packaging of literal information sits well with Searle's approach to metaphor, which sees non-literal language as an invitation to interpretation based on propositional knowledge of the world shared between two interlocutors [35]. Ortony, in his "reconstructivist" theory of metaphor, has even suggested that there must be some sort of mental imagery involved in the interpretation of figurative language: a metaphor evokes a non-literal scene which effects the vivid transference of intension in a way that invites logical inference [31]. This move introduces conscious perception to the explication of metaphor, with the experience of a mental state playing a direct role in the transmission of richly detailed information.

But how can consciousness ever be discussed in a way that is literal or veridical? If qualia, with their intrinsically subjective character, are the substance of conscious experiences, then it seems impossible to describe such phenomenological conditions in terms of truthful propositions about situations in the world. Chalmers has made much of this divide between subjective conscious experience and objective physical reality, focusing in particular on the difficulty of determining the truth conditions of a report of a phenomenological perception [9]. From a phenomenological perspective, the defining characteristic of consciousness is that there is something it is like to experience qualia, and this very "likeness" of the experience immediately suggests the application of analogical conceptualisation and correspondingly metaphoric expression. While a mutually agreed description such as "red thing" might allow two interlocutors to pick out a set of objects with some shared characteristic, it is not clear that there is any way to know that the actual phenomenology of the red experience is similarly shared. Since there is no way to expressively project the actual conscious experience of perceiving an object, a descriptive speaker who wishes to convey something phenomenological is left with no choice but to resort to an act of analogy, giving the world such poetic turns of phrase as "lips as red as blood" or "eyes as blue as the sky".

Along these lines, Everett has highlighted the absence of abstractly quantifiable colour terms in the language used by the Pirahã people of Brazil, who instead employ standardised expressions that are fundamentally figurative: the color term corresponding to what an English speaker would describe as "red", for instance, transliterates to the expression "bloodlike", and "black" becomes the phrase "blood is dirty" [17]. Levinson reports similar findings in his analysis of the Yélî Dnye language spoken by the inhabitants of an isolated island near Australia, who use the terms for various birds and plants to describe other similarly coloured objects [29]. Even if, as Kay and Maffi claim, the lack of fixed absolute colour partitions in a language is anomalous [27], the admission of chromatic descriptions such as "chartreuse", "coral", or "eggplant" in English illustrates the ease with which a perceptual experience of one thing can be converted into a classification of something else. There is an inherent process of analogising occurring when cognitive agents turn to language to express the subjective characteristics of their perceptual existence.

This perpetual trafficking of intension from one perceptual or conceptual domain to another extends especially into more general descriptions of consciousness. The difficulty of discussing qualia in objective and material terms has compelled philosophers to resort time and again to thought experiments involving components fantastically removed from reality – beetles in boxes, homunculi in theatres, deceptive demons – in order to allude circumspectly to what it is like to be conscious. Even Dennett, who has questioned the efficacy and indeed the existence of qualia [14], acknowledges that it is generally necessary to employ analogical reasoning when dealing with descriptions of mental processes [15].

There is a temptation to take the necessity of analogy in discussion of consciousness one step further by way of construing consciousness itself as a process of metaphor-making. In the 1970s, Jaynes proposed his bold "bicameral" theory of mind based on the idea that pre-literate humans had perceived their own consciousness as a mentally external expression of instructions and proclamations experienced as ongoing auditory hallucinations [26]. To a mind sundered in such a way, the modern experience of self as realised through subjective phenomenology was supposedly replaced with a personal fictive narrative that cast the consciously feeling component of the mind in the role of a god or a commanding spirit. This controversial theory has received some recent support, at least implicity, in Carruthers' formulation of "interpretive sensory-access" based on the mindreading faculties that facilitate the acts of interpretation at the centre of consciousness [6]. In a propositional reversal that nonetheless maintains some of the core tenets of Jaynes' bicameral mind, mindreading capacities can be applied not only to introspection, but also to the interpretation of the mental states of other people and even as the projection of mind-like faculties on objects that are obviously actually inanimate. So, for instance, it seems quite reasonable to metaphorically discuss the temperament of things like computers, cars, appliances, or the weather without the presumption that these types of objects actually have minds.

If these projective theories of mind are to be taken seriously, then the essential role of metaphor in consciousness must be considered. There certainly seems to be a case to be made for the idea that consciousness necessarily involves a transgression of literal conceptualisation of the world, a transference of a feature from one mental object to another that results in an expression of the experience of a thing as something other than what it actually is. There are three propositions at stake here. The first is that the only feasible mechanism for communicating about the experience of consciousness is to cast the description of that experience out onto some universally accessible entity with qualitative attributes that will hopefully simulate the experience. The second is that the mind can only be understood in terms of things other than minds, things that have mind-like properties and therefore analogically corroborate an explanation of what it is like to have a mind. The third is that having a conscious mind necessarily involves the projection of phenomenological characteristics onto external entities, some that presumably are likewise conscious and others that almost certainly are not. In each of these cases, through experiential transference, through analogical descrip-

tion, and through projection of the self onto another, an essentially metaphoric process is at play: knowledge of the mind seems to consist of a network of proxies and equivalences that trace the outline of the thing that they don't quite touch.

3 Words Are Objects

The recent history of theoretical approaches to metaphor has been characterised by an intellectually productive tension, with both sides notably departing from any notion that the figurative use of language should somehow be treated as an exceptional case. On the one hand, there are those who would describe metaphor as a transference or projection of intensionality from the conceptual space of a source to the similarly oriented space of a target, a view that found an early champion in Black and his "interactionist" theory of metaphor [3, 4]. By this account, metaphor involves conceptual mappings that place a non-literal source at the centre of the "implicative complex" of a targeted conceptual system, so that characteristics of the way the source does things are projected onto similar activities undertaken by the target. On the other hand, a dissenting contingent of theorists have argued that the metaphoric use of language stands entirely outside the realm of conceptualisation, and that the meaning of any sentence can only be interpreted literally—an idea originally expounded by Davidson [12], with early support coming from Rorty [34].

In the early 1960s, Hesse argued for the importance of analogy as a tool for scientific understanding [25]. At the root of her argument was the idea that all theories are ultimately models of the world, and that, in terms of the extreme scales involved in, for instance, the study of physics, these models could only be grasped in terms of metaphors: so, for instance, a distributed gas bears an analogy with a space full of colliding and rebounding balls. The study of metaphor subsequently underwent a Renaissance of sorts, with a flurry of research throughout the 1970s (see [32] for a compendium of exemplars), culminating in Lakoff and Johnson's case for an understanding of metaphor as a mapping between isomorphic conceptual schemes [28]. This theory presented metaphorical language in terms of its relationship to an embodied cognitive experience of the world, so, for instance, the analogy which maps the conceptual situation between "up" and "down" to the situation between "happy" and "sad" is a product of the actual culturally loaded experience of orientation in the real world. A lattice of networked spaces, extending from the world through perception and conceptualisation into language, allowing for the transference of entire isomorphic conceptual complexes: if a surgeon is a butcher, then hospitals become abattoirs and patients become animals.

Davidson, however, offered a dissenting interpretation of metaphor, springing from his rejection of the idea that language should be talked about as a system for conceptual representation in the first place [11]. Instead, he proposed that the meaning of a metaphor could only be considered in terms of the literal proposition made by a metaphorical statement, and that the operation of a metaphor in the process of communication must be considered as something altogether outside the realm of meaning [12]. This stance has met with considerable resistance, finding an early opponent in Bergmann, who argued that Davidson's critique only applied to decontextualised encounters with metaphor; once the metaphor is put into the context of a situation involving a speaker with an intention, it can be clearly seen to have a meaning [2]. Hesse also revisited her case for metaphor as a fundamental cognitive operation, arguing that all language is metaphoric in that all language plays a protean role in a nebulous network of meaning [24]. Rorty, on the other hand, came to Davidson's defence, interpreting his approach as placing metaphor actually in the world of natural events rather than consigning it to an essential role in an interplay of symbols that is ancillary to reality [34]. By this reading, language is not to be considered as a model or representation of reality, but rather as a component directly in reality, existing on the same level of abstraction as impressions and ideas.

The debate over metaphor in subsequent years has involved a back and forth between those who see metaphor as by-product of an essential cognitive operation and those who claim that language plays a more fundamental role in perception of the world, though Davidson has arguably been broadly misinterpreted. In an expansive consideration of metaphor as evidence of "the poetic structure of mind", Gibbs suggests that Davidson places emphasis on first determining the literal meaning of a metaphor and then accepting that the potential non-literal meanings of the phrase are somehow infinite and unknowable [22], perhaps a misreading of Davidson's contention that "there are no unsuccessful metaphors". As a recent proponent of the non-cognitive take on metaphor, though, Carston has recast Davidson's rejection of cognitive content in terms of a more fundamental "imagistic" feature of language [7]. In particular Carston considers the metaphor "Bill is a bulldozer": the interpretation of this phrase as a description of a man who is grossly aggressive and inconsiderate is clear, but upon further analysis there is no literal property of a piece of equipment such as a bulldozer that bears the inherently human intensions being drawn out in Bill [8]. At best there might be an argument that a double metaphor is being employed here, with a bulldozer standing in for something aggressive and then Bill being described as one of those things, but this introduces a combinatorial explosion of ways to frame all but the simplest metaphors and in so doing seems to miss the point of the cogency of figurative language. Instead, it seems reasonable to say that the metaphor evokes something that is not purely in the realm of language, a direct perception of Bill as a potentially destructive machine.

In this analysis, Davidson and his acolytes emerge as something of the arch-pragmatists. Rather than keeping the construction and interpretation of metaphor on a symbolic level, where language models the world it describes, here the very meanings of the words employed in a metaphor become implements to be handled and used to accomplish communicative goals in the same *ad hoc* way that a more overtly physical object might be picked up and used. Meanings exist, but as the features of elements of language that suggest their functionality: in fact, the meanings of words themselves become the intensions of those words, suggesting potential uses of language in the way that, for instance, the solidness and heaviness of an object might recommend it as a weapon to an attuned perceiver in need of such a device. Just as a shoe might present itself as a hammer under the right circumstances, or a stick or rock as a writing instrument, the word "bulldozer" offers itself as the right term to convey Bill's comportment in the same grasping process of perception and cognition, because language is actually happening on exactly the same level as the rest of existence, not in an abstract secondary space.

At this point, language can be situated in the context of Gibson's theory of affordances, which holds that cognition arises in the process of the perception of opportunities for action in an environment [23]. Clark has worked towards expanding environmentally situated approaches to cognition into the domain of linguistics, describing the "persisting but never stationary material scaffolding" of language [10]. A picture emerges of language use as a process of scavenging a shifting space of meaning for the words that can be used to accomplish some expressive task. These meanings are not representational models that stand in a relationship of signification to perceptions and

conceptions of the world; they are the cognitive detritus of entanglement in an environment that involves communication with other linguistic agents, sitting right alongside other mental experiences of reality.

So an alternative approach to modelling metaphor emerges, one that does not involve considering the language involved in metaphor-making as simply a corollary to mappings between isomorphic conceptual spaces. Instead, metaphor can be envisioned as a process of searching a space of linguistic percepts for the sounds or symbols that can be arranged to fulfil some communicative requirement. The challenge then becomes defining this space of meanings and understanding how word-objects are selected from it. This theory does not refute the descriptive power of Lakoff and Johnson's ideas about conceptual metaphors; in fact, it seems clear that there must be some discernible aspect of meaningful entities that allows them to be cobbled into a pragmatically efficacious structure, and it seems likewise reasonable to construe this act of construction as an aligning of mental objects. As an explanatory device, though, the idea that metaphoric language simply corresponds to congruent concepts seems, upon closer analysis, insufficient.

Hesse's quip about all language being metaphoric also follows from this revised approach: all language use involves grabbing meanings that present themselves as functionally appropriate for the communicative act at hand, and, while some constructions may challenge interpretation more than others, there is no clear reason to draw a definitive line between the literal and the figurative use of meaning. The ubiquity of metaphor takes on a more distinctly Peircean character, though, when word-objects are recognised as existing in the same cognitive space as other percepts. Peirce's claim that all thought is realised through signs [33] seems of a piece with Davidson's pragmatic approach to metaphor once the difference between considering objects as symbols of the mind versus considering symbols as objects of cognition becomes a relatively minor point of contention. To Peirce, reality was a lattice of ubiquitous signification, with meaning manifesting itself through a "life in signs", by which all thought results from the inherently interpretable interplay between things, and all physical interactions are characterised by this kind of life. The perpetual life cycle of event, perception, and interpretation means that signs are always exploding outward from the thing that they signify, becoming themselves the object of a further signification in the instant of their interpretation, even as the interpretation becomes a sign of the thing it interpreted. This endless sequence of becoming something else, accomplished by means of the transformative faculties of symbols, points to a fundamental and enduring process of metaphor-making in the experience of existence.

And here consciousness re-enters the consideration of metaphoric language: consciousness as the thing that can only be objectively grasped through metaphor, or metaphor as the mechanism that facilitates the subjective experience of consciousness. By Peirce's account, the world is conscious, an audacious asseveration that nonetheless lines up well with the idea that being conscious involves the perpetual invocation of the fundamental metaphor that everything else is conscious, as well. If the Peircean variety of panpsychism is perhaps a bit strong, a consideration of the metaphoric nature of individual consciousness at least offers an explanation of why the rest of reality would seem that way, as well. In fact, in accepting that language is wrapped up in a pragmatic process of meaning-grasping, and that all use of word-objects is essentially a ready-to-hand encounter with linguistic percepts, the experience of perpetual metaphor and therefore of imminent and ubiquitous consciousness becomes a less alarming outcome.

4 Meaning Is Geometric

Computational models of metaphor have tended to embrace symbolic approaches that treat language as a representation of cognitive content. As a first approximation, this is not unreasonable, given that computers are symbol manipulating machines: a semantic formalism is precisely the kind of information processing model that is tractable to such a machine. Consequently, van Genabith has found success programming computers to analyse metaphors using type theoretical constructs where source and target both belong to a supertype by virtue of their joint properties, and the intensions transferred by the metaphor are categories specific to the type of the source [38]. Similarly, Veale has built a computational system that handles analogies in terms of "dynamic types" culled from prefabricated conceptual networks such as WordNet [39]. And Gargett and Barnden have described a metaphor generating system that applies information processing instructions to conceptual structures [21], in this case implemented through the contextually sensitive typed schema of Feldman's embodied construction grammar [18].

These kinds of systems treat words as indices to concepts, where the logical structure of concepts can be aligned so as to indicate the affiliated linguistic expression that conveys the projection of properties from source to target. In this way, they are implementations of the conceptual approach to metaphor outlined by Lakoff and Johnson: their success stems from their recourse to abstract representations of concepts, and language is treated as a kind map of the mappings inherent in the dynamics of the conceptual space, metaphoric precisely because of the analogical aspect of cognitive content. In the case of Veale's system, the conceptual schema are, compellingly, built in an *ad hoc* way, even if this ongoing construction is based on a pre-established network. With Gargett and Barnden's system, the underlying formalisms are specifically designed to contextualise conceptual representations in terms of the physical world. By the same token, though, these models are intrinsically committed to the cognitive-content approach to metaphor, treating language as a secondary feature merely pointing to the world model of a conceptual space.

It is not clear how such a system could, for instance, model the direct imagistic experience of perceiving an aggressive person as a bulldozer. The inescapable figurativeness of consciousness, that property by which there is a bulldozer-like quale in the encounter with this unpleasant individual, is lost to a system that depends on conceptual constructs removed from encounters with the percepts – the language and the imagery – that become the symbolic index to those concepts. If the project of computationally modelling metaphor is to be pursued further, it seems necessary to formulate a way in which a space of meanings can be constructed directly from an encounter with language in the world, based on the actual statistical features of the language rather than on predetermined rules regarding the processing of symbols. But how can a computer go about realising this kind of language model?

In fact, symbol manipulating machines seem like exactly the right tools for engaging with this task, and a viable methodology already exists in the form of ongoing work on vector space models of distributional semantics. This approach to language modelling involves the geometric representation of words as points in a high dimensional space [40]. Words are construed as vectors, with the dimensions of these vectors corresponding to the contexts in which a word is likely to occur: in the most straightforward implementation, a dimension of a word-vector corresponds to a term, and the scalar value of that dimension indicates the likelihood of the word co-occurring with

that particular term. When the co-occurrences of the words found throughout a large scale corpus are computed, the result is a space in which the proximity of word-vectors to one another corresponds to the similarity of the contexts in which those words have been found. The intuition behind work in this direction has been that words that are found in a similar context will naturally be likewise semantically similar [37].

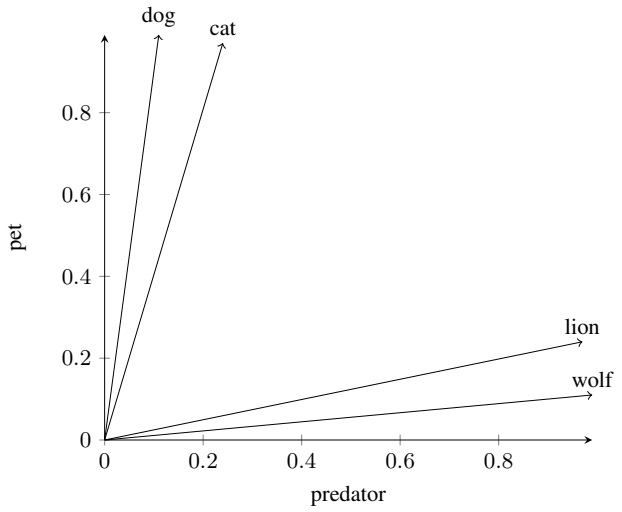


Figure 1: In a highly simplified (two dimensional) vector space model, the words "dog" and "cat" are seen to be semantically relatively similar by virtue of their frequent co-occurrence with the term "pet", whereas "wolf" and "lion" are more likely to occur in the context of the term "predator".

Furthermore, the mathematically tractable properties of a geometric space have been exploited in the modelling of compositionality, with linear algebraic operations between word-vectors producing statistical structures corresponding to the meaning of larger segments of language [5, 30]. A similar intuition can be applied to the construction of metaphor, though with the philosophical caveat, informed by Davidson's take on metaphor, that, where meaning applies to the space of words, the compositions constructed from this space are properly understood only in terms of their use in acts of communication. All the same, it is the geometry of the space of words that suggests ways in which sets of meanings can be pragmatically constructed as metaphors: if proximity corresponds to similarity, then regional clusters of related terms should be discoverable within the vector space. Moreover, the relationship between the terms within such a space indicates a particular geometry, and a congruence in the configuration of terms between two regions might be interpreted as an indication of a potential metaphor. So, for instance, the constellation of word-vectors indicated by the sequence $\{surgeon - patient - hospital - scalpel\}$ would be expected to line up with the shape described by $\{butcher - animal - abattoir - cleaver\}$.

Underwriting this statistical prediction is the theoretical intuition that the way in which a computer encounters symbols in a corpus stands in a synecdochical relationship to the way in which a cognitive agent encounters percepts – including linguistic symbols – in an environment. The hope is that treating large scale corpora as a kind of native habitat for computers serves as a more veridical simulation of the process by which cognitive agents directly grapple with language in the physical world than does the construction of abstract conceptual representations. Just as an agent maintains a shifting lexicon of meaning based on a continuous entanglement with language percepts, a computer can establish a network of relationships based on the statistics of its ongoing encounter with symbols in a textual environment. The statistical relationship of words learned by a corpus traversing computer becomes its knowledge base, its space of meanings that can be invoked in a disengaged way when the definition of a particular term is sought, but that at the same time stand ready-to-hand waiting to be grabbed as affordances in the construction of communicatively effective language. When the moment comes for the system to compose an expression, it grasps for the combination of terms that fulfil the required criteria, and these criteria are specifically modelled in terms of the geometric alignment of regions within the space of meanings.

Some preliminary work has been done exploring the relationship between established conceptual metaphors construed in terms of the arrangement of their components within a vector space model, comparing, for instance, the region of butchery to the region of surgery, or the region of orientation (e.g. $\{up - down - in - out\}$) to the region of emotion (correspondingly $\{happy - sad - inclusion - loneliness\}$). Early results have invited cautious optimism: the geometry of the compared vector regions has remitted a high degree of congruence in the anticipated alignments. Future research will have to examine the way in which regions of vectors, corresponding to the construct of conceptual spaces [20], can be defined within a vector space, and this direction of inquiry will in all likelihood motivate a close consideration of the techniques employed in the construction of the vector space itself, as well. The prospective outcome of this project is a system that will use corpus analysis to facilitate a program outputting novel and useful metaphors based on inputs that are perceived as being relatively literal.

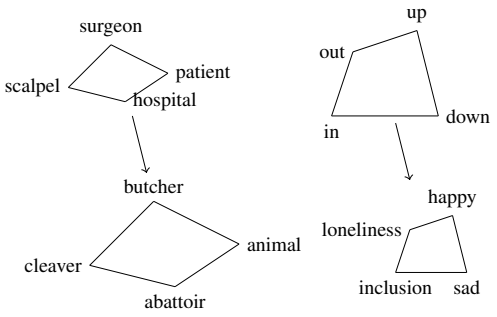


Figure 2: Congruences discovered in subregions of a vector space model suggest metaphoric mappings. The regions do not necessarily have to be of the same scale in order to identify a possible alignment.

5 Conclusion

In weighing the merits of considering the use of words as distinct from the meaning of words, it is worthwhile to observe the extreme ease with which people produce and digest figurative language: metaphor is so universal that almost nothing makes sense if it is taken absolutely at face value. Such a linguistic environment might appear particularly hostile to so formal and literal an agent as a computer. It would seem that the relationship between language and the situations described by language is much messier than some semantic

formalisms would suggest, and the role that meaning plays in the process of communication cannot be easily situated in a denotational relationship to some sort of mental content outside of language. In order for a computer to have a chance in a scenario where all language is open to interpretation, it is necessary for the information processing system to have recourse to its own semantic constructs, and these naturally take the form of statistical interpretations of the bearing of words in their compositional contexts.

Using a computer to model the pragmatic dynamics of metaphor reveals nothing about how consciousness works or why consciousness exists. In this regard, the most that can be said about the system described in this paper is that it attempts to simulate a process with which consciousness is concerned—and this much is true of any computer program that presents data in a way that is designed to be interpretable to a conscious user. Nonetheless, the project of constructing a metaphorical framework within a symbol manipulating system takes on added resonance when considered in the scope of the ineluctably analogical modality of the understanding of the conscious mind. Even if the model that has just been proposed doesn't shed any light on the nature of consciousness, it does address some of the questions about the linguistic operation involved in conceptualising consciousness. It is the very ineffability of consciousness that forces a philosopher to resort to analogy and metaphor when discussing this hard topic and indeed when describing the experience of it. In understanding the construction of metaphor as a utilisation of meaning towards the goal of expression, it becomes clear how a cognitive agent must be constantly involved in this operation, always grasping for the combination of meanings that work when put out into the world as the communication of a mental experience. In the process of constructing the sustained sense of self at the core of a conscious experience of the world, a cognitive agent must necessarily cast the idea of the self out into the world to reflect upon it; it is only natural, then, that an essential feature of consciousness should be to imagine that other things are conscious.

So this pragmatic reconsideration of metaphor and the computational implementation of the redesigned model offer at least the beginning of an explanation for the mind's propensity to figuratively project its own consciousness onto the entities that it encounters in the world. This final observation regarding the relationship between metaphor and consciousness can be turned into a possible stance in the debate regarding the controversial construct that reimagines the mind as a computer: if anything, it is the mind that projects consciousness onto the computer, not the computer that stands in as a model for what the mind does. The conceit of the mind as computer seems to easily forget that the operations of a computer are only meaningful by virtue of the values assigned to its inputs and outputs by some agent who is plugged into reality in a deeply intentional way—but then the mysteriousness of consciousness likewise evades the question of what exactly it is that is doing the conscious sensing, leaving only the fanciful notion that all nature of other things can consciously sense, as well. And so in the end, the metaphor of the mind as a computer is perhaps actually just a reversal of the metaphor of a computer as a kind of mind, a lending out of the self which is actually just a specific case of what conscious minds, in their incessant and incurable projecting, do to everything in the world.

ACKNOWLEDGEMENTS

This research has been supported by EPSRC grant EP/L50483X/1.

REFERENCES

[1] Aristotle, *The Poetics*, Macmillan and Co, London, 1895.
[2] Merrie Bergmann, 'Metaphorical assertions', *The Philosophical Review*, **91**, 229–245, (1982).
[3] Max Black, 'Metaphor', in *Proceedings of the Aristotelian Society*, volume 55, pp. 273–294, (1955).
[4] Max Black, 'More about metaphor', in *Metaphor and Thought*, ed., Andrew Ortony, 19–41, Cambridge University Press, 2nd edn., (1977).
[5] William Blacoe and Mirella Lapata, 'A comparison of vector-based representations for semantic composition', in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 546–556, (2012).
[6] Peter Carruthers, *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford University Press, 2011.
[7] Robyn Carston, 'Metaphor: Ad hoc concepts, literal meaning and mental images', in *Proceedings of the Aristotelian Society*, volume CX, pp. 297–323, (2010).
[8] Robyn Carston, 'Metaphor and the literal/nonliteral distinction', in *The Cambridge Handbook of Pragmatics*, eds., K. Allan and K. M. Jaszczolt, Cambridge University Press, (2012).
[9] David J. Chalmers, *The Conscious Mind*, Oxford University Press, 1996.
[10] Andy Clark, 'Language, embodiment, and the cognitive niche', *Trends in Cognitive Sciences*, **10**(8), (2006).
[11] Donald Davidson, 'On the very idea of a conceptual scheme', in *Proceedings and Addresses of the American Philosophical Association*, volume 47, pp. 5–20, (1974).
[12] Donald Davidson, *Inquiries into Truth and Interpretation*, chapter What Metaphors Mean, Clarendon Press, Oxford, 2nd edn., 1978.
[13] Donald Davidson, 'A coherence theory of truth and knowledge', in *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, ed., Ernest LePore, (1986).
[14] Daniel C. Dennett, *Consciousness Explained*, The Penguin Press, London, 1991.
[15] Daniel C. Dennett, *Intuition Pumps and Other Tools for Thinking*, W. W. Norton and Company, 2013.
[16] Fred I. Dretske, *Knowledge and the Flow of Information*, CSLI Publications, 1981.
[17] Daniel L. Everett, 'Cultural constraints on grammar and cognition in pirahã: Another look at the design features of human language', *Current Anthropology*, **46**(4), 621–646, (2005).
[18] Jerome Feldman, 'Embodied language, best-fit analysis, and formal compositionality', *Physics of Life Reviews*, **7**(4), 385–410, (2010).
[19] Luciano Floridi, *The Philosophy of Information*, Oxford University Press, 2011.
[20] Peter Gärdenfors, *Conceptual Space: The Geometry of Thought*, The MIT Press, Cambridge, MA, 2000.
[21] Andrew Gargett and John Barnden, 'Gen-meta: Generating metaphors using a combination of ai reasoning and corpus-based modeling of formulaic expressions', in *Proceedings of TAAI 2013*, (2013).
[22] Raymond W. Gibbs, Jr., *The Poetics of Mind*, Cambridge University Press, 1994.
[23] James J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Miffline, Boston, 1979.
[24] Mary Hesse, 'The cognitive claims of metaphor', *The Journal of Speculative Philosophy*, **2**(1), 1–16, (1988).
[25] Mary B. Hesse, *Models and Analogies in Science*, Sheed and Ward, New York, 1963.
[26] Julian Jaynes, *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, Penguin Books, 1976.
[27] Paul Kay and Luisa Maffi, 'Color appearances and the emergence and evolution of basic color lexicons', *American Anthropologist*, **101**(4), 743–760, (1999).
[28] George Lakoff and Mark Johnson, *Metaphors We Live By*, University of Chicago Press, 1980.
[29] Stephen C. Levinson, 'Yélî dnye and the theory of basic color terms', *Journal of Linguistic Anthropology*, **10**(1), 3–55, (2001).
[30] Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver, 'Evaluating neural word representations in tensor-based compositional settings', in *Proceedings of EMNLP 2014*, (2014).
[31] Andrew Ortony, 'Why metaphors are necessary and not just nice', *Educational Theory*, **25**(1), 45–53, (1975).

[32] *Metaphor and Thought*, ed., Andrew Ortony, Cambridge University Press, 2nd edn., 1993.
[33] Charles Sanders Peirce, *Peirce on Signs*, The University of North Carolina Press, Chapel Hill, NC, 1991.
[34] Richard Rorty, 'Unfamiliar noises: Hesse and davidson on metaphor', *Proceedings of the Aristotelian Society*, **61**, 283–296, (1987).
[35] John R. Searle, 'Metaphor', in *Metaphor and Thought*, ed., Andrew Ortony, Cambridge University Press, (1979).
[36] Alfred Tarski, *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, Hackett, Indianapolis, IN, 2nd edn., 1983. Translated by J. H. Woodger.
[37] Peter D. Turney and Patrick Patel, 'From frequency to meaning: Vector space models of semantics', *Journal of Artificial Intelligence Research*, **37**, 141–188, (2010).
[38] Josef van Genabith, 'Metaphors, logic and type theory', *Metaphor and Symbol*, (2001).
[39] Tony Veale, 'Dynamic type creation in metaphor interpretation and analogical reasoning: A case-study with wordnet', in *Conceptual Structures for Knowledge Creation and Communication*, 146–159, Springer, (2003).
[40] Dominic Widdows, *Geometry and Meaning*, CSLI Publications, Stanford, CA, 2004.

A Formal Model of Metaphor in Frame Semantics

Vasil Penchev[1]

Abstract. A formal model of metaphor is introduced. It models metaphor, first, as an interaction of "frames" according to the frame semantics, and then, as a wave function in Hilbert space. The practical way for a probability distribution and a corresponding wave function to be assigned to a given metaphor in a given language is considered. A series of formal definitions is deduced from this for: "representation", "reality", "language", "ontology", etc. All are based on Hilbert space. A few statements about a quantum computer are implied: The so-defined reality is inherent and internal to it. It can report a result only "metaphorically". It will demolish transmitting the result "literally", i.e. absolutely exactly. A new and different formal definition of metaphor is introduced as a few entangled wave functions corresponding to different "signs" in different language formally defined as above. The change of frames as the change from the one to the other formal definition of metaphor is interpreted as a formal definition of thought. Four areas of cognition are unified as different but isomorphic interpretations of the mathematical model based on Hilbert space. These are: quantum mechanics, frame semantics, formal semantics by means of quantum computer, and the theory of metaphor in linguistics.

1 INTRODUCTION

The thesis of the paper is fourfold: (1) Metaphor can be seen as the interaction of at least two frames in a sense of frame semantics. (2) Then representation can be interpreted as the particular case of zero interaction between the frames. (3) In turn, this allows of the frames to be interpreted formally as correspondingly "reality" and the "image of reality", and language as an (even one-to-one) mapping between those two universal and formal frames of "reality" and its "image". (4) Metaphor can be further represented formally as the "entanglement"[2] of two or more frames and thus in terms of quantum information.

That thesis has advantage (or disadvantage from another viewpoint) to be self-referential and paradoxical: Indeed the so-defined concept of metaphor is in turn the interaction between two frames: both that of frame semantics and that of formal semantics and consequently it would be "only" a metaphor if the frame semantics and formal semantics can interact as this text advocates; and vice versa: if any scientific notion is expected to be a representation of reality, this text should be zero-content for the set of its extension should be empty. Nevertheless, that explicit paradox is rather an advantage as the analogical paradox generates the development of language and thus perhaps this text as a live part of it.

The mathematical formalism of quantum mechanics (the so-called quantum mathematics) can serve for a formal theory of metaphor and thus for a serious technical formulation applicable to AI. However, the demonstration of the latter is absolutely impossible in the volume of the present paper. Its purpose is restricted only to *outlining the possibility* of a "quantum theory of metaphor".

That "quantum theory of metaphor" can be defined as that mathematical model of metaphor, which is based on Hilbert space very well utilized already by quantum mechanics.

Thus the suggested "quantum theory of metaphor" would share a common mathematical formalism with quantum mechanics. If that is the case, the representation of metaphor in terms of quantum mechanics is neither merely a loose analogy nor any metaphor of "metaphor", but rather a mapping between two different interpretations of the underlying model of Hilbert space.

Furthermore, the notions, approaches and even visualisations of quantum mechanics are exceptionally well developed in detail. They allow of that theory of metaphor called quantum to be represented immediately by a complete language including both mathematical model and huge practical area such as quantum mechanics.

Some of the most essential concepts of quantum mechanics necessary also to that theory of metaphor are "entanglement", "quantum information", and "quantum computer" defined bellow. Besides them, still a few terms need some specification, namely: "frame semantics", "frame" "formal semantics":

"Frame semantics" is meant in the sense of Charles J. Fillmore: "Frame semantics offers a particular way of looking at word meanings, as well as a way of characterizing principles for creating new words and phrases, for adding new meanings to words, and for assembling the meanings of elements in a text into the total meaning of the text" [1].

"Frame": "The idea is that people have in memory an inventory of schemata for structuring, classifying and interpreting experiences, and that they have various ways of accessing these schemata and various procedures for performing operations on them" [2]. "By the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits ..." [1]. The "frame" already linked to formal semantics is specified as a set of well-orderings referring to something as its "logic", in which any property, relation, part or feature of that something can be understood by somebody or by some group. Consequently, that formal and semantic "frame" means the relation between the wholeness of that something and the "logic" of it as a collection of well-orderings.

[1] Dept. of Logical Systems and Models, Institute for the Study of Societies and Knowledge at the Bulgarian Academy of Sciences, Sofia, Bulgaria, email: vasildinev@gmail.com.

[2] Entanglement can be interpreted as a kind of interaction due to wholeness: If two or more entities constitute a common system, they can interact with each other by the whole of the system itself, i.e. holistically, rather than only by some deterministic and unambiguous mechanism.

"Formal semantics" is a term used both in logic and in linguistics[3] but in partially different meanings [3]. The common is the utilization of mathematical and logical models. However, the logical "formal semantics" addresses the natural entailment in language in terms of logical sequence while the linguistic "formal semantics" discusses rather the correspondence both of linguistic units and the wholeness of texts to reality in terms of mathematical mappings, set theory, and logic [4, 5]. These meanings will be "entangled" in this paper by the mathematical concept of well-ordering, which can refer both to any logical sequence, and thus to any entailment in language, and to set theory including the axiom of choice, and thus to any one-to-one mapping of language and reality, such as a presentation.

"Entanglement" is a term in quantum mechanics, meaning the information interaction between two or more quantum systems and thus being fundamental for the theory of quantum information. However, the formal and mathematical definition of "entanglement" as that Hilbert space[4], which cannot be factorized to any tensor product of the Hilbert spaces of subsystems, allows of the term to be generalized to any model utilizing Hilbert spaces. For the formal and semantic model used here is based on Hilbert space(s), the concept of entanglement is applicable. It is the mathematical base for the model of metaphor.

"Quantum information" is a term initially coined by quantum mechanics to describe the base of a generalized kind of information underlying all quantum mechanics. So, quantum information can be interpreted as both transfinite series of bits and finite or infinite series of qubits. A bit is the elementary choice between two equally probable alternatives, and a qubit (i.e. quantum bit) can be interpreted as the elementary choice among an infinite set of alternatives though it is initially defined in quantum mechanics as the normed superposition of two orthogonal subspaces of Hilbert space. The quantity of information whether classical or quantum is the quantity of the corresponding elementary choices (whether bits or qubits) necessary for transforming a well-ordering to another (both, whether finite or transfinite). Thus quantum information can be interpreted as the quantity of elementary choices necessary to transform a frame into another and consequently the information of any metaphor formalized as above.

"Quantum computer" [7, 8, 9] is a mathematical model involved by quantum mechanics to interpret its formalism as a generalized kind of calculation, processing quantum information. Thus all physical states and processes may be also seen as computational.

The advantages of the suggested theory of metaphor would be the following:

It relies on a developed and utilized model though in a rather different scientific area.

It can be applied practically as this is sketched (only roughly) in Section 2.

It would aid the formal reconstruction of semantic interactions as a whole as well their historical change by investigating the correlations in the uses in texts and discourses.

It allows of far reaching unifications, generalizations, and philosophical conclusions.

A section (6) is devoted to the unity of thesis as a single, coherent and contextual whole consisting of the distinguished parts (namely the four "folds" of the fourfold thesis above). The mathematical model lent by quantum mechanics is the common base.

Nevertheless some ideas can be considered in their own right even out of the model, e.g. representation as a particular, borderline and limiting case of metaphor.

However this seems to be impossible as to others, e.g. the converse relation of model and reality, proposed near the end of Section 4. Those are logical corollaries from the utilized model.

The argumentation for the thesis has four corresponding points:

(1) Metaphor can be understood as the appearance of a new frame by interaction of two or more initial frames for some essential part of each of them is shared by all. Thus the understanding of each of them separately generates immediately the understanding of the metaphor as a new whole [10, 11] demonstrating therefore the appearance of a new frame, which is not the simple additivity of the sub-frames composing it. The set of well-orderings formalizing semantically a frame can be substituted by a point of Hilbert space [12], and interpreted as a wave function[5] of a quantum system [13]. Any possible frame is measurable as a single value of quantum information. Then the metaphor will be interpretable as the entanglement of the quantum systems corresponding to each sub-frame composing it.

(2) Representation can be interpreted after that as a particular and borderline case of metaphor, a "zero" metaphor, or just as the simple additivity of the sub-frames composing it. The corresponding wave functions are orthogonal to each other and there is no entanglement between them.

(3) Language is reduced to an infinite countable set (A) of its units of meaning, either words or propositions, or whatever others [14]. It includes all possible meanings, which can be ever expressed in the language rather than the existing till now, which would always a finite set. The external twin of reality is introduced by another set (B) such that its intersection with the above set of language to be empty. The union of them ($C=A \cup B$) exists always so that a one-to-one mapping ($f: C \leftrightarrow A$) should exist under the condition of the axiom of choice. The mapping (f) produces an image (B (f)) of the latter set (B) within the former set (A). That image (B (f)) serves as the other twin of reality to model the reality within the language as the exact representation [15] of the reality out of language (modelled as the set B). In the model, the necessity and sufficient condition of that representation between reality both within and out of the language is just the axiom of choice: If the axiom of choice does not hold, the relation between the sets B (f) and B cannot be defined rigorously as an exact representation but rather as some

[3] Some authors doubt the relevance of formal semantics to natural languages [6].

[4] The complex Hilbert space is the fundamental mathematical structure underling quantum mechanics. It is a vector space defined over the field of complex numbers. Hilbert space can be thought as the infinitely dimensional generalization of the usual three-dimensional Euclidean space where furthermore the real numbers are replaced by complex ones. Just the complex Hilbert space is meant for "Hilbert space" in the paper. It allows of: arithmetic and geometry to be generalized and thus unified into a single structure; the possible and actual to be not more than different interpretations of a single mathematical structure.

[5] The term "wave function" is used bellow without quotation marks also a synonym of an element of the complex Hilbert space. Exactly speaking, the former is the common interpretation of that element in quantum mechanics.

simile and the vehicle between the two twins of reality can be only metaphor[6].

(4) Metaphor formalized as above is representable as the wave function of the frame compounded by two or more sub-frames, which interact between each other by means of the shared nonzero intersection. The quantity of quantum information of a metaphor is different from that quantity of the corresponding representation. Thus the metaphor demonstrates the entanglement of the composing sub-frames after they have been formalized as points in Hilbert space [16].

The intuitive sense for metaphor to be represented as the entanglement of its terms is the following. The meaning of any term in a metaphor influences the meanings of the rest.

Consequently, their meaning within the metaphor is essentially different from those of the terms by themselves.

Any mathematical model of metaphor needs a certain relevant quantity of that influence. Once that model involves Hilbert space(s), the entanglement and the corresponding quantity of quantum information are the most natural applicant for describing the degree of that influence.

However, the metaphor itself being already mathematically modelled serves to describe the degree of entanglement between different formal realities (or "languages") in Section 4 and Section 5. Then the formal concept of language is accordingly generalized from a simple representation of reality, i.e. its identical "twin", to a metaphorical image of both reality as a whole and its separate elements such as "things".

The paper is organized as follows. The sections from 2 to 5 argue for the four "folds" of the thesis: (1) to (4) above. Section 6 unites them into a single viewpoint. Section 7 presents the conclusions and provides directions for future work.

2 METAPHOR AS INTERACTION OF FRAMES

Metaphor can be seen as the interaction of two or more frames as follows. Any frame corresponds of some unit of meaning such as a word. The meaning is understood as a whole, i.e. all links between this unit and other units in the frame are actually given according to frame semantics. One can suppose language as the maximal frame containing all other frames as sub-frames. Anyway the most part of language remains absolutely or almost irrelevant to the understanding of any given term. The other, quite small part most relevant to the understanding can be used for its definition. Consequently, the understanding of a meaning can be thought as an exactly determined position in the maximal frame of language, in which the neighbour links are crucial, the next links are less crucial, and the significance of further links weaken very fast, but gradually, moving away from the position in question and converging to zero as to the most part of the language [17].

The same picture can be repeated for arbitrarily many meanings, and particularly for one more:

Let us figure that both meanings are simultaneously active and their joint understanding is supposed. If both meanings are neighbour or at least relevant in definition, this is rather a proposition than a metaphor. The link between them is explicit in the frame of each of them.

However that is not the case of a proper metaphor where the link connects two areas, each of which is relevant for the understanding of one term, but irrelevant for the other one.

Obviously, the transition between the compound frame of a proposition and that of a metaphor is gradual [10].

Metaphor can be seen as a generalization of proposition referring to remote meanings in the maximal frame of language. Proposition does not generate any radically new meaning irredundant to those of its parts. The meaning of a proposition can be called "analytical" in a *broad and linguistic sense*[7].

Any metaphor appeals to some implicit meaning relevant to the pathway frame between the connected ones. However, that pathway frame of a metaphor is not objective. It depends not only on the connected frame, but also on the person(s) who understand(s). The pathway and thus the implicit frame are not unambiguously determined: it includes also the personality and biography of who understands. The meaning of a metaphor can be called "synthetic" in a broad and linguistic sense:

One can utilize the picture of the maximal frame, in which are chosen two positions as two points. Furthermore, the proposition connects them by a single "classical trajectory" while, the metaphor does the same by *all possible trajectories*, each of which is differently probable. Any understanding chooses only one of them. The mapping analogy to the Feynman interpretation[8] of quantum mechanics [18, 19, 20, and 21] is obvious. It addresses further the idea for the mathematical formalism of quantum mechanics to be only adapted to the relevant terms of frame semantics:

Indeed any measurement in quantum mechanics corresponds to a given understanding of what the metaphor mean. The metaphor unlike any proposition does not predetermine how it should be understood, however it defines implicitly a wave function of all possible understandings as the set of pathways, in any of which it can be interpreted equally justifiably.

Entanglement and the Feynman interpretation are both deduced from the mathematical formalism, but historically independent of each other. Nevertheless, there exists the following rigorous logical link between them:

The Feynman interpretation implies entanglement:

[6] The axiom of choice is independent of the other axioms of set theory in the usual systems of its axioms. The former case corresponds to the systems with the axiom of choice, the latter without it. However in fact, the utilized model of Hilbert space is invariant to it without being independent of it in a sense: Quantum mechanics uses Hilbert space both with and without the axiom of choice in two interpretations of quantum mechanics, which identify to each other and anyway distinguish from each other. This is rather a special and inherent property of Hilbert space than an accidental one brought in by quantum mechanics for interpretation.

[7] That "broad and linguistic sense" means that the proposition is a series, the elements of which are ordered in a whole. Anyway this is not the rigorous formal and logical deduction, which is analytical in a *narrow sense* for the premise implies the conclusion necessarily. The analyticity of a proposition is pragmatic and due to the possibility and probability of a rather expected link being usual and more or less often used. Metaphor is rather unexpected and nevertheless understandable.

[8] The essence is any motion or change to be generalized as done in infinitely many paths simultaneously rather than in a single one. The metaphor can be thought in the same way as the motion from a term to another or others in "many paths", each of which is an interpretation of the metaphor in questions and can be realized by somebody.

Indeed any "path" between two or more quantum entities means that they share at least one of their own possible states as common. And vice versa: if there is not entanglement, the Feynman interpretation would be impossible for this means that the entities are orthogonal to each other and thus they are not able to share any common states.

Furthermore, the exact mathematical formalism, which the Feynman interpretation implies, considers Hilbert space only as an approximation or as a limit after *infinitely* many "paths". In fact, that approximation and thus the nonzero difference between Hilbert space and the proper formalism of that kind are inherently necessary for that interpretation because this allows of entanglement to "sneak" implicitly into it.

Consequently, the Feynman interpretation is a stronger statement than the standard mathematical formulation about single, independent and thus non-entangled Hilbert spaces, which are all equivalent to a single Hilbert space[9].

Once the Feynman interpretation is involved for the mathematical model of metaphor as above, this implies immediately that entanglement is also though implicitly introduced and should be discussed in the framework of that model.

The Feynman interpretation further means that if it is universal, all quantum systems are entangled, and the standard consideration of quantum mechanics by single and non-entangled Hilbert space is not more than a working idealization and simplification.

That states of affairs in quantum mechanics can be forthwith interpreted in terms of the utilized model of metaphor: Representation is not more than a working idealization and simplification of metaphor: one statement, which will be discussed in detail in the next section.

The situation of two terms can be continued to more than two, even to arbitrarily many, and one is able even to consider the case of the metaphor of metaphors [22] as well that of the "proposition of metaphors". The method for that continuation is the relevant interpretation in terms of quantum mechanics in order to be borrowed the very well developed mathematical model.

Practically, one needs some relevant, reliable, and relatively unambiguous method for any given metaphor in a given language with its use and history to be adequately determined its wave function. This method can involve the following stages:

1. Determining a broad set of associative series, which can connect the terms of the investigated metaphor.

2. Structuring this set as a directed graph [23].

3. Determining the combinatory frequency of each vertex in the entire dictionary of the language or in any as contemporary as historical sub-dictionary if need be.

4. Calculating the frequency and probability in any possible pathway in the graph.

5. Summarizing these data as a probability distribution.

6. Approximating this probability distribution [24] by a wave function.

7. Eventually interpreting and modelling this wave function as a state of a quantum system and thus of a quantum computer.

Only stage 1 depends crucially on the human creativity to be figured all thinkable and unthinkable associative series, which can connect the terms of a metaphor. All rest stages can be accommodated for relevant software.

However, ever this first stage might be replaced by a formal frequency use analysis of common terms in the frames of all terms constituting a given metaphor. One should consider those frames as frequency use in the context of a given term and consisting of two, three, four and so on words. Consequently, the following stages «1′» and «1″» can substitute the above «1»:

1': Formally determining the frame of each term constituting the given metaphor as frequency uses of two, three, four, five, and so on words, containing the term in question.

1": Determining the frequency use of common terms in the frames of the terms of the investigated metaphor.

Those stages can be quite roughly illustrated by an imaginary example for their application about a real metaphor, e.g. "The moon is sad".

First of all, this is an obvious metaphor, which connects a celestial body, which is impossible to be sad, with a human mood, that to be sad: Who is sad cannot be anything inanimate such as the moon.

Furthermore, "Google" shows that the exact phrase as above is used in 59,000 web sources (retrieved on 14.03.2015). Nevertheless, the phrase is found in no case in the huge data base of English literature in "Ngram Viewer" of "Google books" (again then). Consequently, this is a real contemporary metaphor rather than a "white metaphor" coining Derrida's metaphor about any too used metaphor.

There are at least two different practical methods, which would give also different results perhaps, to be determined the paths and their corresponding probabilities for the latter term, "sad", to be reached starting from the former term, "moon".

The one method would construct the frames of both terms by means of main frequency uses of small contexts containing the terms and would search for coincidences of terms belonging to both frames.

One can figure as an imaginary example that the pair (moon, round) has frequency use "f_1" and probability "p_1" calculable as the ratio of "f_1" to the number of all considered frequency uses in the frame of "moon". Furthermore, the triple (round, face, sad) is analogically featured by "f_2" and "p_2" in the frame of "sad". "Round" is the searched coincidence. It allows of constructing some relevant function "P_1 (f_1, p_1, f_2, p_2)", which would suggest a value of the composed path (moon, round, face, sad) connecting both terms of the metaphor in a possible way.

The other method would consider only the frequency uses of those pairs, the series of which starts from "moon" and finish to "sad".

In the above example, those would be: (moon, round), f_3, p_3; (round, face), f_4, p_4; (face, sad), f_5, p_5. They would imply some P_2 (f_3, p_3, f_4, p_4, f_5, p_5) of the same path however calculated by the latter method.

[9] However one has to mean that any quantum system referring to a single Hilbert space can be always exactly and equivalently represented as consisting of two or more entangled subsystems and correspondingly Hilbert spaces. Then the viewpoint of the system differs from that of any subsystem. The Feynman interpretation is a way the viewpoint of the quantum whole to be represented as a certain function (namely its wave function) of the viewpoints of its virtual classical "parts", each of which is featured by a single classical "path". The suggested model of metaphor being considered as a whole would consist of the virtual parts of its interpretations, any of which is featured by its own proper associative path and a corresponding probability of this path calculable by relevant frequency uses.

If that procedure either in the former or in the latter method is repeated as to many enough paths, one can yield the probability distribution, which refers to the metaphor "The moon is sad" in English, with any preliminarily defined exactness. Then, the characteristic function of that probability distribution will represent the searched wave function of the metaphor in question.

The above two methods can be further modified and mixed in different proportions. However, they reflect two different ways for the model of metaphor to be understood: either as the entanglement of the frames of terms constituting a given metaphor or as a single frame of the metaphor as a whole, which is practically reduced to a set of series corresponding to paths between the terms of the metaphor.

Anyway the goal of the paper is only the possibility in principle as well as a schematic diagram of how the metaphors first interpreted in terms of frame semantics to be further modelled mathematically and then computationally.

3 REPRESENTATION AS A PARTICULAR CASE OF METAPHOR

The next step refers to representation: How the representation to be grounded on metaphor? The usual way is the reversed: How the metaphor to be founded by representation, which is granted as a self-obvious base?

However, the above mapping to quantum mechanics leads just to the metaphor to be the starting point. The end point is not so the representation by itself, but the concept of reality to be obtained in a formal and mathematical way [25] in order to be modeled.

The representation can be considered as a particular and borderline case of metaphor following the method for quantum mechanics to be reduced to classical mechanics by the principle of correspondence.

The problem is the following. Some metaphor is given. Which are the boundary conditions, on which its wave function can be transformed into that of a corresponding representation? The wave function of a representation is degenerated in a way so that the corresponding probability distribution is reduced to a single infinite pick in a single point, i.e. to the Dirac δ-function.

That result for the probability distribution in all associative ways of the metaphor in question can be obtained so: the interval of nonzero probabilities converges to the limit of a single point.

The process of convergence requires both decreasing the associative "distance" between the connected terms of the metaphor (which are at least two) and increasing the extension of the generalization of the terms so that the set of all associative pathways to be able to be reduced gradually to a single one. If that is the convergence, the corresponding directed graph of the metaphor will degenerate to a directed segment and even to a directed segment of zero length. The latter in turn is equivalent to a bit of information [26]: the "cell" of the segment possesses two equally probable, but alternative state of each of the two ends.

This would correspond to the degenerated or "ontological" metaphor: '"A" is A' decodable as the dialectic judgment that both "A" is A, and "A" is not A. The two ends of the "zero segment" are: "A" and A (whatever A is).

The directed segment of zero length (or a bit) means an elementary choice as well as an identical mapping. If these concepts are applied to an infinite set, they require the axiom of choice and even a special case of invariance in relation to it. That invariance consists in this, any subset of any set not only to be able to be enumerated by virtue of the axiom of choice, but also the set and the enumerated image of it to be identified.

The mathematical model of representation deduced from the metaphor should include all aforesaid formal properties.

Let us now interpret these mathematical features of representation in terms of frame semantics, i.e. as an interaction between two frames, which relation can be even identical. That interaction is zero in both opposite cases: both where the frames are absolutely independent of each other and where they coincide.

Even more, both cases can be identified by the above formal properties of representation as the "two ends of a directed segment of zero length" or as the "ontological metaphor": '"A" is A'.

Then the "class of all representations" can be defined as 'reality' in terms of the formal frame semantics. Reality can be deduced from representation, which in turn can be deduced from metaphor.

The formal and mathematical concept of reality is crucial for modeling any intellect able to be standalone. The demarcation line between a machine however "clever" and an intellect however "stupid" is just the concept of reality, which is inherent for the latter and somebody else's for the former. Thus the machine however "intelligent" remains a machine in somebody else's reality, e.g. a human being's.

Reality equivalent to the class of all representations is equivalent also to the aforesaid invariance to the axiom of choice for the class of all representations coincides with that invariance. However, it can be defined only on infinite sets.

Practically, this means that the formal concept of reality defined as above can be modeled only by some quantum system, i.e. on a quantum computer rather than on a Turing machine (i.e. on any standard computer independent of its power) always representing always a finite series after finishing effectively by any result .

A representation modeled on a quantum computer is a measurement of it. Any direct measurement means for a quantum computer to be irreversibly demolished, though:

This means that the superposition of all possible states, which is essential for its definition, is reduced to a single one, namely what is measured. Indeed the processing of a quantum computer consists in a reversible and smooth change of all elements of a set of probability distributions. Thus the statistical probabilities of the corresponding ensemble of measured results are changed as the output of that computer. However, the measurement of any state cancels irreversibly its work and it is destroyed in fact.

Consequently, the attempt to be modeled that formal concept of reality on a quantum computer fails for the set of representations, i.e. measurements are not infinite: even if the measurements are done of a collection of quantum computers. Furthermore, that collection is not only finite, but also a statistical ensemble rather than a coherent state.

One has to search for other, nondestructive ways for mappings of a coherent state into another or other of a quantum computer rather than into the elements of a statistical ensemble.

This requires the correspondence of reality and image to be first reformulated in a generalizing way allowing of the communication between them by means of entanglement.

4 HILBERT SPACE: REALITY AND ITS MAPPING WITHIN A QUANTUM COMPUTER

The next step refers to the formal concept of language again by means of Hilbert space [27, 28]. The goal of that step addresses reality to be generalized in way allowing of sharing reality not to lead to demolishing the quantum computer. The constraints and quantitative laws of that sharing are further problems.

Once reality is defined formally as a special set of mappings, one can continue generalizing to broader and broader sets of mappings. They can be also considered as "languages" mapping the so defined "reality" in different ways. Furthermore, each that language offers a different metaphor in general[10] for each "element of reality" being a representation. Then any collection of metaphors about those "elements of reality" is a language obviously defined already formally.

In other words, the language is defined as a particular set of primary (or "elementary") metaphors, in which at least one term is necessarily an "element of reality" while the others designate or define it. Two frames correspond to them in frame semantics being linked to each other by a wave function, i.e. by a point in Hilbert space according to the model introduced in section 2.

This means that any language should be consider as a state of the quantum field over reality. The term of "quantum field" is meant as usual in quantum mechanics, i.e. as a mapping of a set (the set of all representations, or "reality") into Hilbert space.

The "set of all possible states of the so-defined quantum field" including all possible languages will be designate as 'ontology'[11].

Consequently, the concept of ontology is implied much broader than that of reality. If any image of reality in any language is interpreted as another reality, then ontology is the class of all realities or of all possible worlds.

One can demonstrate that those formal concepts are able to be modelled entirely within Hilbert space in a quite natural way. Indeed "representation" corresponds to the relation of two coinciding elements of the two dual spaces. They are both identical and complementary.

Consequently, the so-defined formal concept of reality is inherent to Hilbert space. If Hilbert space is considered as a model shared e.g. by quantum mechanics, that reality is internal rather than external to it. It is complete to that reality.

The interrelation of model and reality (more exactly, the so-defined reality as a formal model) is rather extraordinary in comparison with classical physics, science, and epistemology, being "reversed" in a sense. Model contents the model of reality rather than reality contents the reality of model.

Then any language is a mapping of Hilbert space [29] into itself, and thus any physical quantity[12] is a language defined formally as above (but not vice versa).

Furthermore, Hilbert space can be considered as a quantum computer, and any point in it as a state of it. So that quantum computer should content reality in the sense of the above formal model of reality within itself being therefore standalone rather than a machine within somebody else's reality.

However, there is a considerable problem of how two or more different realities are able to communicate. Particularly, how is a quantum computer able to transfer a result to us without demolishing itself and thus destroying also that other reality within it and different from ours?

As we will see: only "metaphorically".

5 METAPHOR IN TERMS OF ENTANGLEMENT

The next step requires the relation of any two "languages" to be defined in terms of Hilbert space(s) therefore involving entanglement between them. The goal is: some nondestructive way for transmitting information between two or more realities identified as languages to be outlined. The way of measurement has already excluded above as destructive.

Let there are two different "metaphors" of one and the same "element of reality" in two languages, i.e. two wave functions. The "element of reality" can be excluded and any of the two metaphors can be directly referred to the language (reality) of the other. Those language and reality in the neighborhood of the metaphor are unambiguously defined by the corresponding wave function. Thus the metaphor will "seem" or "appear" as the entanglement of both wave functions from the viewpoint of each of the languages.

One can compare the formal definition of a metaphor in Section 2 as a single wave function with the present definition as the entanglement of two ones. Obviously, these definitions do not coincide: There are two different definitions of one and the same metaphor therefore each one needing some different, but relevant interpretation:

The metaphor defined as in Section 2 as a single wave function should be interpreted as that in the common system of the language or in the universal reality to the particular realities of each term.

The metaphor defined as here, in Section 5 as the entanglement of two or more wave functions should be interpreted as seen from the particular viewpoint of each term of it and thus in the corresponding particular reality.

However, that mismatch is just the nondestructive way for a quantum computer to transmit a result, as we see, only "metaphorically". The transfer is "less metaphorical", i.e. more precious, the quantum computer will be more influenced by the transfer, even demolished after any absolutely exact transmission of its result. The mismatch depends on the quantity of entanglement, in particular, on that of the quantum computer and our reality.

If one of the terms of the metaphor is permanent, e.g. anchored in our reality, the change of the others can be interpreted as the metaphorical "message" thus poetically [30, 31]. The quantum computer turns out to be a "poet".

Practically, the transmitted result will be a change of the rest frames to an anchored frame postulated as that of reality as to our reality. That change of a few frames being also a change of metaphor and an arbitrary[13] operator in Hilbert space can be defined as a single elementary thought [33].

Consequently, a quantum computer cannot report a result in a nondestructive way, but can communicate a thought just as a

[10] Particularly some metaphors in some languages can coincide.

[11] T. Giraud offers a fundamentally different ontological perspective [32].

[12] In the way as it is defined in quantum mechanics.

[13] That is neither self-adjoint, nor linear in general.

human being can. If the thought is clearer, the computer is more "obsessed" by it: i.e. its state and thus future work will be more influenced by its communication.

6 THE UNITY OF THESIS

One can deduce the following from summarizing Sections 2 – 5:

From 2: Metaphor can be represented as an interaction of frames in terms of frame semantics, and then modelled formal as a "wave function", i.e. as an element (point, vector) in Hilbert space.

From 3: Representation can be defined as a particular case of metaphor, namely as the directed segment between two coinciding frames with a corresponding probability distribution degenerated to a Dirac δ-function. The set of all representations is a formal definition of reality.

From 4: That reality turns out to be inherent and internal to Hilbert space and thus to any quantum computer. It can be also considered as identical to a formal concept of language. The class of all languages (or "realities") defines formally the concept of ontology.

From 5: A quantum computer can report a result only "metaphorically" or "poetically". The report is more precise, the quantum computer is more influenced; and even demolished in the borderline case of absolutely exact report. That report is a change of a metaphor to an anchored term and can be considered as a formal definition of thought.

Conclusion from 2 to 5: Any quantum computer being furthermore standalone and supplied by reality can think. Human thinking can be exhaustedly modelled by a quantum computer.

The unity of the thesis includes a few heterogeneous fields of cognition: quantum mechanics as a theory of nature, frame semantics as a theory of human thinking, the theory of metaphor and representation as a theory of language, quantum computer as a theory of artificial intellect. The four can share a common mathematical model based on Hilbert space(s). This allows of a uniform and even mathematical description both of thinking whether human or artificial and of states and process whether physical or linguistic. These four can be considered as not more than different interpretations of a single model and thus isomorphic to each other.

7 CONCLUSIONS & FUTURE WORK

This paper shows how one can use the concept of frame in frame semantics to define metaphor as an interaction of frames. The Feynman "many-paths" interpretation of quantum mechanics allows of the metaphor to be represented by a wave function and thus the mathematical model of Hilbert space to be involved.

One can demonstrate a general approach for any given metaphor in any given language to be assigned a relevant probability distribution and then a wave function. Though the approach is shown by the example of two terms, it can immediately extend to more than two terms following the pattern of quantum mechanics: any separate position in the Feynman model corresponds one-to-one with a term of the metaphor.

The formal model of metaphor implies that of representation as a particular and borderline case of the "ontological" metaphor '"A" is A', and the Dirac δ-function as the corresponding probability distribution. This allows of a formal definition of reality as the set of all representations. That reality is inherent and internal to Hilbert space. Thus any quantum computer turns out to be supplied by its inherent and internal reality. Its reality is what guarantees for it to be standalone rather than a machine in somebody else's reality. However, a quantum computer cannot report us any absolutely exact result without self-demolition.

One can define a formal concept of language within Hilbert space as the mapping of "reality", being internal to the Hilbert space, to the same Hilbert space. That mapping can be considered as a quantum field in the standard definition of quantum mechanics. However, it can be also interpreted as a language mapping any element of reality (signified) into another (signifier) by means of that metaphor (sign), the wave function of which is the value of the quantum field for this element of reality. Furthermore one can define 'ontology" as the "class of all languages" and therefore of all realities or all possible worlds.

This allows of another formal definition of metaphor as a compound "sign" (i.e. two or more entangled wave functions) consisting of two or more signs referring to different signifier in different languages, but of a single common signified.

That formal concept of language is a "quantum field" on "reality", i.e. as a mapping of the set of the formally defined reality in Hilbert space into the same Hilbert space. Any "element of reality" is a "signified" mapped by the "sign" of a metaphor (i.e. a wave function) into another (in general) "element of reality" as a "signifier". Any "language" is also interpreted as another and different "reality" again formally defined. 'Ontology' is further defined as the "class of all languages" and thus that of all realities.

The other, new, and different formal definition of metaphor is given as the relation between different signifiers of a single element of reality as a signified and therefore modeled by two or more entangled wave functions corresponding to the sign of each term in each language.

There will be two distinct definitions of one and the same metaphor: as a single wave function according to Section 2 and as a few entangled wave functions according to Section 5. The quantitative mismatch (being due to the entanglement) between the two definitions can be represented back in terms of frame semantics as a change of a frame to another, after which all rest terms will change their position to one anchored to that reality (language) chosen as a reference frame, e.g. ours.

That "frame change" being also a "metaphor change" can be defined as an 'elementary thought' [34].

Any quantum computer can transmit any result in a nondestructive way only "metaphorically" or "poetically" rather than literally, i.e. as an elementary thought. The thought transmits the result more exact, it is more "obsessive" for the computer: that is its state and thus reality is more influenced by the event of transmission. The borderline case of an absolutely exact report of the result is tantamount to its demolition.

One can also say that quantum computer thinks in this sense of transferring a message between realities (or languages) metaphorically. Furthermore, the essence of thought turns out to metaphorical and thus poetical in the frame of the present paper.

The unity of the thesis demonstrates that a single and common mathematical model based on Hilbert space can be shared by four scientific fields: quantum mechanics describing nature; frame semantics describing human cognition; linguistics

describing metaphor and representation; theory of quantum information describing quantum computer.

That unity implies the following five directions for future work. Four ones for each of the four fields enumerated above and still one, the fifth for their synthesis developing the underlying mathematical model.

REFERENCES

[1] C. J. Fillmore. Frame semantics. In: *Linguistics in the Morning Calm*. Linguistic Society of Korea (Ed.) Hanshin, Soeul, Korea: 111-137 (1982).

[2] C. J. Fillmore. Frame semantics and the nature of language. In: *Origins and Evolution of Language and Speech*. S. R. Harnad, H. D. Steklis, J. B. Lancaster (Eds.) (Annals of the NY Academy of Sciences, Vol. 280) New York Academy of Sciences, New York, USA: 20-32 (1976).

[3] J. Woleński. What is Formal in Formal Semantics? *Dialectica*, 58(3): 427–436 (2004).

[4] J. Woleñski. From Intentionality To Formal Semantics (From Twardowski To Tarski). *Erkenntnis* , 56(1): 9-27 (2002).

[5] R. Rogers. A survey of formal semantics. *Synthese*, 25 (1): 17-56 (1963).

[6] A. Galton. Formal semantics: is it relevant to artificial intelligence? *Artificial Intelligence Review*, 2(3): 151-165 (1988).

[7] R. P. Feynman. Simulating Physics with Computers. *International Journal of Theoretical Physics,* 21(6/7): 467-488 (1982).

[8] R. P. Feynman. Quantum Mechanical Computers. *Foundations of Physics*, 16(6): 507-531(1986).

[9] D. Deutsch. Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 400 (1818): 97-117 (1985).

[10] L. J. Cohen. The semantics of metaphor. In: *Knowledge and Language: Selected Essays of L. Jonathan Cohen*. J. Logue (Ed.) Dordrecht, Springer Netherlands: 27-40 (2002).

[11] F. Guenthner. On the semantics of metaphor. *Poetics*, 4(2-3): 199-220 (1975).

[12] A. Y. Khrennikov. *Ubiquitous Quantum Structure*. Springer, Berlin; Heidelberg (2010).

[13] D. Aerts and M. Czacho. Quantum aspects of semantic analysis and symbolic artificial intelligence. *Journal of Physics A: Mathematical and General*, 37 (12): 123-132.

[14] M. Jubien. Formal Semantics and the Existence of Sets. *Noûs*, 15(2): 165-176 (1981).

[15] J. H. Dreher. On the semantics of formal representation. *Philosophia: Philosophical quarterly of Israel*, 8(1): 71-78 (1978).

[16] D. Aerts and L. Gabora. A theory of concepts and their combinations I: The structure of the sets of contexts and properties. *Kybernetes*, 34(1&2): 151-175 (2005).

[17] C. Gauker. Contexts in Formal Semantics. *Philosophy Compass* 5(7): 568–578 (2010).

[18] R. P. Feynman. Space-Time Approach to Non-Relativistic
Quantum Mechanics. *Reviews of Modern Physics*, 20(2): 367-387 (1948).

[19] R. P. Feynman and F. L. Vernon, Jr. The theory of a general quantum system interacting with a linear dissipative system. *Annals of Physics* (N.Y), 24: 118-173 (1963).

[20] R. P. Feynman; A. R. Hibbs. *Quantum mechanics and path integrals*. New York: McGraw-Hill (1965).

[21] H. D. Zeh. Feynman's interpretation of quantum theory. *The European Physical Journal H*, 36(1): 63-74 (2011).

[22] M. Bergmann. Metaphor and formal semantic theory. *Poetics*, 8(1-2): 213-230 (1979).

[23] W. J. Hutchins. Semantics in three formal models of language. *Lingua*, 28(1): 201-236 (1972).

[24] P. Bruza, J. Busemeyer , and L. Gabora. Introduction to the special issue on quantum cognition. *Journal of Mathematical Psychology*, 53: 303-305 (2009).

[25] D. Pearce and V. Rantala. Realism and formal semantics. *Synthese,* 52(1): 39-53 (1982).

[26] L. V. Arshinskii. Substantial and Formal Deductions in Logics with Vector Semantics. *Automation and Remote Control*, 68(1): 139–148 (2007).

[27] A. Khrennivov. Classical and Quantum Mechanics on Information Spaces with Applications to Cognitive, Psychological, Social, and Anomalous Phenomena. *Foundations of Physics*, 29(7): 1065-1098 (1999).

[28] D. Aerts and L. Gabora. A state-context-property model of concepts and their combinations II: A Hilbert space representation. Kybernetes, 34(1): 192-221 (2005).

[29] C. Begz. Mathematical Approaches to Cognitive Linguistics. *International Journal of Applied Linguistics & English Literature*, 2(4): 192-199 (2013).

[30] T. Van Dijk. Formal semantics of metaphorical discourse. *Poetics*, 4(2-3): 173-198 (1975).

[31] E. Bruss. Formal semantics and poetic meaning. *Poetics*, 4(14-15): 339-363 (1975).

[32] T. Giraud. Constructing formal semantics from an ontological perspective. The case of second-order logics. Synthese, 191(10): 2115-2145 (2014).

[33] N. Chino. Hilbert space theory in psychology (1) - Basic concepts and possible applications. *Bulletin of the Faculty of Letters of Aichi Gakuin University*, 28: 45-65 (1998).

[34] A. Ortony. Metaphor, language, and thought. In: *Metaphor and thought*. A. Ortony (Ed.) Cambridge University Press, Cambridge; New York: 1-16 (1979).

How can metaphors be interpreted cross-linguistically?

Yorick Wilks[1]

Abstract. Research on metaphor as a phenomenon amenable to the techniques of computational linguistics received a substantial boost from a recent US government (the iARPA agency) funding initiative that set up a number of teams in major universities to address the issues of metaphor detection and interpretation on a large scale in text. Part of the stated goal of the project was to detect linguistic metaphors (LMs) computationally in texts in four languages and map them all to a single set of conceptual metaphors (CMs). Much of the inspiration for this funding was the classic work (Lakoff and Johnson, 1980) which posited a set of universal metaphors used across cultures and languages.

I wish to examine the assumptions behind this goal and in particular to address the issue of how and in what representation such CMs can be expressed. I shall argue that a naïve approach to this issue is to make very much the same assumptions as the work of Schank and others in the 1970s (including the present author): namely that there can be a universal language of "primitives" for the expression of meaning, which in practice always turns out to be a form of simple English (or in the case of Schank, atoms like PTRANS, very close to English words). In none of those system was the sense ambiguity of the English-like terms every tackled in a systematic way (though see: Guo 1989). Reviving that assumption for the study of metaphor raises additional issues since, even if the *senses* of the terms in those CM representations could be added, by annotation from a standard lexicon for the CM representations, metaphors often considered to deploy new senses of words which will not be found in existing sense inventories like computational lexicons which, if true, might make such annotation impossible (though later in the paper I shall argue against just that novel deployment of sense in metaphor). This paper is not intended just to present a negative conclusion; I also argue that the representation of metaphors in a range of languages can be brought together within some CM scheme, but that simply reviving the *English-as-interlingua* assumptions of forty years ago is not a good way to make progress in this most difficult area of meaning computation.

In what follows I first discuss first the representation of CMs and ask: in what language are they stated? I argue the need for some inclusion in the representation of the senses of their constituent terms within the CM, or at least a default assumption that the major sense (with respect to some lexicon such as WordNet) is the intended one. I then consider the issue of conventional metaphor and its representation in established lexicons (again such as WordNet) and [1]the effect that can have on detection strategies for metaphor, such as selectional preference breaking. I then argue that the mapping of text metaphors to CMs, as well as the empirical, rather than intuitive, construction of CM inventories requires further use of preference restrictions in lexicons by means of a much-discussed process called projection or coercion. I conclude that only the use of (computable) procedures such as these for metaphor detection and mapping can lead to a plausible program for the large-scale analysis of metaphor in text, and that Lakoff's views on metaphor lack these empirical underpinnings.

[1] Florida Institute of Human and Machine Cognition, 15 SE Osceola, Ocala FL 34471. Email: ywilks@ihmc.us

1 INTRODUCTION

Understanding prose in any natural language rests first on it being in a language one understands, let us say English for the purposes of this paper. But problems in understanding arise even for native speakers of English as well as with translations, human or mechanical, from other languages. One way of capturing the additional understanding needed that goes "beyond knowing the words and the grammar" is expressed by the term "metaphor". This notion conveniently expresses aspects of culture and figurative expression that go beyond literal or ostensive meaning and are crucial to understanding. These phenomena are sometimes opaque even to those who are experts in the language concerned. Metaphor also has the advantage that it has been an area of research in computer language processing for decades, and one that has yielded real results. That research has been driven in part by the writings of George Lakoff at Berkeley [1] who has developed an approach to metaphor that rests on the following assumptions (in my terms, but I think fairly uncontentious):

- There are similar metaphors found in all cultures that are crucial to understanding language.
- These metaphors can be discovered and listed, even if not exhaustively.
- We can proceed with analysis as if these metaphors can be not only paraphrased but expressed in English.

For example, such a universal metaphor might be expressed (in English) as LIFE IS A JOURNEY and we shall refer to items like this as Conceptual Metaphors (CM). There is then an initial analytic question of how to detect metaphors in text, possibly related to or "expressing" that CM such as *The pensioner was nearing the end of his road*. After locating this sentence as a metaphor there is then the task of matching it to such a stored generalized CM form. We shall refer to linguistic strings like the one in italics as Linguistic Metaphors (LM), There may then be the problem, if one believes in the universal nature of CMs, of how to locate expressions of "similar" metaphors in, say, Farsi to that same CM. The capitalised words in the English form of the CM may themselves have many senses and the question

immediately arises as to how an algorithm is to determine which sense is intended by "LIFE" in that CM: that it is not, say, a "a life as in a children's game of hide and seek, a score token".

One problem with metaphor research, at least from a computational or Natural Langauge Processing (NLP) perspective, is that universal theories like the one above (expressed by the three bullets) have proved resistant to computational implementation, which has not been the case with other, quite different, empirical approaches based on bottom-up detection of LMs in text (e.g. [3], [4]), rather than starting from a set of a priori CMs. We shall now turn to questions about *the representational language in which CMs are stated* and how they to be intuitively understood, since their terms (e.g. LIFE) do not disambiguate themselves

2. THE LANGUAGE OF CONCEPTUAL METAPHORS (CMs)

I shall argue that a crucial aspect of the research problem, which many seem to believe is a solution, is that CMs are classically expressed in English words but without any realization of what that entails. When this is pointed out, a frequent response is that this is an accidental fact of no significance and we can just carry on since though they appear to be English words they are not, but rather some form of symbol outside ordinary natural language. I believe this is profoundly inadequate response. It is in fact a recrudescence of the early discussions in AI and NLP in the 1960s and 1970s on the role of interlinguas in machine translation and in cognitive representations generally. There was a fashion at that time for limited languages (expressed by English primitives terms) within systems for the semantic representation of language content (e.g. in the work of Schank [5]; Wilks, [6] and many others). I am not here defendingthat approach, only pointing out that the extended discussion forty years ago (e.g. in [7]) of the adequacy or otherwise of this limited language of (English-like) primitives to carry the general meaning of language expressions has many similarities to what we are discussing now, nearly fifty years later, in regard to CMs.

There was no real resolution to that controversy of long ago: key references are Pulman's [8] attack on the practice from a linguistic perspective, and Lewis [9]from a philosophical one, in the course of which Lewis invented the term "markerese" for the self-description of language in linguistics (e.g. by Fodor and Katz, [10]) by means of word-like *markers* with no illumination or benefit. But the critiques were not heeded and much such representational work continued, simply because researchers in semantics could see no alternative (outside radical connectionism) to continuing to use symbols to represent the meanings of other symbols. Montague [11] was a philosopher who reacted against markerese but his representations of mean, although more replete with logical forms that those of Fodor and Katz, still were expressed in symbols including English-like words, though now usually expressed in lower case and with an apostrophe attached. Language content had to represented somehow, theorists reasoned, so why not in this English-like language? Dictionaries, after all, describe word meanings using the very language they describe, and so the practice has continued, ignoring the waves of philosophical and linguistic criticism, simply because there seemed to be no alternative. What has happened is that the language terms used for representation have been embedded in more logical and formal-seeming structures so as to make them palatable, but the underlying issue has not gone away. That issue is: How can I describe semantic content with a term such as MAN, HUMAN or ANIMATE and be confident I know what it means, and not just "means in English"? I shall now turn to how problems of CM representation problems can be ameliorated with the aid of a sense-lexicon.

3. REPRESENTING CMs UNAMBIGUOUSLY WITH MAJOR WORD SENSES

If we are to use CMs at all, no matter how derived or expressed, they must be in as word-sense-neutral a form as we can manage. To my knowledge this has never yet been fully considered as problem, perhaps an insurmountable problem, let alone a solved problem. We cannot just ignore this as we do when we say, for example, that [POVERTY IS A GAP] is a CM, and underlies the metaphor "poverty gap", and that we just know what the senses of the words in the CM are present in that expression and that they make up a CM. Just suppose that we had two CMs in our inventory of universal metaphors that could be written as:

POVERTY IS A GAP

POVERTY IS AN ABYSS

Now suppose we want to locate Russian metaphors and find the text string (LM) containing the keywords : *бедность провал*, which mean roughly "poverty" and "failure". But, and here is the problem "*провал*" can also means "abyss" and "gap" in English; in which case how do we know which of these two so-called universal CMs to match the Russian LM to? Or should we seek for or construct a third CM [POVERTY IS FAILURE]? It seems clear to me that either:

1) The CMs are in some language other than English, in which case how do we know what English word senses the terms above correspond to, since the English words "poverty", "failure" and "abyss" may all have multiple senses in, say, WordNet [12]. If, however, the terms are not English but some universal language of indeterminate syntax and semantics, how can LMs ever be matched to CMs as any serious theory of metaphor seems to require?
2) If however, the terms in the two CMs above *are* in English, and they certainly appear to be, then we need to know what senses those words have in those particular forms, so as to match any word in an English or Russian LM to them.

A natural way of carrying out the requirement in (2) is to tag the English words in the CMs (and the words in any putative LMs) with WordNet senses. Since the EuroWordNet project [12] in which the present author participated, we now have a convenient

way of setting up such a match since that project took the core Princeton WordNet for English as, essentially, an interlingua, and linked senses in the Wordnets for other languages to those core senses. So, for example (and the correctness of these correspondences does not matter for the argument): there may well be an English WordNet sense of "failure", namely failure#1 that is deemed by a EuroWordNet mapping to be the same sense as Провал#1 in the Russian WordNet. Again, there may be a "Провал#3" that similarly corresponds to "abyss#1".

What do we want to say about universal CMs and their ability to support the analysis of metaphor instances in such a case? The first natural thing to say---given the above WordNet assumptions---- is that the original Russian string "*бедность провал"* can express both CMs and we cannot decide which. But that is only true if we cannot decide which sense the last word bears in the Russian LM. If it bears only one of the two noted senses then the Russian LM matches one and only one of the CMs—assuming now the CM terms are tagged with WordNet senses. Russianists should note here that I am ignoring the case issues for the proper expression of that string in Russian and just concentrating on the main forms of the words. Also, I am not suggesting it would be problematic if a LM were to match to two possible CMs, though I do not believe that need be the case here. It could be that other, perhaps pragmatic, factors outside the text would settle the choice. My only point here is that a systematic empirical account of mapping LMs to CMs should take account of this possibility and standard contemporary metaphor theories do not consider the issue at all.

Now a Russian speaker may take that (LM) phrase to have one and only one of those senses in context—assuming the Russian speaker can understand the distinction we are making with the words "failure" and "abyss" in English—let us assume they can, even though the string may be too short and vague for a wordsense disambiguation program to determine the sense in that LM context.

Or, and this is a quite different possibility, is it the case that, in a metaphorical string such as the LM "Poverty is failure" we cannot rely on the normal psychological or computational methods to resolve a word sense for us. Since the content is, more or less, novel, at least on first encounter, the standard disambiguation techniques may well not work because they are all, to some extent, based on redundancy, which does not apply to novel utterances? So, to use an old and hackneyed example, if someone says *The shepherd swung his crook*, we infer that "crook" is a tool for shepherds not a gangster, simply because of the redundant presence of "shepherd". But in LMs this may not be available, unless the metaphor is dead, or lexicalized or otherwise familiar (in which case wordsense disambiguation hardly applies). What I am suggesting is that perhaps in metaphors, especially novel ones, the words must be taken in their basic senses by default, as it were, *because in a metaphor we lack the familiar context to resolve a participating word to any non-basic sense.*

This conclusion is perhaps not very striking but rather obvious: words of a real language, like English, can only function in an interlingua (such as CMs constitute) on condition that they bear their "basic" senses, which will, in WordNet terms, usually mean #1 for any given word. This implies that in the capitalized English CMs above, each term implicitly has whatever its #1 sense is in WordNet.

So to return to the purported sense correspondence in Euroword-net style:

failure#1 is deemed by a EuroWordNet mapping to be the same sense as Провал#1. Again, there may in addition be a "Провал#3" that similarly corresponds to "abyss#1".

This line of reasoning would imply that we should take the CMs (and LMs, with the caveat above) in their default #1 senses, since we have no information to allow us to do anything else. Hence "Провал" should be taken in the context above to be Провал#1, its first sense, and so as a CM about failure not about an abyss, even thought the latter could conceivably be indicated by another context for the same words. This suggestion that the senses in a CM are major senses of the relevant words also implies that the two CMs above are different from each other, which preserves the insight of the tradition that metaphors are strictly speaking lies (attributed variously to Mark Twain, Nietzsche et al.) rather than the less acceptable alternative that CMs are tautologies, where the constituent senses simply recapitulate each other.

This risk of tautology in the expression of CMs is very real even if we are wary and assign (implicitly as main senses) interpretations to the symbols in CMs. If, in the CM [POVERTY IS A GAP], we allow the first WordNet sense interpretation to "gap" we get:

S: (n) **gap**, spread (a conspicuous disparity or difference as between two figures) "gap between income and outgo"; "the spread between lending and borrowing costs"

Thus, and depending on the sense assigned to "poverty", we have a very real risk of tautology since this sense of "gap" is itself abstract (and not, say, a gap between two pieces of wood) and itself very close to any definition of poverty, or at least "relative poverty" the currently fashionable version. This unfortunate fact can be dismissed, or simply accepted as a weakness or error in WordNet, or, perhaps, as a reason for excluding [POVERTY IS A GAP] as a CM.

One important inference from this discussion, if it has any value, is that we cannot just say, as many researchers in the Berkleyan universal metaphor tradition seem to want to, that some particular metaphor "in one language" is commoner than in another. As we have seen, it is a very sophisticated matter to establish whether LMs in two languages point to a single CM or not, given the problems of how any CM is to be unambiguously represented and, given the need for some lexical resource of at least the size and scope of (Euro)WordNet in order to do that. In the example above, the LM word strings in question in the two languages—Russian and English ---actually point to different CMs in the common interlingua, a conclusion that, we argued, undermines the foundation of the Berkeley approach to understanding metaphor, since the LMs could clearly be interpreted as "meaning the same thing". At this point, let us step

back and review the basic role of "preference" in detecting, then mapping, metaphors.

4. THE ROLE OF PREFERENCE IN DETECTING AND MATCHING METAPHORS

An exception to the "rule of main senses" we have just stated, as far as LMs are concerned, is the situation we have defined elsewhere as one of "conventional metaphor" [13] This is where a lexical resource such as WordNet actually encodes a metaphorical sense as a (dead or) conventional metaphor. Our approach to detecting metaphor has been that an initial *sufficient* criterion for a surface (LM) metaphor to be present is that a verb or adjective "preference" is broken [6] e.g. in the simplest case the verb does not receive the agent or object it expects (whether that last notion is unpacked linguistically or statistically) in a stereotypical case. Verbs and adjectives will, of course, have multiple senses in the lexicon, each with its own preferences. So to write *fall into poverty* is to break the preference for a spatial-container-like object for the basic sense of "fall into". This general criterion reappears frequently in the literature (e.g. the recent work of Shutova [4]) indeed it is not clear there is any alternative to it as a basic criterion for metaphor recognition, unless one believes that metaphors are detected by direct matching to stored CMs. As we have seen above this a notion whose very intellegibilitys dissolves somewhat under scrutiny.

If such preferences, and the associated noun-senses for fillers, are thought of as stored in a respository like WordNet or VerbNet, then what counts as a broken preference depends crucially on the state of lexicon at a given time, since sense inventories extend with time and indeed often come to store senses that were in origin metaphorical. Where that is the case, a dead, or as we would prefer to say conventional, metaphor will not result in a broken preference with respect to WordNet because in such a case the metaphorical sense is itself stored in WordNet and so will fit the demands of the corresponding verb.

So, to take a very simple and uncontentious example:

Public employees' unions have built a fortress around their pension systems

In VerbNet [14] we find the following:

[[VerbNet: build

Member of

§build%2:31:03 (member of VN class base-97.1)

§build-26.1-1

•WordNet Sense 1

•Agent [+animate | +machine]

So "<u>Unions" violates Agent restriction for build</u>

•WordNet Sense 8

•Agent [+animate | +organization]

<u>"Unions" satisfies the Agent restriction ---as an organization—for build</u>]]

The situation is one where the primary sense of "build" is not satisfied by the first sense of the agent the sentence contains but is satisfied by a "lower' (in this case #8) sense. In [13] I proposed that this could serve as a useful heuristic (i.e. main sense failure but some lower sense a successful match) for detecting conventionalized metaphors of the sort this sentence contains, since such metaphors would be missed by any "preference breaking" heuristic for metaphor detection as there is a (lower) sense of "build" available for which the agent preference here is satisfied. The heuristic was that a main sense fails and a lower sense satisfies; and both parts must be true. Its main defect is that it relies on the ordering of senses in WordNet as carrying information, which is generally true but as always with this database has many errors and omissions.

The point here is not to draw attention to this metaphor detection heuristic against a large lexicon for its own sake, but only to show a limitation on the earlier suggestion that metaphor detection (and as we shall discuss below, metaphor mapping to CMs) must depend on the main senses, as listed in a lexicon. Our claim here is that this heuristic for detecting conventional or lexicalized metaphor does not compromise the general value of that rule. In the case of the above example, there are arguably two CM metaphors present: the major one is to do with barriers and the protection of assets, however expressed, and the other is more simply (and even though it is, more strictly, a meronym, though such differences are not crucial here):

ORGANIZATIONS ARE PEOPLE

which is expressed (in major senses of the relevant words) by the process of detection we have described.

The latter move is the basis of how preferences, and their violations in metaphor, are also central to the subsequent process of mapping from a detected metaphor to some stored form, which we are calling CMs. If we were again dealing with "He fell into poverty" we might expect the broken preference for the object of "fall into" to be some coding for hole/abyss/gap/aperture. The inference from that detection to the underlying metaphor in play is generally to assert that the metaphor's object (poverty in this case) is being asserted to be equivalent to the preferred filler that is made available in the lexical coding (e.g. in VerbNet, see [14]) but not in the sentence itself. This would lead directly to some form such as:

POVERTY IS AN ABYSS

as a potential CM, empirically derived from this example text rather than a linguist's intuition. The interesting difficulty is to determine at exactly what level its last term is to be expressed,

since “abyss” is, in general, a very magnified form of hole. The mapping process from a metaphor instance, or LM, to a CM, however expressed, will require an ontology of the kind that underlies WordNet to navigate from what appears in a VerbNet coding (perhaps “hole”) to an item in an already stored CM (perhaps, as here, “abyss”). This method, merely sketched here, can in principle serve to map LMs to CMs, and to create potential CMs from text.

This process, making use of the preferred constituents of lexical codings, has been central to a number of systems based on inferences within lexical semantic structures and under names such as “projection” and “coercion” (e.g. Wilks, [6]; Pustejovsky,[15]; Nirenburg and Raskin, [16] and Hanks [17]) among many others. It provides at least the beginning of a process of determinate empirical construction of CMs from text cases quite different from the intuitive creation of CMs in the Berkeley tradition. Moreover, [22] contains a sophisticated analysis of some of the cross-lingual issues raised here. Further possible examples of the method would be with a failed subject+verb preference in *Israel has inflicted this wound on itself*. There we can get (from the stored VerbNet subject preference for “inflict” as PERSON) we can link the existing target (Israel) to the preferred subject (as source), namely PERSON, and then the WordNet type of “Israel” as COUNTRY to give as a possible CM: COUNTRY IS PERSON. We could do the same for verb+object failure as in: *The bank hyenas are feeding on money*, assuming we have access to “feed on” as a verb with its own preferences FOOD or EDIBLES. Then, using similar reasoning to that for subjects above, and again combining the assigned object and the preferred object, we can derive directly a potential CM: MONEY IS FOOD. For adjective+noun preferences, similar processes are possible, as in *Brazil’s economic muscle will become increasingly important*. If we have a preference established for the preferred type of noun associated with the adjective “economic” as COMPLEX-SYSTEM, then from the existing adjective object “muscle” (and taking its semantic type from WordNet as BODY) we then have directly a CM: COMPLEX-SYSTEM IS BODY. Many metaphor theorists would want to argue that equations of target and source CMs produced by a process such as this must be brought under some higher level generalization on both sides of the assertion in the CM, as we shall now show.

Notice though that no claims here depend on the actual quality or completeness of resources such asVerbNet or WordNet. These are always variable, depending on the language used, and will always contain errors and omissions, as well as being constantly changing with the language itself. The only claim is that some such resource will be needed to carry out the processes described here, even if augmented in practice by statistical corpus computations (some of which augmented these resources in the work described in [13]).

There has been criticism of processes of this sort applied to the empirical construction of CMs in this manner: during a recent large-scale metaphor detection and interpretation project a project manager wrote:

“[CMs that were] proposed….. were inconsistent and generally unmotivated. For the most part, the relationship of an LM (for a Target) and a proposed CM was semantically extremely shallow with generally no mapping at all. This process caused a huge proliferation of “lexical” CMs, often dependent on a synset label from WordNet.”[18]

It is odd, in the current empirical climate, to criticise a linguistic process for being grounded in data, rather than linguistic intuition. One must also respond (a) that there is no known correct *level* for the expression of CMs beyond the intuitions of metaphor theorists, so no level is demonstrably “too lexical” and (b) more fundamentally, the CMs are inevitably in some language (usually English) and require sense disambiguation of their terms, as we argued at length above. They are not in a language that is self-disambiguating, since nothing is. Hence the presence of WordNet labels, even if implicit, so as to indicate main senses as we suggested above, is inevitable. That would be a feature not a bug.

The problems of the appropriate level for the expression of CMs, their distance and separation from LMs and their very origins in intuition, are not ones that preoccupy only NLP researchers, as is clear from Deignan’s:
“.... at some points in the development of CMT [Conceptual Metaphor Theory], there has been a tendency for researchers to propose new conceptual metaphors using limited linguistic evidence. For instance, [19] take the idioms "he really couldn't swallow it" and "[leave] a bad taste in the mouth" as instantiations of a conceptual metaphor termed ACCEPTING SOMETHING IS EATING IT. It is not clear how many other realizations there might be of this conceptual metaphor, and in what way it differs from the more-often cited IDEAS ARE FOOD. Kovecses [20] lists as a conceptual metaphor CONSIDERING IS CHEWING, which again is difficult to separate from IDEAS ARE FOOD. If this tendency becomes widespread, the notion of a conceptual metaphor loses clarity, along with any predictive power it may have had.” ([21] p.105)

I take the force of this comment, from a corpus linguistic standpoint, to be consistent with the NLP processing critique advanced in this paper, and indeed with the internal project critique quoted earlier above. However, there is a difference of emphasis here: Deignan argues that CMT theorists in fact make up CMs from data, no matter what they say about intuition, and I have argued that they should be constructed by a determinate process from data since there is no other reliable route. But the internal project critique earlier seems to say that derivation from data in any such way is a istake and leads to shallow CMs and “real” CMs come only from intuition. I hope I have set out reasons for thinking tis comment profoundly wrong and out of line with all modern thinking on linguistics and data.

5. THE LAKOFF BERKELEY VIEW OF METAPHOR REVISITED

This view, against which I have argued, seems to me to rest on the following, very questionable, assumptions:

1. There is a set of universal CMs, determinable by linguistic intuition and underlying all languages.

There is no suggestion this set should be small, even fixed, as Schankian primitives were once held to be, and certainly some

depend on developments in technology, economics etc. Yet, as I have argued, there is no empirical evidence for their existence or how many of them there are, and intuition as a source of linguistic insight is no longer considered reliable, taken alone. However, there may be a discovery procedure for them from text along the lines suggested here (and in [6]).

2. CMs can be expressed in an English-like language, whatever their real underlying representation.

I have argued that they are in fact in English, as they appear to be, and not as an inevitable approximation; this is made clear by the problem of expressing exactly what senses their constituent words are to be taken in. This situation is only tolerable as a heuristic if some form of cross-lingual sense representation is incorporated into the representation, as suggested here.

3. Surface metaphors (LMs) in languages can be mapped to these CMs in a determinate way.

I have argued that no definitive procedure is ever given, within this tradition, for performing this crucial step and it can only be attempted at all with the aid of some fairly reliable, cross-sense mapping of the languages concerned, such as (Euro)WordNet.

If LMs can be matched bottom-up to CMs in something like the way sketched here---as opposed to being the subject of some direct matching top-down from stored CMs to LMs in text---- it should be possible to count how many LMs correspond to a given CM. That would then make it possible to estimate the frequency of occurrence of CMs in a reliable manner. That analysis could be extended cross-lingually and cross-culturally if parallel text were available. Suppose we had an English-Spanish parallel text in which sentences are aligned. We could then ask whether LMs are detected in parallel (putatively synonymous) sentences and, if so, do they map to the same CMs. If they do, that would be independent confirmation of the utility or universality of such a CM. Quantitative and distributional questions about universal metaphor can only be asked, it seems to me, if procedures of this kind I sketch here are developed, but these are not obviously compatible with standard Lakoffian approaches to metaphor, though there is no reason in principle, or course, why it could not develop so as to incorporate some empirical theory of sense ambiguity like the present one.

My main conclusion is that, for these reasons, Berkeley metaphor theory cannot easily be the basis of an empirical exploration of metaphors in texts in multiple languages, and that any research program aimed at the interpretation and translation of metaphor instances so based will have been mistaken.

ACKNOWLEDGEMENTS

The paper is indebted to comments from Patrick Hanks, Robert Hoffman, Brian MacWhinney, Jaime Carbonnel, Tomas By and Sergei Nirenburg, though the errors are all mine as always.

This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0020. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

REFERENCES

[1] Lakoff, G., and Johnson, M. 1980 *Metaphors we live by*. University of Chicago Press: Chicago.

[2] Guo, C-M. 1989. *Constructing a machine-tractable dictionary from "longman dictionary of contemporary English*. Doctoral dissertation, Computer Science, New Mexico State University.

[3] Fass, D. and Wilks, Y. 1983. Preference semantics, ill-formedness and metaphor. *Journal of Computational Linguistics,* 9, pp.1798-187.

[4] Shutova, E., Teufel, S. and Korhonen, A. 2012. *Statistical Metaphor Processing*, Computational Linguistics, 39(2).

[5] Schank, R (Ed.,) 1975. *Conceptual Information Processing*. Elsevier: Amsterdam.

[6] Wilks, Y., 1968/2007. Making preferences more active. Reprinted in Ahmad, Brewster and Stevenson (Eds.) *Word and Intelligence I.* Springer: Berlin.

[7] Wilks, Y. 1977/2007. Good and bad arguments for semantic primitives. Reprinted in Ahmad, Brewster and Stevenson (Eds.) *Word and Intelligence I.* Springer: Berlin.

[8]Pulman, S. 1983. *Word meaning and belief*. Croom Helm: London.

[9] Lewis, D., 1972. General Semantics, In: Davidson and Harman (eds.), Semantics of natural language. Reidel: Dordrecht.

[10] Fodor, J. and Katz, J. 1963. The structure of a semantic theory. *Language*, *39*(2), Apr-Jun, 170-210.

[11] Montague, R. 1974. *Formal philosophy : selected papers of Richard Montague* / ed. and with an introd. by Richmond H. Thomason. New Haven: Yale Univ. Press.

[12] Vossen, P. (ed.) 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer: Amsterdam.

[13] Wilks, Y., Dalton, A., Allen., J.. Galescu, L. 2013. Automatic Metaphor Detection using Large-Scale Lexical Resources and Conventional Metaphor Extraction. *Proc.1st Workshop on Metaphor in NLP (Meta4NLP 2013).* Atlanta, GA.

[14] Windisch Brown, S., Dligach, D., and Palmer, M. 2011 VerbNet Class Assignment as a WSD Task. In IWSC 2011:

Proceedings of the 9th International Conference on Computational Semantics, January 12 - 14, 2011, Oxford, UK.

[15] Pustejovsky, J. 1995. *The Generative Lexicon.* MIT Press: Cambridge, MA.

[16] Nirenburg, S. and Raskin, V. 2004. *Ontological semantics.* MIT Press: Cambridge, MA.
[17] Hanks, P. 2013. *Lexical analysis: norms and exploitations*. MIT Press: Cambridge, MA.

[18] iARPA: http://www.iarpa.gov/Programs/ia/Metaphor/metaphor.html

[19] Gibbs, R., Bogdonovich, J., Sykes, J., & Barr, D. 1997. Metaphor in idiom comprehension. *Journal of Memory and Language*, 37, pp. 141-154.

[20] Kovecses, Z. 2002. *Metaphor: a practical introduction*. Oxford University Press: Oxford.

[21] Deignan, A. 2005. Metaphor and corpus linguistics. Benjamins: Amsterdam.

[22] Prior, A., Wintner, S., MacWhinney, B., & Lavie, A. (2011). Translation ambiguity in and out of context. *Applied Psycholinguistics, 32*, 93-111.

Relevance Theoretic Comprehension Procedures: Accounting for Metaphor and Malapropism

Zsófia Zvolenszky[1]

Abstract. According to Sperber and Wilson, relevance theory's comprehension/interpretation procedure for metaphorical utterances does not require details specific to metaphor (or nonliteral discourse); instead, the same type of comprehension procedure as that in place for literal utterances covers metaphors as well. One of Sperber and Wilson's central reasons for holding this is that metaphorical utterances occupy one end of a continuum that includes literal, loose and hyperbolic utterances with no sharp boundaries in between them. Call this the *continuum argument about interpreting metaphors*. My aim is to show that this continuum argument doesn't work. For if it were to work, it would have an unwanted consequence: it could be converted into a continuum argument about interpreting linguistic errors, including slips of the tongue, of which malaprops are a special case. In particular, based on the premise that the literal–loose–metaphorical continuum extends to malaprops also, we could conclude that the relevance theoretic comprehension procedure for malaprops does not require details specific to linguistic errors, that is, details beyond those already in place for interpreting literal utterances. Given that we have good reason to reject this conclusion, we also have good reason to rethink the conclusion of the continuum argument about interpreting metaphors.

1 INTRODUCTION

Mrs. Malaprop, a character in Sheridan's (1775) play *The Rivals* had a tendency to make linguistic errors of a special sort: she would describe people as being "the pineapple of politeness" (when she meant *pinnacle*); or "as headstrong as an allegory on the banks of the Nile" (when she meant *alligator*). Such slips of the tongue have since come to be called malaprops. In a framework like relevance theory, how might we characterize the process of interpreting malaprops as opposed to interpreting literal utterances? We will see that addressing this question exposes a challenge for the relevance theoretic treatment of *metaphorical* utterances.

Within philosophy of language as well as rhetoric the following claims are widely held, considered platitudinous even: the distinction between literal and figurative discourse carries theoretical importance, and metaphorical utterances clearly fall on the figurative side of the divide, constituting departures from literality. Relevance theory calls into question these time-worn claims.

Relevance theory [1, 2] has become, over the past three decades, a leading research program in pragmatics. Its founders', Dan Sperber's and Deirdre Wilson's [3] most recent position on metaphorical utterances is that (i) the interpretation/comprehension procedure for metaphors does not require resources beyond those already needed to account for literal utterances (call this the *procedure claim*), and (ii) metaphorical utterances occupy one end of a continuum that includes literal, loose and hyperbolic utterances (call this the *continuum claim*). Relevance theorists seem to regard the continuum claim as one reason to hold the procedure claim; call this the *continuum argument about interpreting metaphors*.

Sperber and Wilson subscribe to this continuum argument: "We see this continuity of cases, and the absence of any criterion for distinguishing literal, loose, and metaphorical utterances, as evidence not just that there is some degree of fuzziness or overlap among distinct categories, but that there are no genuinely distinct categories, at least from a descriptive, psycholinguistic or pragmatic point of view. Even more important than the lack of clear boundaries is the fact *that the same inferential procedure is used in interpreting all these different types of utterance*" [3, p. 111–112, emphasis added].

In this paper, I aim to show that the continuum argument about metaphors, if it were to work, would face an unacceptable consequence: the argument would license *a continuum argument about interpreting malapropisms* (and more generally, a continuum argument about linguistic errors):

> *Continuum premise for malaprops*: The literal–loose–metaphorical continuum extends to malaprops.
>
> *Procedure conclusion for malaprops*: The relevance theoretic comprehension procedure for malaprops does not require details beyond those needed to account for literal utterances.

We have good reason to resist the malaprop conclusion: surely, when we manage to interpret Mrs. Malaprop as having meant 'alligator' when she said 'allegory', the fact that the lexically encoded meaning of 'allegory' becomes wholly irrelevant is a detail that is bound to be featured in a full description of our process of interpreting her. And if we want to resist the malaprop conclusion, then we have to find fault with the continuum argument about interpreting malapropisms. There are two strategies we could follow: we could fault the premise or fault the argument itself as non-truth-preserving. I will argue that the former strategy is not open to us, so our remaining option is to regard the malaprop argument as non-truth-preserving. But then we have to say the same about the continuum argument about interpreting metaphors also. Whether the comprehension

[1] Dept. of Logic, Institute of Philosophy, Eötvös Loránd Univ. (ELTE), Budapest, Hungary 1088. Email: zvolenszky@elte.hu.

[2] Our concern here is with acts if *linguistic* communication, but the communicative principle and the relevance theoretic framework are intended to apply to a broader range of cases: acts of ostensive communication which include, besides linguistic utterances, certain kinds of non-linguistic acts also.

procedure for interpreting metaphors includes any details specific to metaphor (or nonliteral discourse) therefore remains an open question.

2 RELEVANCE THEORY ABOUT THE LITERAL–LOOSE–METAPHORICAL CONTINUUM

Dan Sperber and Deirdre Wilson's influential framework for the study of communication, relevance theory [1, 2] outlines an inferential comprehension procedure that hearers follow in arriving at an interpretation of speakers' linguistic utterances. Crucially, the comprehension procedure is delimited and guided by specific assumptions about relevance (i)–(iii), accepted by speakers and hearers alike. (i) Cognition (generally, not just in the case of communication) aims to maximize relevance (this is the *cognitive principle* of relevance). (ii) Linguistic utterances communicate a presumption of their own optimal relevance (this follows from the *communicative principle* of relevance[2]). And (iii) an utterance is *presumed to be optimally relevant* if and only if it is at least relevant enough to be worth the speaker's effort to process it, and it is the most relevant utterance compatible with the speaker's abilities and preferences. The kind of inference involved in the relevance theoretic comprehension procedure is inference to the best explanation [4]. The concepts encoded by the words the speaker has used on a given occasion are mere starting points for arriving, via inferential steps, at an interpretation of her utterance: her utterance's explicit content (the speaker's explicit meaning) on the one hand, and its implicit content (which consists of implicit premises and conclusions) on the other.

By explicit and implicit content, we mean content that was *intended* as such by the speaker. The hearer's task is to *reconstruct* the explicit content and implicit premises and conclusions that the speaker has intended to communicate. Of course, rarely, if ever do hearers converge on the very same concepts as those that speakers actually meant. Nor is this required for successful communication. It suffices that the concepts reconstructed by the hearer be ones that allow him to draw (nearly enough) the same inferences as those intended by the speaker; it is enough that the reconstructed concepts "activate contextual implications that make the utterance relevant as expected" [3].

A recurring example of Sperber–Wilson's [3, 5, 6] exemplifies *loose use*: 'Holland is flat' uttered in the context of the following conversation: Peter and Mary are discussing their next cycling trip. Peter has just said that he feels rather unfit. Mary replies: "We could go to Holland. Holland is flat." Sperber–Wilson [5] illustrate the inferential comprehension procedure via which Peter interprets Mary's second sentence as follows.

(a) Mary has said to Peter, 'Holland is flat'.	*Decoding of Mary's utterance.*
(b) Mary's utterance is optimally relevant to Peter.	*Expectation raised by the recognition of Mary's utterance as a communicative act, and acceptance of the presumption of relevance it automatically conveys.*
(c) Mary's utterance will achieve relevance by giving reasons for her proposal to go cycling in Holland, which take account of Peter's immediately preceding complaint that he feels rather unfit.	*Expectation raised by (b), together with the fact that such reasons would be most relevant to Peter at this point.*
(d) Cycling on relatively flatter terrain which involves little or no climbing is less strenuous, and would be enjoyable in the circumstances.	*First assumption to occur to Peter which, together with other appropriate premises, might satisfy expectation (c). Accepted as an implicit premise of Mary's utterance.*
(e) **Holland is FLAT*** (where FLAT* is the meaning indicated by 'flat', and is such that Holland's being FLAT* is relevant-as-expected in the context).	*(Description of) the first enriched interpretation of Mary's utterance as decoded in (a) to occur to Peter which might combine with (d) to lead to the satisfaction of (c).* ***Interpretation accepted as Mary's explicit meaning.***
(f) Cycling in Holland would involve little or no climbing.	*Inferred from (d) and (e). Accepted as an implicit conclusion of Mary's utterance.*
(g) Cycling in Holland would be less strenuous, and would be enjoyable in the circumstances.	*Inferred from (d) and (f), satisfying (b) and (c) and accepted as an implicit conclusion of Mary's utterance.*

Table 1. Interpretation of Mary's utterance 'Holland is flat'.

As indicated on line (e) (in boldface), the explicit content of Mary's utterance 'Holland is flat' is 'Holland is FLAT*'. FLAT* is an *ad hoc concept* Peter arrived at that is distinct from, broader[3] than the lexicalized concept encoded by the word 'flat': say, FLAT. Unlike FLAT*, the extension of FLAT doesn't include imperfectly flat surfaces like the Dutch landscape.

Loose use, as in 'Holland is flat' is a type of literal discourse[4] that involves some departure from the lexically encoded concept. While the departure is greater than in many other instances of literal discourse, Sperber–Wilson [3] stress that the comprehension procedure for *some* literal utterances (to wit: cases of loose use) already involves the formation of *ad hoc* concepts. They suggest further that even in literal utterances that do not involve a departure from the lexically encoded concept, the process of disambiguating the expressions used involves inferential steps similar to those in Table 1. For example, Mary's and Peter's idiolect may have (at least) two senses associated with the word 'flat', one of which amounts to, say, "having a smooth, even surface" while the other, to "is in a horizontal position"; Sperber–Wilson [3, p. 111] suggest that if Mary uttered "My computer screen is flat", the process of interpreting her utterance and deciding that she has in mind the first and not

[3] Alternatively, according to another prominent relevance theorist, Robyn Carston [7], the formation of *ad hoc* concepts involves conceptual narrowing as well as broadening.

[4] Sperber–Wilson [1, pp. 234–235; 3] stress the literal status of instances of loose use.

the second sense of 'flat'would take a similar inferential procedure as the one seen in Table 1.

Sperber and Wilson [3] gradually build up a continuum a cases with no clear boundaries in between them. The continuum includes cases of disambiguation like ("My computer screen is flat"), various examples of

- loose use (or broadening), covering a broad range:
  - *Approximation*: "Holland is flat"';
  - *Limited category extension*: "Here is a Kleenex", said of a piece of non-Kleenex-brand tissue;
  - *Creative category extension*: "For luggage, pink is the new black";
- *Hyperbole*: "Joan is the kindest person on earth";
- *Nonpoetic metaphor*: "Joan is an angel";
- *Poetic metaphor*: "The fog comes on little cat feet" (from Carl Sandburg's poem *The Fog*).

A central claim of relevance theory (besides Sperber and Wilson, see also Carston [7]) is that each of the listed cases involves the formation of an *ad hoc* concept, one that—as we go down the list of examples—exhibits a gradually greater degree of departure from the concept lexically encoded by the word used, that is, the concept that serves as one of the starting points for the comprehension procedure. The *ad hoc* concepts are then featured as part of the explicit content attributed to the speaker (as in line (e) in Table 1). The *ad hoc* concepts for the listed examples (except for poetic metaphors, to be discussed in detail in Section 4) are as follows:

- FLAT*, whose extension includes imperfectly flat surfaces like the Dutch landscape;
- KLEENEX*, whose extension includes paper tissues that aren't Kleenex brand;
- BLACK*, whose extension includes (roughly) objects of a fashionable, trendy color, among them pink suitcases;
- KINDEST PERSON ON EARTH*, whose extension includes people who are very kind, but not even close to being among the *kindest*;
- ANGEL*, whose extension includes nonangelic human beings who are very kind.

We are now in a position to formulate in far more depth and detail Sperber–Wilson's (and other relevance theorists') argument about interpreting metaphors:

THE CONTINUUM ARGUMENT ABOUT INTERPRETING METAPHORS

Continuum premise for metaphors:
All metaphorical utterances (poetic and nonpoetic alike) can be located on a continuum of cases that includes loose use (a kind of literal use) as well as hyperbolical and metaphorical uses. Further, the process of forming *ad hoc* concepts to arrive at the explicit content attributed to the speaker is a tool that is readily applicable to all metaphorical utterances (not just to instances of loose use and hyperbole).

Procedure conclusion for metaphors:
Equipped with the relevance theoretic comprehension procedure and the *ad hoc* concept formation tool, both already required for interpreting literal utterances like loose use, we have all the resources needed to describe the comprehension procedure at play during the interpretation of metaphorical utterances. No further details specific to metaphor (or figurative language use) are needed in a comprehensive account of interpreting metaphors.

In Section 3, I will raise an objection intended to show that the continuum argument about interpreting metaphors is flawed: even if we accepted its premise, that is not reason enough to accept its conclusion also. I will motivate this by giving what I think is an analogous argument about malaprops with a clearly false conclusion. Someone might then raise a counterobjection: the argument about malaprops has a false conclusion because its premise is false. So as long as we can maintain (as relevance theorists do) the continuum premise for metaphors while resisting its analog about malaprops, we are entitled to keep the continuum argument about interpreting metaphors and maintain that its conclusion is true because its premise is. In Section 4, I will elaborate this counterobjection and deflect it by showing that the malaprop premise and the metaphor premise are equally plausible. My objection therefore has traction and there is room to reject the procedure conclusion for metaphors, despite relevance theorists' arguments to the contrary.

3 AN OBJECTION TO THE CONTINUUM ARGUMENT ABOUT INTERPRETING METAPHORS

Once we have accepted the continuum argument about interpreting metaphors, along with its premise and its conclusion, we have, I claim, no reason to resist making the same moves with respect to an analogous argument about malaprops (and more generally, about linguistic errors):

THE CONTINUUM ARGUMENT ABOUT INTERPRETING MALAPROPS

Continuum premise for malaprops:
All malaprops can be located on a continuum of cases that includes loose use (a kind of literal use) as well as hyperbolical and metaphorical uses. Further, the process of forming *ad hoc* concepts to arrive at the explicit content attributed to the speaker is a tool that is readily applicable to all malaprops.

Procedure conclusion for malaprops:
Equipped with the relevance theoretic comprehension procedure and the *ad hoc* concept formation tool, both already required for interpreting literal utterances like loose use, we have all the resources needed to describe the comprehension procedure at play during the interpretation of malaprops. No further details specific to slips of the tongue (or more broadly: linguistic errors) are needed in a comprehensive account of interpreting malaprops.

But—my objection goes—there is a flaw in this argument: (i) its conclusion is clearly unacceptable and (ii) it remains unacceptable even if we accept its premise. And if we accept all this, we have exposed a flaw in the original continuum argument about interpreting *metaphors*. In the rest of this section, I aim to establish (i), in the next section, (ii).

The procedure conclusion for malaprops leads to the following bizarre results:

- *Allegory example.* In interpreting Mrs. Malaprop's utterance "She is as headstrong as an allegory on the banks of the Nile", the explicit content that hearers arrive at involves an *ad hoc* concept ALLEGORY*, which is constructed by broadening the concept lexically encoded by the word 'allegory' (about a certain kind of trope or figure of speech) in such a way that its extension includes *alligators*. The comprehension procedure is basically the same as that in Table 1, it's just that the degree of departure to get from FLAT to FLAT* is not as great as that from ALLEGORY to ALLEGORY*.
- *Spanking example.* In interpreting George W. Bush's utterance in the context of a speech he gave at a school "I want to spank all teachers" (he meant *thank all teachers*), the explicit content that hearers arrive at involves an *ad hoc* concept SPANK*, which is constructed by broadening the concept lexically encoded by the word 'spank' (about slapping) in such a way that its extension includes acts of *thanking*. The comprehension procedure is basically the same as that in Table 1, it's just that the degree of departure to get from FLAT to FLAT* is not as great as that from SPANK to SPANK*.

As mentioned before, the continuum argument about malaprops is readily extended to linguistic errors of all sorts, including slips of the tongue other than malaprops as well as mistaken translations like the following:

- *Steak example.* In interpreting a German speaker's order in a restaurant "I want to become a steak" ('bekommen' in German means 'get'), the explicit content that hearers arrive at involves the *ad hoc* concept BECOME*, which is constructed by broadening the concept lexically encoded by the word 'become' in English (about 'turning into') in such a way that its extension includes one thing *getting* another. The comprehension procedure is basically the same as that in Table 1, it's just that the degree of departure to get from FLAT to FLAT* is not as great as that from BECOME to BECOME*.

It is bizarre to think that when we manage to interpret successfully the German speaker's request to "become a steak", we are broadening the concept lexically encoded by the English word 'become'. After all, our grasping that he's talking about getting a steak rather than turning into one happens *despite* his use of the English word 'become'. We can say the same about understanding Mrs. Malaprop's and George W. Bush's utterances: it is *despite* the encoded meaning of the words they have used that we manage to interpret them as having said something about alligators and thanking, respectively.

In the light of this, it seems that relevance theoretic comprehension procedures, as they stand, are missing key details that distinguish malaprops (and more broadly, linguistic errors) from utterances that are literal or metaphorical. To wit: the procedure has to specify that in utterances like 'Holland is flat', 'Joan is an angel' (loose and metaphorical uses alike), the speaker has *not* committed a linguistic error; further, that the speaker (and hearer) takes the lexically encoded concept associated with her words to be in force, and would not retract her words when confronted with the concept lexically encoded by her words. By contrast, in the case of linguistic errors including malaprops, the hearer is rerouting the inference such that he sets aside the lexically encoded concept entirely, and the speaker, when confronted with the lexically encoded concept, would retract his or her words: "I didn't mean spanking teachers was desirable, I wanted to talk about thanking them." "I didn't mean there were allegories on the banks of the Nile, I wanted to talk about alligators". But we would have absolutely no grounds for seeking such additional details if we thought the continuum argument about malaprops worked and moreover featured a true premise. If, despite the argument about malaprops, we thought the additional details were needed, then we open the door to seeking additional details with which to supplement the comprehension procedure for metaphorical utterances also. And we thereby open the door to rejecting the conclusion of the continuum argument about metaphors.

An analogy helps illuminate what my objection, if successful, shows with respect to Sperber–Wilson's continuum argument about interpreting metaphors. If you are at Columbus Circle in Manhattan and want to take the subway to the Museum of Natural History (at 81st Street), then don't get on the A train (the 8th Avenue Express); despite the fact that you would initially approach your desired destination, eventually, your train would whizz right past the Museum of Natural History, taking you all the way to 125th Street in Harlem, far away from your desired destination. Likewise: if you don't want an inferential comprehension procedure for malaprops (and other linguistic errors) that invokes no more than the formation of *ad hoc* concepts at work in the comprehension procedure you posited for cases of loose use, then don't apply the continuum argument to metaphorical utterances, for you won't be able to get off there but will be whisked straight to a place where you don't want to be: the continuum argument about interpreting malaprops.

4 A COUNTEROBJECTION DEFLECTED

It seems natural to respond to the foregoing objection as follows: a distinguishing feature of linguistic errors, malaprops included, is that the speaker makes a mistake about which *word form* is associated with the lexically encoded concept that he or she wants to express: G. W. Bush has said 'spank' even though his intended concept is expressed by the word form 'thank'; Mrs. Malaprop has said 'allegory' even though her intended concept is expressed by the word form 'alligator'. Those voicing such a counterobjection may then continue: of course the swapping of word forms, and the fact that the hearer recognizes the swap and reroutes the inference accordingly, will be part of the comprehension procedure via which he interprets malaprops and the like. We are in no way forced to regard the alligator, spanking and steak examples as cases involving simply the formation of *ad hoc* concepts with extreme degrees of departure from the lexically encoded concepts that had served as starting points for the construction of the *ad hoc* concept. This is how the counterobjection goes.

Someone could maintain this line while holding on to the continuum argument for *metaphors* and its conclusion, by denying the premise of the continuum argument about *malaprops*. This would amount to showing either that—in the context of relevance theory—extending the literal–metaphorical

continuum to malaprops (and other linguistic errors) is unfounded, or that—again, in the context of relevance theory—extending the tool of *ad hoc* concept construction to malaprops (and other linguistic errors) is unfounded. In what follows, I will show that neither of these will work and hence the counterobjection fails. My response consists of three parts:

> (A) In the case of poetic metaphors, the *ad hoc* concept departs greatly from the lexically encoded one, yet Sperber–Wilson (and others) do not doubt that here, too, explicit content is arrived at via the construction of an *ad hoc* concept.
>
> (B) With respect to malaprops (and other linguistic errors also) we can talk about a continuum of cases ranging from limited to extreme degrees of discrepancy between the intended concept and the lexically encoded one. And the limited-discrepancy cases fit squarely the *ad hoc* concept formation paradigm and can be readily placed on the literal–metaphorical continuum Sperber–Wilson had posited.
>
> (C) In formulating the continuum argument about metaphors, Sperber–Wilson appealed to considerations (about there being a continuum of cases that encompasses various types of loose use, including approximation, limited and creative category extension, along with hyperbole, nonpoetic metaphor and poetic metaphor) based on which there is no reason to deny that the continuum and the process of *ad hoc* concept formation extends to all other examples that (i) themselves form a continuum, (ii) are candidates for being accounted for via the already posited inferential comprehension procedure featuring the formation of *ad hoc* concepts, and (iii) include clear candidates for inclusion on the literal–metaphorical continuum.

In Section 2, I have already given reasons for holding (C). In what follows, I will, in turn, motivate (A) and (B).

(A) concerns poetic metaphors. We've already encountered the example from Sandburg's poem "The fog comes on little cat feet". According to Sperber–Wilson, the explicit content arrived at in the comprehension procedure for interpreting this line of the poem involves the *ad hoc* concept: ON-LITTLE-CAT-FEET*. What Sperber–Wilson say about this concept signifies that it involves a great degree of departure from the lexically encoded concept: the *ad hoc* concept is supposed to help convey that the fog is spreading in a smooth, quiet, stealthy and deliberate way. Yet it remains quite vague what this *ad hoc* concept is, in what direction it takes off from the lexicalized concept, what does and does not belong in its extension. The authors offer us limited guidance on these matters: ON-LITTLE-CAT-FEET* "is the concept of a property that is difficult or impossible to define, a property possessed in particular by some typical movements of cats (though not all of them—little cat feet can also move in violent or playful ways) and, according to the poem, by the fog" [3, p. 122].

As Sperber–Wilson see it, the great distance between lexicalized and *ad hoc* concepts and the vague description of the latter is no obstacle to applying the *ad hoc* concept formation paradigm to highly creative, poetic metaphors. Then comparably great distances and vagueness characterizing ALLEGORY* (whose extension includes certain reptiles) and SPANK* (whose extension includes acts of thanking) should be no obstacle to applying the *ad hoc* concept formation paradigm to malaprops (and other linguistic errors).

Turning to (B), about examples involving limited-discrepancy between the encoded concept and the intended one. Examples like the following form a continuum with the extreme-discrepancy examples about allegory, spanking and becoming a steak. Meanwhile, these examples fit squarely within the *ad hoc* concept formation paradigm, comparable to the "Here is a Kleenex" and "For luggage, pink is the new black" type examples.

Ocean example (a slip of the tongue involving limited discrepancy). G. W. Bush said once: "I didn't grow up in the ocean—as a matter of fact—near the ocean—I grew up in the desert. Therefore, it was a pleasant contrast to see the ocean. And I particularly like it when I'm fishing." In interpreting the first portion of Bush's utterance, via *ad hoc* concept formation, from the encoded lexical meaning IN-THE-OCEAN, we arrive, by broadening, to IN-THE-OCEAN*, whose extension includes events and things *near* the ocean.

Library example (a mistaken translation involving limited discrepancy). A French speaker says: "There is a library around the corner" to mean that there is *bookshop* around the corner (in French 'libraire' means bookshop). In interpreting the utterance, via *ad hoc* concept formation, from the encoded lexical meaning of LIBRARY, we arrive, by broadening, to LIBRARY*, whose extension includes bookshops. (Such an utterance could also exemplify a slip of the tongue involving limited disrepancy.)

In the ocean example, the distance between IN-THE-OCEAN and IN-THE-OCEAN* is no greater and no less vaguely delineated than that between KLEENEX and KLEENEX*. The same can be said about LIBRARY and LIBRARY* also. And we can envision a continuity of cases from such limited-discrepancy examples to the more extreme ones like in the allegory, spanking and steak examples.

This concludes my justification for (A)–(C), which together show that the counterobjection about swapped word forms does not undermine the objection I had formulated against the continuity argument about interpreting metaphorical utterances. After all, the limited-discrepancy examples of linguistic error make clear that the continuum premise for malaprops (and other linguistic mistakes) is just as plausible as the continuum premise for metaphors. We therefore have at hand two analogous arguments, both with true premises, and the one about malaprops boasting a clearly false conclusion. Hence, the other argument, about metaphors, is also undermined: the truth of its premise is no guarantee for the truth of its conclusion.

5 CONCLUSION AND FUTURE WORK

The continuum argument about interpreting metaphorical utterances is central to Sperber–Wilson's conclusion that "[t]here is no mechanism specific to metaphors, no interesting generalisation that applies only to them. In other terms, linguistic metaphors are not a natural kind, and 'metaphor' is not a theoretically important notion in the study of verbal communication" [3, p. 97]. My aim has been to show that we need not accept this conclusion given that the continuum

argument about interpreting metaphors is flawed, as shown by its application to malaprops (and other linguistic errors).

In the wake of my objection to the continuum argument, several questions arise.

First, what shall we make of empirical considerations about metaphor processing, according to which, for example, the interpretation procedure for simpler metaphors is similar to that for literal utterances, while interpreting highly creative or novel metaphors involves a markedly different procedure [8]?[5] The dialectical situation is as follows: such considerations support or undermine, *independently of the continuum argument about interpreting metaphors,* the claim that a similar comprehension procedure applies to literal utterances and certain types of metaphorical utterances. The continuum argument doesn't—cannot—provide an objection to or further support for such claims, because (as I have tried to argue, successfully, I hope) if it were to work, it would show too much, so it doesn't work. Therefore the tenability of the claim about a literal–loose–metaphorical continuum and the application of *ad hoc* concept formation in the interpretation of metaphorical utterances will depend on *other* (experimental-data-driven) arguments.

Second, it is worth considering a positive proposal about how to supplement the relevance theoretic comprehension procedure for interpreting metaphors. I address this question in work in progress [11, 12], drawing in part on some of the considerations that provide missing details to supplement the comprehension procedure for interpreting malaprops and other linguistic errors (these were briefly discussed in Section 3). In the case of metaphorical utterances (but not malaprops), the speaker (and hearer) takes the lexically encoded concept associated with her words to be in force, and would not retract her words when confronted with the concept lexically encoded by her words. "The fog doesn't really walk on feline legs," someone might challenge the poet. And he might reply: "I was speaking metaphorically. But I stand by my words: The fog does come on little cat feet". By contrast, Mrs. Malaprop, when challenged, "There are no such things as pineapples of politeness," would (likely) respond: "I retract my previous words; I meant to speak about a *pinnacle* of politeness".[6] Such differences in the response to being challenged about the lexically encoded concepts associated with one's words do, I think, offer a promising starting point for the sorts of details that a relevance theoretic comprehension procedure might incorporate in an account of metaphor. Such an account would part ways with Sperber–Wilson's stance, claiming instead that there are, after all, interesting details and generalizations specific to metaphors. More generally, the various ways in which lexically encoded concepts systematically constrain speakers' meaning in the case of loose, hyperbolic and metaphorical utterances is a worthy area of inquiry within the relevance theoretic framework.[7]

[5] More recent experimental results [10] cast doubt on earlier views positing a marked difference in the processing of novel metaphors and literal utterances. Carston [9] a central figure of relevance theory parts ways with Sperber–Wilson [3] and posits two distinct modes of processing metaphorical utterances.

[6] See Camp [13, 14] about how deniability reveals distinctive features of metaphorical utterances.

[7] I have received many incisive comments in connection with this research project. I thank audiences and organizers at two conferences (Philosophy of Linguistics and Language X, in the Special Session on Dan Sperber és Deirdre Wilson's Philosophy of Language, Interuniversity Center Dubrovnik, September, 2014; Meaning and Experience (a conference in Hungarian) held at Kaposvár University, January 2015), one workshop (Metaphor workshop, in Hungarian, organized by the Hungarian Coaches Association and Erasmus Collegium's Language Research Group, Budapest, May, 2013), and two invited presentations (the MASZAT Colloquium Series (Round Table Society of Hungarian Semanticists), MTA Research Institute for Linguistics, Budapest, September 2014; the Institute of Philosophy of the Czech Academy of Sciences, Prague, Czech Republic, October 2014). Special thanks are due to Nicholas Allot, Tibor Bárány, Ágnes Bende-Farkas, Robyn Carston, Bálint Forgács, Hans-Martin Gaertner, Michael Glanzberg, Karen Lewis, Nenad Miscevic, Paolo Santorio, Adam Sennet; and, last yet foremost, Craige Roberts, Dan Sperber and Deirdre Wilson for invaluable comments, help and encouragement. The present research was supported by Grant No. K-19648, entitled Integrative Argumentation Studies, of the Hungarian Scientific Research Fund (OTKA) and travel funds from UiO CSMN.

REFERENCES

[1] D. Sperber and D. Wilson. *Relevance: Communication and Cognition*, Blackwell, Oxford, UK (1986) (2nd edition 1995).

[2] D. Wilson and D. Sperber. *Meaning and Relevance*, Cambridge University Press, Cambridge, UK (2012).

[3] Sperber, Dan – Wilson, Deirdre. A deflationary account of metaphors. In: *The Cambridge Handbook of Metaphor and Thought.* R.W. Gibbs (Ed.).Cambridge University Press, Cambridge, UK 84–105 (2008) (reprinted in Wilson–Sperber 2012).

[4] N. Allott. Relevance Theory. In: *Perspectives on Linguistic Pragmatics* (*Perspectives in Pragmatics, Philosophy & Psychology 2*). A. Capone, F.L. Piparo, M. Carapezza (Eds.). Springer, Berlin, Germany 57–98 (2013).

[5] D. Wilson and D. Sperber. Truthfulness and Relevance. *Mind* 111:583–632, 2002 (reprinted in Wilson–Sperber 2012).

[6] D. Sperber and D. Wilson. Pragmatics. In: *Oxford Handbook of Contemporary Analytic Philosophy*. F. Jackson, M. Smith (Eds.), Oxford University Press, New York, NY, US 468–501 (2005) (reprinted in Wilson–Sperber 2012).

[7] R. Carston. *Thoughts and Utterances*: *The Pragmatics of Explicit Communication*, Blackwell, Oxford, UK (2002).

[8] R.W. Gibbs, *The Poetics of Mind: Figurative Thought, Language and Understanding*, Cambridge University Press, Cambridge, UK (1994).

[9] R. Carston. Metaphor: Ad Hoc Concepts, Literal Meaning and Mental Images. *Proceedings of the Aristotelian Society* 110:295–321.

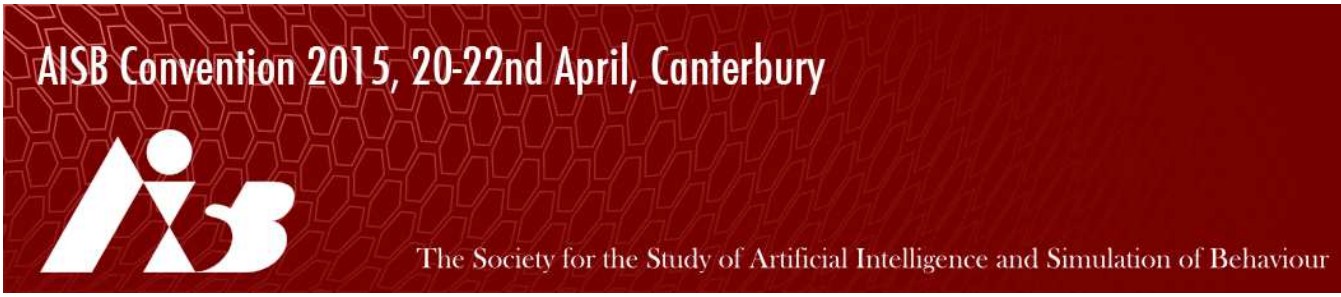
[10] B. Forgács, Á. Lukács and C. Pléh. Lateralized Processing of Novel Metaphors: Disentangling Figurativeness and Novelty. *Neuropsychologia* 56:101–109 (2014) doi:10.1016/j.neuropsychologia.2014.01.003

[11] Z. Zvolenszky. Inferring Content: Metaphor and Malapropism. Manuscript, Department of Logic, Institute of Philosophy, Eötvös Loránd University (ELTE), Budapest, Hungary (2015).

[12] Z. Zvolenszky. Why Inflate Sperber & Wilson's Account of Metaphor? Manuscript, Department of Logic, Institute of Philosophy, Eötvös Loránd University (ELTE), Budapest, Hungary (2015).

[13] E. Camp. Showing, Telling and Seeing. *The Baltic Yearbook of Cognition, Logic and Communication*, vol. 3, *Figure of Speech*, 1–24 (2008).

[14] E. Camp. Sarcasm, Pretense, and The Semantics/Pragmatics Distinction. *Noûs* 46(4):587–634 (2012). doi: 10.1111/j.1468-0068.2010.00822.x



AISB Convention 2015

University of Kent, Canterbury

Proceedings of the AISB 2015 Symposium on Embodied Cognition, Acting and Performance

Edited by Mark Bishop, Deirdre McLaughlin, Experience Bryon and Marco Gillies

ISBN: 978-1-908187-54-3

Introduction to the Convention

The AISB Convention 2015—the latest in a series of events that have been happening since 1964—was held at the University of Kent, Canterbury, UK in April 2015. Over 120 delegates attended and enjoyed three days of interesting talks and discussions covering a wide range of topics across artificial intelligence and the simulation of behaviour. This proceedings volume contains the papers from the *Symposium on Embodied Cognition, Acting and Performance*, one of eight symposia held as part of the conference. Many thanks to the convention organisers, the AISB committee, convention delegates, and the many Kent staff and students whose hard work went into making this event a success.

—Colin Johnson, Convention Chair

Contents

Image theatre and digital story-telling: Towards a research method called 'Collaborative Embodied Participant Analysis' (CEPA)

Michael Carklin[1]

This paper reports on research that I have been undertaking investigating the use of image theatre and digital storytelling with groups of university staff to gauge their thoughts, perceptions and experiences of the creative industries in higher education. In piloting these approaches as research methods specifically, I have been interested in comparing the responses that emerge from such active, participatory activities, which have ideas of embodied cognition at their centre, with the kind of material that emerges from focus groups and one-to-one interviews.

As an academic and manager within a faculty of creative industries, the overarching focus of my research has been on critically exploring aspects of the rise of this multidisciplinary field within higher education. In our current HE context, much credence is given to student voice; my concern with staff voice being marginalised or lost within institutional decision-making has led me to search for research approaches which might help to articulate the multiplicity of thoughts and views of staff. At the same time, such approaches help to address three key challenges: a) carrying out insider research in a faculty in which I also hold a management position; b) subverting the dominant language of the meeting room which is often filled with jargon and cliché; and c) contributing to encouraging dialogue and interaction amongst staff across disciplines within the faculty.

I have called the method that I am piloting 'Collaborative Embodied Participant Analysis' (CEPA), which has involved critically re-investigating each of these constituent terms. Practically, the method is initially rooted in an active, participatory drama-based approach known as image theatre, seeking to investigate how applying processes of embodied meaning-making and interpretation, linked to a heightened need for reflexivity by the participants, might lead to insights and perspectives that would differentiate this approach from other, more dominant research methods. At its core is the notion of collaboration in meaning-making, but also in interpretation and re-interpretation. Participants collaborate with other participants, but are also collaborative research partners to some degree. And fundamentally, this collaboration is carried out through a physical, embodied, drama-based process.

An extension of this work has involved digital story-telling in which, following a recorded interview in a meeting room, individual academic staff are recorded talking about their approaches to teaching and learning within the actual spaces that they normally teach in. This is useful in extending notions of embodiment through the linking of experience to place and space; investigating the impact of being physically present in a space on the ways those participants might think about and articulate their experiences.

In both cases – the image theatre and the digital storytelling – we are concerned with performance which demands a physical engagement and interaction. Whilst participants are not actors per se, there are levels of enaction and physical expression demanded which open up further possibilities for considering relationships between embodiment, experience and understanding. This paper highlights the ways in which the various facets of these activities might be qualitatively analysed and understood.

[1] School of Drama and Music, Univ. of South Wales, CF24 2FN, UK.
E-mail: michael.carklin@southwales.ac.uk

Stanislavsky's Mindful Actor: The System as a Guide to Experiencing Embodiment

Ysabel Clare[1]

ABSTRACT. This paper proposes that embodiment, ostensibly the subject of the second part of Stanislavsky's actor training course [1] [2] actually forms the experiential foundation on which the first [3] [4] is based, and provides the framework and the terms of reference around which the whole is designed. Discovering how this framework underpins the work elucidates meaning by exposing conceptual and actual relationships between experiencing and embodiment, opening up new possibilities for the understanding and thus the practice of both.

The concepts of *Perezhivanie* (experiencing) and *Voploshchenie* (embodiment) are central to Stanislavsky's work. Both resist verbal description, definition or explanation. Stanislavsky has addressed this problem with considerable strategic ingenuity in his fictionalized training diaries. Examining how he did so provides practical insights into how to recognize, learn, teach, and facilitate embodiment.

Research comprising detailed analysis of the action outlined in these texts has uncovered complex narrative patterning evidencing underlying conceptual constructs that, once revealed, clearly articulate an embodied experiential framework. The most complete text, *An Actor Prepares*, is not just a series of exercises with justifications and explanations, but a subtle and nuanced sequence of actions and effects (in a Socratic, dialogic form of exercises and responses) cleverly engineered to deliver a systematic encounter with an orderly underlying model of subjective (and necessarily embodied) experience.

This implicit conceptual framework both originates in and is a re-presentation or projection of human experience. Original diagrams are supplied that in turn re-present the deep structure of Stanislavsky's model in its own terms, graphically illustrating its roots in embodiment. These demonstrate the irrevocable conceptual links between the core concepts of *Perezhivanie* and *Voploshchenie*, showing how they can be operated to create and maintain a stable, coherent state in which the actor is dynamically experiencing embodiment: mindful - 'in the moment'.

Stanislavsky's underlying model is consistent with an experiential realist view such as that of Lakoff and Johnson [5]. While superficially different, it also shares deep structure with other contemporary frameworks for understanding human process, such as those of Pinker [6], Damasio [7] and Fauconnier [8]. Stanislavsky, however, shows us how to manipulate the phenomena of human process deliberately, at will. While language might not serve his purpose, and he cannot actually give the reader of his books an embodied experience, he does the next best thing by cleverly engineering the form and the narrated events. In this reading, results are as important as exercises, for the patterns in which the fictional students' responses occur express essential aspects of embodied experience that otherwise resist description.

In conclusion, the paper asserts that despite the passing of time, Stanislavsky still has something to contribute to actor training in the 21st Century because he offers practical strategies for actors to learn, manage and manipulate their embodied experience for the purpose of mindful performance.

REFERENCES

[1] K. Stanislavsky. *Building a Character,* trans. E. Reynolds Hapgood, Elizabeth, Methuen, (2008).

[2] K. Stanislavsky. *An Actor's Work Part II,* trans. Jean Benedetti, Routledge, (2008).

[3] K. Stanislavsky. *An Actor Prepares,* trans. Elizabeth Reynolds Hapgood, Methuen, (2008).

[4] K. Stanislavsky. *An Actor's Work Part I,* trans. Jean Benedetti, Routledge, (2008).

[5] G. Lakoff, & M. Johnson. *Metaphors We Live By*, University of Chicago Press, (1980).

[6] Steven Pinker,*The Stuff of Thought: Language as Window into Human Nature*, Penguin, (2008).

[7] Antonio Damasio, *The Feeling of what Happens : Body and Emotion in the Making of Consciousness*, Vintage, (2000).

[8] Gilles Fauconnier, *Mental Spaces : Aspects of Meaning Construction in Natural Language*, Cambridge University Press, (1998).

Better Than Life; testing techniques for an online audience to influence and participate in a live performance

Nicky Donald, Marco Gillies

Abstract. This work introduces the mixed reality show Better Than Life, testing techniques for an online audience to influence and participate in a live performance.
This show combines aspects of online multiplayer game, live theatre and reality television. Participants described it as immersive theatre, Alternate Reality Game (ARG) and Live Action Role Play (LARP).
The aim is to provide a set of interaction mechanisms for the online users to affect the storyworld in real time, alongside the data gathering and analysis tools to assess the ludic/narrative effectiveness and user experience of those mechanisms.

1 INTRODUCTION

Goldsmiths worked with Coney, who make live games and Showcaster, who stream live events. Coney created a storyworld of a cult built around the clairvoyant Gavin, testing new recruits for psychic abilities. These tests were designed by Pan Studios, who worked alongside magician Jon Armstrong to create two classic pieces of stage magic, a substitution and a disappearance. The participants in the live studio space were the recruits, and the online participants were tasked with choosing which of them would lead the cult on Gavin's departure.

The interactions had to form a seamless part of the narrative; the user interface had to enable a feeling of participation in the live event and the storyworld. This meant a very fast signup and entry process, so users could start watching and interacting very quickly. This in turn meant that we had to gather user data on-the-fly with short simple questionnaires that didn't detract from the flow of the live experience. We also looked to engender a feeling of presence and embodiment through multiple avenues:

- Online users could navigate the space by switching cameras at will, discovering additional locations, actors and scenarios.
- They could influence the action by means of live chat and mouse movement.
- They could chat to each other (often sharing things seen on other cameras) to actors (influencing their script and costume) and to live participants after the show (piecing together a shared picture of the show).

2 LATENCY

At two points online users' movements were captured. During a group breathing exercise, mouse movements became a DMX value controlling the brightness of lights in the real space. In the finale we projected a spot of light that embodied each user, moving as they moused over the video. In both instances, we were faced with an extremely variable system latency, i.e. the interval between an event in the live space and its appearance in the viewed video feed at remote locations was completely unpredictable. This was down to several factors:

- The commercial servers used to stream the data were under pressure from the World Cup and Wimbledon 2014 coverage and associated live streams.
- The commercial infrastructure (ISPs) delivering data to users was under similar pressure
- Users were viewing through a variety of domestic, office, public and academic connections rated at differing speeds
- Users were using a variety of devices and platforms, from hard-wired desktop machines to handheld devices operating on wifi.

This meant that the gap between an online users movement and the resulting scenographic change was subject to a similar delay. When we asked for concerted action, the input was spread over a period of between 10 and 60 seconds.

3 PERFORMERS

The actor playing Gavin was improvising constantly, incorporating input from the online users, addressing the online and live groups individually and simultaneously, maintaining a complex narrative with other actors and performing a vanishing act.

4 DATA

We collected a very large data set from 70 live and 262 online participants over 8 shows in a three-week period. Some of this data has to be animated since it is too complex and multidimensional for conventional visualisation. The initial findings are that the online experienced a growing social presence and collective agency, a sense of sharing and doing, which exceeded that of the live participants.

5 CONCLUSIONS & FUTURE WORK

Going forward, we want to model much larger user numbers and implement the technology in adventure games for heritage sites that connect small groups of visitors with large numbers of online participants in homes and classrooms.

The Cognitive Dynamics of Performance Generating Systems: Deborah Hay through Christopher House

Pil Hansen[1]

Abstract. Performance generating systems are rule- and task-based dramaturgies that systematically set in motion a self-organizing process of dance or theatre creation. The resulting performance is not generated from the performers' impulses or choices, as in the case of improvisation, but rather from the ways in which a system directs, limits, and adapts the performers' conscious attention, perceptions, and interactions. At present we are unable to archive and remount these systems. The reason is that a valid blueprint needs to capture the dramaturgical and cognitive principles through which the systems generate performance instead of recording the actual performance that is danced or acted.

Performance Generating Systems (an international research project hosted by the University of Calgary) seeks to develop a tool for dramaturgical analysis and notation based on Dynamical Systems Theory; a tool that enables dramaturgs and scholars to script the most relevant components of performance generating systems and the dynamics of interaction and perceptual manipulation they generate. This paper will outline the project and present its first case study, the DST analysis and notation of Christopher House's (Toronto Dance Theatre) adaptations of Deborah Hay's solo performance scores.

Expressed in cognitive terms, Hay's scores and praxis challenge the performer to continuously and consciously register a larger and less selective amount of perceptual stimuli than normative cognitive processing involves. Otherwise implicit reliance on memory in the present is inhibited and replaced with attempts to avoid accumulation, patterning, sequencing, anticipation, and other forms of recycled movement responses to stimuli. The task is impossible; self-organizing movement patterns are attracted over time, yet the attempt results in a differently earned presence.

Hay articulates her praxis as a belief system. Thus one of my main challenges when applying DST to the work of Hay and House is to honour both their vocabulary and my observations of the embodied cognition of this praxis, while using DST to distill principles of performance generation that can be transferred between artists over time. In other words, I am negotiating the interdisciplinary positions of specificity versus generalization in search of an operational, and dramaturgically productive, compromise.

1. School of Creative and Performing Arts, University of Calgary, CA. pil.hansen@ucalgary.ca

Acted Emotion: a performance experiment in psychology and actor training

David Jackson

Robert Harnish's 'narrow construal' of cognitive science envisages the mind as a kind of computer, a model that is closely related to efforts to build artificial intelligence. His 'broad construal' expands this definition to incorporate, in addition to computer science, philosophy, anthropology, neuroscience, psychology and linguistics. Each of these disciplines approaches the human mind from a different perspective, generating a wide range of theoretical models. It seems that the mind is such a complex topic that efforts to understand its workings tend to transcend the limits of a single discipline. Recent studies of brain architecture suggest that it is not just disciplinary boundaries that are collapsing under the weight of new discoveries. Hard and fast distinctions between cognition and emotion are also under threat. The work of neurologists Joseph LeDoux and Antonio Damasio demonstrates not just the interaction of thought and emotion, but also the role of the body in both cognition and feeling. Emotion and the body, therefore, must be welcomed into the fold of cognitive studies.

The explosion of interest in emotion research over the last twenty years has generated a host of ground-breaking accounts which place the emotional process in an ecological and somatosensory context. Moreover, new technology and research methods have developed to facilitate the investigation and understanding of the topic. These developments create an ideal climate for a reassessment of the specialised function of emotion in acting and performance and for addressing some key questions of long-standing theoretical and practical interest: what is the nature of acted emotion? Is it different from spontaneously occurring emotion? Do actors feel the same emotions as their characters?

This paper offers a timely response to this propitious moment for addressing the area of acted emotion. It outlines some of the influential theories that dominate discourse in the scientific and performance research communities, thus establishing the context for an investigation of the topic. I describe an innovative 'Performance Experiment', a term

which deliberately combines the language of science and performance. Using video documentation, I discuss the experiment in terms of its two principle aims: firstly, comparing two strategies for arousing and expressing emotion, (Method Acting and Alba Emoting) and secondly, integrating research methodologies drawn from psychology and actor training. Student actors engage with a series of exercises and I assess their impact using a range of techniques, including both self-report and external observation.

Finally, I present the results of the data analysis and consider a number of related questions: which technique has a greater impact on the actors from a phenomenological, physiological and observer's perspective? Is there a difference between actors' perception of their emotions and the unconscious evidence provided by the body? Can such interdisciplinary investigation bring us closer to an understanding of the nature of acted emotion? Can performance practice inform science as much as science can inform performance practice?

Enacting Desire: Constructing Social Flexibility through Somatic-informed Processes

Thomas Kampe (PhD)[1]

This paper discusses the facilitation of actor training as a holistic education effected through somatic-informed processes of embodiment. It will draw on applications of the work of Moshe Feldenkrais (1904-1984), understood as an enactive and ecological model of reflective self-creation through movement, within actor training contexts.

It examines the construction of a Feldenkrais-informed educational practice which draws on Feldenkrais' practices of 'Awareness through Movement' and 'Functional Integration'. In addressing the themes of this conference, this paper considers Doidge's (2015) writings on neuroplasticity which places Feldenkrais' non-dualist practices, within the development of 'flexible minds' (Feldenkrais 2010) and at the forefront of learning approaches that use embodiment as a vessel for transformation of brain functions.

The paper places Feldenkrais as a radical pioneer within the enactivist paradigm, whose practical educational modalities empower learners to access possibilities for 'self-education'(Feldenkrais 1992). This includes a heightened self-awareness and expressive potentiality, and an emerging 'Enactive Social Understanding' (Di Paolo et al 2014:60) of their lived environment. Feldenkrais developed a use of 'self-imaging' (Beringer 2001) within his practice, which is multi-modal and synergistic. It includes verbal and sensory imagery, motor-imagery as in imagining movement without moving, and an 'enactivist approach to imagery' (Thomas 2011) where sensation and image are generated through movement and self-observation in interaction with the material and social environment.

More so, this paper explores the probing of underlying assumptions and principles informing the above practices as modes for an embodied co-creation of the actor as a flexible, relational and desiring social creature. It examines Feldenkrais-specific notions of thoughtful-doing as felt-embodied enquiry, Feldenkrais' use of touch-interaction as a questioning of the cognitive closure of the human being, his 'theory of reversibility'(Feldenkrais 2010), and his eco-proposition of a 'functional unity between body, mind, and environment' (2005:149) - environment understood as a bio-psycho-social structuralisation - as departure points for creative pedagogic inquiry.

Ultimately, this paper argues for a construction of a Feldenkrais-informed practice as a critical, transformative and emancipatory pedagogy which questions hierarchical and reductionist modes of actor training. It suggests that such pedagogy supports a co-enactive process of 'organic learning' (Feldenkrais 1981) that facilitates conditions for shared artistic inquiry.

Performance theorist Gesa Ziemer asserts that such conditions for embodied inquiry 'where linguistic eloquence is being slowed down, where we are disoriented and touched at the same time to perceive something'(2009) are socially transformative and empowering for the participants. In his discussions with Richard Schechner, Feldenkrais (2010[1972]) proposes that such slowing-down enables the actor to engage in 'an awareness of action' which fosters 'greater clarity and ease', a capacity to 'listening to the other person' and the possibility for 'rediscovery', which for Feldenkrais provides the potential towards a bio-psycho-social flexibility and the forming of new behavioural patterns.

The paper suggests that an acquired Feldenkraisian flexibility which includes psycho-social competencies and a heightened ability of a 'learning to learn'(Feldenkrais 2010), supports the student actor in their personal, creative and professional development. The author draws on practice-led research, his own pedagogical practice, and on student-feedback from the BA Acting program at Bath Spa University, while referring to writings by Moshe Feldenkrais, social-theorist Cornelius Castoriadis(2005,2011) and ecologist Edgar Morin (1999;2007) - all three informed by the paradigm of Autopoiesis (Varela 1995,1999).

References

[1] Bales, M.& Nettl-Fiol, R., eds. (2008) *The Body Eclectic – Evolving Practices in Dance Training*, Illinois: University of Illinois Press (2008)
[2] Beringer, E. 'Self Imaging'; in*: Feldenkrais Journal* 13; pp. 33–38 (2001).
[3] Blair, R. *The Actor, Image, and Action: Acting and cognitive neuroscience*. New York: Routledge (2009).
[4] Castoriadis, C. *Postscript on Insignificance: Dialogues with Cornelius Castoriadis;* London: Continuum(2011).
[5] ---*.Figures of the Thinkable;* http://www.notbored.org/FTPK.pdf (2005)
[6] ---.*The Imaginary Institution of Society*; Cambridge: MIT Press (1998).
[7] Evans, M. *Movement Training for The Modern Actor*; London: Routledge(2008).
[8] Doidge, N*The Brain's Way of Healing;* London: Penguin(2015).
[9] Feldenkrais, M. *Embodied Wisdom: The Collected Papers of Moshe Feldenkrais*, Berkeley: North Atlantic Books ; with Beringer, E.(ed) (2010) .
[10]---. *Body and Mature Behaviour: A Study of Anxiety, Sex, Gravitation and Learning*, Madison: International Universities Press (Second edition) (2005).
[11] ---. *The Potent Self*. Berkeley: Somatic Resources (Second edition) (2002)
[12] ---. *Awareness Through Movement: Health Exercises for Persona Growth*, New York: Harper & Collins (Second edition) (1992).
[13] ---.*The Elusive Obvious*. Cupertino, California: Meta Publications (1981).
---.*Feldenkrais on thinking independently from words*
[14] http://www.youtube.com/watch?v=1V_5O7KANWI(1981a).
[15] ---. *The Case of Nora*; New York: Harper and Row (1977). [16] Feldenkrais, M.& Schechner R. *Image, Movement, and Actor: Restoration of Potentiality;* http://www.feldenkraismethod.com/wp

[1] **Department of Performing Arts ; Bath Spa University;**

email: T.Kampe@bathspa.ac.uk

content/uploads/2014/11/Image-Movement-and-Actor-Moshe-Feldenkrais.pdf [accessed 20/02/2014] (2010)
[16] Ginsburg C. Body-Image, Movement and Consciousness: Examples from a Somatic Practice in the Feldenkrais Method in *Journal of Consciousness Studies*. 6(2-3): 79-91(1999).
[17] ---. *The Intelligence of Moving Bodies: A Somatic View of Life and its Consequences*; Albuquerque: Awareing Press(2010).
[18] Kampe, T.Eros and Inquiry: The Feldenkrais Method as a Complex Process; in: *Theatre, Dance and Performance Training*, Vol. 6.2; London: Taylor and Francis (2015).
[19] ---. The Art of Making Choices: The Feldenkrais Method as a Soma Critique, in: Whatley, S., Garrett-Brown, N., Alexander, K. (2015) *Attending to Movement: Somatic Perspectives on Living in this World*; Axminster: Triarchy (2015a).
[20] ---.*The Art of Making Choices: The Feldenkrais Method as a Choreographic Resource*; PhD documentation; London Metropolitan University (unpublished)(2013).
[21] ---. Recreating Histories: Transdisciplinarity and Transcultural Perspectives on Performance Making, *The Korean Journal for Dance* Vol 67 (2011).
[22] Klein, G. & Noeth, S. (eds.) *Emerging Bodies: The Performance of Worldmaking in Dance and Choreography;* Bielefeld: Transcript (2011*)*.
[23] Morin, E. *On Complexity*, Cresskill, NJ: Hampton Press (2007).
---.*Seven Complex lessons in education for the future*, Paris: Unesco Publishing (2001).
[24] Steward, J.& Gapenne, O.& Di Paolo, E.A. *Enaction: Towards a New Paradigm for Cognitive Science*; Cambridge Mass: MIT Press (2014).
[25] Varela, F. Large Scale integration in the Nervous System and Embodied Experience; in: *Report -1st European Feldenkrais Conference;* pp.12-14 ; Paris: International Feldenkrais Federation (IFF) (1995).
[26]---. 'Neurophenomenology : A methodological remedy for the hard problem, Journal of Consciousness Studies', *Special Issues on the Hard Problems*, with J.Shear (Ed.) (1999).
[27] Varela, F, Thompson E, & Rosch E. *The Embodied Mind: Cognitive Science and Human Experience* , Cambridge, MA: MIT Press(1991).
[28] Ziemer, G. Was kann die Kunst? Forschen anstatt wissen In: *Zwölf . Die Zeitschrift der Hochschule für Musik und Theater (Hg.*).(5). http://www.gesa ziemer.ch/pdf/Was_kann_die_Kunst.pdf [accessed 12/06/12] (2009).

Watergait:

Designing Sense Perceptions for Individual Truth

Esthir Lemi, Marientina Gotsis, Vangelis Lympouridis[1]

Abstract. *Watergait* is an experimental meditation in the form of a sonified experience of walking with shoe sensors that translate shifting foot pressure into sound within an aural environment. This experiment was collaboratively designed by three artists, Esthir Lemi, Marientina Gotsis and Vangelis Lympouridis, influenced by different yet complementary theoretical, aesthetic, and technical domains. The quintessential adage for all three artists is best summarized by the sentence: "all sense perceptions are true" and a mutual adoration of water-related themes and design minimalism. Perhaps not by coincidence, our mutual ethnic backgrounds kept bringing us back to implicit knowledge and shared context of history and experiences that informed our design and pre/post discussion of the experiment.

In this essay, we explore Epicurean tradition, holistic design models, empirical dialectic systems, historical uses of water as a playful theme, and its implications in human computer interaction. The instrumentation of Watergait depends on some "objective" truths that had to be measured and be agreed upon. The sensing array of the shoes measure pressure. Placed right below the insoles and imperceptible to the wearer, the pressure sensors send data to the computer via Bluetooth®technology. What follows is a philosophical perspective of design on how sensing and art intersect through human-computer interaction, and why some contextual bridge between the two is needed.

The ancient Greek philosopher Epicurus advocated for the awakening of the senses through mindful observation of the felt and sensed experience (Letter to Menocoes and Herodotus). For Epicurus, relative and absolute truth can coexist while trying to make sense of the world from a human-centered point of view as he presents one of the first integrative viewpoints of psychology and perception, placing value in how belief influences perception and thus introducing the placebo effect as a quantifiable unknown that produces an effect and contributes to one's own perception of reality. This type of discourse is legitimized through everyday habits toward the pursuit of happiness. While manufacturing happiness, or pleasure, does it matter what the signal is or does it matter more what it is being perceived as, or does it matter at all? We, the artists of Watergait wanted to immerse participants into a simple narrative fantasy through the aural environment and to enable them to follow a path that can excite their imagination through the senses. Making the apparatus simple makes it more prone to several interpretations, and therefore more successful to stimulate the imagination.

Lastly, we discuss the manifested coincidence of summoning our mutual "otherworldly" experience within water: an encounter with whales, which started in the virtual and happened in real life.

[1] Creative Media & Behavioral Health Center, Interactive Media & Games Division, USC School of Cinematic Arts, 900 W 34 ST, SCI 201U, Los Angeles, CA, USA. Email: (Esthir Lemi) lemi@esthir.info (Marientina Gotsis) mgotsis@cinema.usc.edu (V. Lympouridis) vangelis@lympouridis.gr

Participatory enaction of music: Key points towards radicalizing the notion of embodiment in music

Juan Loaiza 1

Cognition -sense-making- is an affective-laden activity that takes place across 'brain, body, and environment' [1], [2], [3]. Strong naturalistic –yet non reductive- claims about the continuity between life and cognition distinguish Enactivism from other theories and implementations of the notion of embodiment [4]. Enactivism, in 'Varela style', rejects the received view of the body as contingent and the social environment as contextual, a view still held by 'mentalists' conceptualizations of embodiment. Enactivism, on the contrary, sees the body as continuously constitutive of sense-making processes [2], [3], [4], [5] and social interaction as the domain where 'higher cognitive processes' –such as linguistic use- take place [6].

The presentation will explore the idea that an enactive notion of music embodiment needs to be qualified by the introduction of a more precise –naturalized- definition of (social) Participation. This definition requires repositioning the level of analysis within the social encounter. Enactivism offers a refined account of participatory sense-making that does not reduce cognitive processes to the aggregate of pre-given individual agents; moreover, it offers an understanding of interactions as autonomous and generative in their own terms [6], [7], [8]. Thus, starting from a social level of analysis, Musicking (a term coined by Small 1998 [9]) is rethought as a class of enactive participation vis-à-vis other participatory genres such as Languaging.

The presentation will expand the discussion with some contrasting points:

"Biographical" vs. "snapshot": Critique to narrow time scales. The snapshot-like, laboratory approach to understand musical activity makes it easier to assume the individual experience as paradigmatic. In contrast to this, an ecologically valid approach brings to the foreground an agent's history of social relationships and patterns of participation.

Enactive organisms vs. "epistemic minds": Critique to mentalist and skull-bound explanations of cognition. Accounts of musical experience often portray the individual finding herself as if left in the middle of an opaque environment that has to be disentangled via mental epistemic moves. Enactivist approaches, in contrast, dis-localize cognition emphasizing the co-constitution of active autonomous organisms and its medium via sense-making.

Complex and adaptive vs. "tidy" ordered systems: Critique to linear approaches to musical interaction. Theorizations and practices often rely on tight modeling and prediction; these however lack the flexibility to address social dynamics. Interactions may be better understood within its own emergent normativity and relative autonomy.

The presentation will bring to the table Enactive notions that stretch beyond the sensorimotor approach to music cognition, namely: agency, autonomy, emergence, identity, sense-making.

REFERENCES

[1] Varela F.J, Thompson E, Rosch E. 1991, *The embodied mind :cognitive science and human experience.* (Cambridge, Mass.; London: MIT Press).
[2] Thompson, E. 2007, *Mind in life: Biology, phenomenology, and the sciences of mind,* Harvard University Press, Cambridge.
[3] Cappuccio, M. & Froese, T. 2014, "Introduction" in Enactive Cognition at the Edge of Sense-Making: Making Sense of Non-Sense., eds. M. Cappuccio & T. Froese, Palgrave Macmillan, Basingstoke, UK, pp. 1-33.
[4] Kyselo, M. & Di Paolo, E. 2013, "Locked-in syndrome: a challenge for embodied cognitive science", Phenomenology and the Cognitive Sciences, pp. 1-26.
[5] Di Paolo E, Rohde M, De Jaegher H. 2010, "Horizons for the enactive mind: Values, social interaction and play". In: Stewart J.R., Gapenne O, Di Paolo E, editors. Enaction: Towards a new paradigm for cognitive science (Cambridge: MIT Press;).
[6] Cuffari E, De Jaegher H, Di Paolo E. 2014, "From participatory sense-making to language: there and back again". Phenomenology and the Cognitive Sciences, Online 19 Nov (2014).
[7] Froese, T. & Di Paolo, E. 2011, "The enactive approach: theoretical sketches from cell to society", Pragmatics & Cognition, vol. 19, no. 1, pp. 1-36.
[8] De Jaegher H, Di Paolo E. 2007, "Participatory sense-making". Phenomenology and the Cognitive Sciences Vol 6, No. 4, p. 485-507.
[9] Small C. 1998, *Musicking : the meanings of performing and listening.* (Hanover: Wesleyan/University Press of New England;).

1. The Sonic Arts Research Centre, Queen's University Belfast. UK, International and postgraduate scholarship, email: jloaizarestrepo01@qub.ac.uk

The Embodied Brain: An Argument from Neuroscience for Radical Embodied Cognition

Julian Kiverstein[1], **Mark Miller**[2]

Abstract. In this programmatic paper we develop an account of embodied cognition based on the inseparability of cognitive and emotional processing in the brain. We argue that emotions are best understood in terms of action readiness [1, 2] in the context of the organism's ongoing skillful engagement with the environment [3, 4, 5]. States of action readiness involve the whole living body of the organism, and are triggered by possibilities for action in the environment that matter to the organism. Since emotion and cognition are inseparable processes in the brain it follows that what is true of emotion is also true of cognition. Cognitive processes are likewise processes taking place in the whole living body of an organism as it engages with relevant possibilities for action.

1 Introduction

Our aim in this paper will be programmatic. We propose a definition of embodied cognition based on the inseparability of emotional and cognitive processes in the brain [6]. Our argument has the following three steps:

(1) Cognition is embodied because cognition and emotion are inseparable processes in the brain.
(2) Emotion is a dynamic process involving the organism's whole body.
(3) From the inseparability of emotion and cognition in the brain it follows that cognition is likewise a dynamic process involving the organism's whole body.

We align ourselves with proponents of radical embodied cognition in endorsing the non-decomposability of the brain-body-environment system. We take this thesis to be implied by the functional integration of emotional and cognitive processing in the brain. We show how recent research concerned with large-scale patterns of connectivity in the brain challenges a decompositional analysis of the brain into regions and components that carry out either emotional or cognitive psychological functions. The current evidence points instead to a theory of brain processes as complex, non-linear, self-organizing processes composed of "intricately interconnected, interacting elements" [7]. We find interconnection, interaction and mutual influence among components (or neural regions) resulting, we argue in processes that are simultaneously both cognitive and emotional.

How can we make an argument from the non-decomposability of cognitive and emotional processes within the brain to the non-decomposability of the larger brain-body-environment system? We begin by providing a tweak to psychological constructionist theories of emotion which interpret the integration of cognitive and emotional neural processes in terms of interactions between domain general neural networks [8]. We suggest (following arguments developed by Luiz Pessoa [9]) that structure-function mappings are not fixed and static properties of networks. Instead structure- function relationships are dynamic, with the functions a given network performs varying over time in a context-dependent manner. It is the latter finding which we take to support the non-decomposability of the brain-body-environment system. To determine the precise functional contribution a network is making to behavior requires zooming out, and having in view the whole organism in its interaction with the environment. Emotional-cognitive processes don't only take place inside of brains, but are processes that involve constant interaction between the brain and the whole living body of the organism in an ecological setting.

The first two steps in our argument establish the inseparability of emotion and cognition in the brain and the deep dependence of emotional processes on the whole body of the living organism in its practical skilled engagement with the environment. We take these two steps to imply a third step: the conclusion that cognitive processes depend on the whole living body in its practical and skilled engagement with an environment of affordances.

[1] Institute for Logic, Language and Computation, Univ. of Amsterdam, Netherlands. Email: J.D.kiverstein@uva.nl
[2] Dept. of Philosophy, Univ. of Edinburgh, UK. Email: s1033091@ed.ac.uk

REFERENCES

[1] Frijda, N. H. The emotions. Cambridge: Cambridge University Press. (1986)
[2] Frijda,. N. H. The laws of emotion. Mahwah: Erlbaum. (2007)
[3] Reitveld, E. Situated normativity: the normative aspect of embodied cognition in unreflective action. Mind 117 (468): 973-1001. (2008)
[4] Bruineberg, J. & Rietveld, E. Self-organization, free energy minimization, and optimal grip on a field of affordances. Frontiers in Human Neuroscience 8: 599. (2014)
[5] Kiverstein, J. & Rietveld, E. The primacy of skilled intentionality. Philosophia. (Forthcoming, 2015)
[6] Pessoa, L. The cognitive-emotional brain: from Interactions to Integration. Cambridge, MA. MIT Press. (2013)
[7] Colombo, M. Moving forward (and beyond) the modularity debate: a network perspective. Philosophy of Science 80 (3): 356-77. (2013)
[8] Barrett, L. F. & Satpute A. B. Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. Curr Opin Neurobio, 23(3): 361-72. (2013)
[9] Pessoa. L. Understanding brain networks and brain organization. Physics of Life Reviews 11: 400-35. (2014)

Stanislavski's System and a Dual-Process Approach to Performer Training

Grant Olson[1]

Abstract. Konstantin Stanislavski's (1863-1938) development of actor training and performance methodology, which he called 'the system', has significantly shaped modern performance theory and practice. Stanislavski was keenly aware that a majority of human experience was shaped by processes not normally available to what he understood as conscious thought. Stanislavski was particularly interested in the role non-conscious processes could be harnessed to achieve his goal of reaching *perezhivanie,* or experiencing through a role. Subsequently, as he developed his approach towards actor training and rehearsal methodology he aimed to access what he considered the unconscious through conscious preparatory methods. In Stanislavski's understanding, he further divided the unconscious into a subconscious equated with instinct, and a superconscious that he associated with intuition. Most of what is currently understood of as intuition finds support in a dual or multiple processor theory of cognitive analysis. William James (1842-1910) first predicted a concept of a multiple or dual processing system in *Principles of Psychology* (1890) wherein he proposed one system of rational thought or true reasoning, and another devoted to impulse or associative thought. As advances in Cognitive Studies have increased understanding in cognitive function, a consensus has emerged of an acceptance of a dual or multiple processing systems divided between the so-called System 1 or fast and intuitive processes, and the System 2 slower analytical processes. This paper identifies several of Stanislavski's theories showing potential correlations with current understandings related to dual-process theories. In addition, I propose several approaches found in Stanislavski's methodology that hold potential to develop an actor's System 1 processing abilities as related to intuition in performance. Moreover, I identify gaps in Stanislavski's system that could benefit from alterations in methodology that would bring his approach in line with current understandings of the dual processing theory of cognitive function. Incorporating approaches aimed at developing the System 1 or fast processing system of cognition into the methodology of performer training and practice holds potential into strengthening performer skills once relegated to the numinous realm of an actor's intuition. In addition, further insight gathered from the performance situation offers a greater understanding of the role emotional response plays in judgement formation and cognitive function.

[1] Faculty of Arts and Social Sciences, Kingston University, KY1 2EE, UK. Email: k1065389@kingston.ac.uk

Attempts on Margarita (multiple drafts): A cognitive dramaturgy generated by voice and space

Christina (Xristina) Penna[1]

Abstract. In the dynamic contemporary theatre and performance landscape of 'immersive', hybrid and interactive production where the boundaries between public and private, performance space and audience space intertwine, alternate or even disappear, scenography is referred to as a process. [1] [2]

The above observation poses a series of questions regarding the critical frameworks that could be used in order to analyse scenography as process and the methods that might be employed to contribute to the creation of dynamic scenographic landscapes where the audience becomes an active co-author of the work.

Through my practice-led research at the University of Leeds I am suggesting a method of staging dynamic scenographic systems using current cognitive theories of consciousness (Baars, Dennett, Edelman and Tononi). These performance-systems engage with the concepts of process, integration of information and complexity inviting the participants to interact in a dynamic bottom-up way with the work.

In the piece 'Work Space I- a scenographic workshop on consciousness' I appropriated Baars' diagram of consciousness known as the Global Workspace [3] to create a workshop-installation in which the participants are invited to share the experience of a performance-game and contribute to the hands-on creation of a multi-authorial artwork.

By reflecting on the above work, which draws and explores the notion of embodiment and the 'socially collaborative, culturally and materially grounded nature of the human mind' [4] I focused on the 'dialogue through making' that occurred during the time of the workshop.

In another practice-led investigation ' Work Space II - Attempts on Margarita (multiple drafts)' I am drawing from Martin Crimps' *postdramatic* work 'Attempts on her Life' and the cognitive theories of consciousness by Dennett, Edelman and Tononi in order to create a multi-layered cognitive dramaturgy in the form of an installation space. A current view on the hard problem of consciousness, largely initiated by neuroscientist/psychiatrist Giulio Tononi, is that 'wherever there's information processing, there's consciousness' [5] In the piece 'Attempts on Margarita (multiple drafts)' aim is to generate a *collective consciousness* in the form of a durational, sound installation by mixing information such as pre-recorded and live - stream voices generated by three types of participants:

- P1: a) Friends/colleagues/acquaintances of mine and b) random passers-by in the university campus who answer the same set of questions regarding 'Margarita'.
- P2: Participants-audience who attended the installation.
- P3: A group of artists working with sound, devising and objects in the main installation space.

In this paper I will focus on the post-show discussion with the participant artists (P3) on their experience of the installation. I will refer to their comments of their experience as 'a reflective space' and of the 'ethics that can be established by a space'. Using as critical framework enactive cognitive science and the ideas of an ecologically extended and socially engaged mind I will then try and analyse this multi-layered process scenography.

REFERENCES

[1] A. Aronson, *Looking Into the Abyss: Essays on Scenography,* Ann Arbor: University of Michigan Press, 2005.

[2] J. McKinney, *The nature of communication between scenography and its audiences*. Ph.D dissertation, The University of Leeds, 2008.

[3] B. J. Baars & S. Franklin, 'An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA', *Neural Networks,* **20**, 955-961, (2007).

[4] C. Sinha, 'Blending out of the background: Play, Props and Staging in the Material World', *Journal of Pragmatics,* **37**, 1537-1554, (2005).

[5] G. Tononi, 'An Information Integration Theory of Consciousness', *BMC Neuroscience,* **5**, 42 (2004). http://www.biomedcentral.com/1471-2202/5/42 accessed 14/11/14.

[1] School of Performance and Cultural Industries, University of Leeds, UK. Email: pccp@leeds.ac.uk.

Extended Body in the Telematics Performance: the perceptual system of remote dancers

Ivani Santana[1]

[1] University of Bahia, Brazil, email: ivanisantana.mapad2@gmail.com

Abstract. This article discusses some artistic proposals of Telematic Performances that explore the relationship of human bodies in distributed dances. Besides several titles of this dance configuration that we can find in the bibliography of this field, we prefer to keep the traditional designation of telematic dance to configure the composition of movements created with bodies (subjects) distributed through discrete nodes from a network and that interact with each other in some level in real time. The dancers can be distant from each other, or be in the same room, what matters is that communication between the nodes must be performed over the network. The nature of this field is a fluid reality, a constant transformed environment where the human being (i.g. the dancer) and her/his milieu (i.g. the network) are co-evolutive, co-dependent and mutually implicated. This article is grounded on the concepts of "Extended Mind" and "Cognitive Artefact" [3, 4], "Actionism" [6, 7] and "Body Image" and "Body Schema" [5]. This approach contributes to re-think the notion of (tele)presence, time, space, distance and of the body (the self). Through these comprehensions we should understand the new perceptual demands from which dancers have to deal with in the context of telematic dance. This new art configuration promotes different sensorimotor experiences than the stage-based dance environment, it affords different skills to the dancers, because this is a transformative art that validates this fluid reality. The discussion of telematic dance will be made through the analyses of the project *"EVD-58 / Embodied Varios Darmstadt 58"* which was created in collaboration with artists from Mexico and Spain (2013), and Portugal and Chile (2014). EVD-58 was created to develop the concept of "(tele)sonorous body" from the theoretical and aesthetic point of view, and to explore the telepresence beyond the relation with the image. The research about sonorities in my artistic process in telematic context began in 2011 and was deepened during the investigation in my post doctoral at the Sonic Arts Research Centre (UK). In this article, it is assumed that a human being knows the world through her/his sensorimotor skills when s/he interacts with the environment [6] and understand the milieu. The digital culture brings some important transformations which are embodied including the notions of negotiation, construction, context and distributed mind which overlap the conventional ideas of reception, representation, content and autonomous brain [1, 2]. Our action in face of telepresence brings different ways of how to perceive one another and how to perceive oneself, because the body image and body schema play an active role in shaping our perceptions [5]. During my trajectory working with Telematic Dance since 2001, my goal has been to investigate new relationships between remote dancers in performances embedded into this digital culture.

REFERENCES

[1] R. Ascott. *Telematic Embrace: Visionary Theories of Art, Technology, and Consciousness.* Berkeley and Los Angeles: University of California Press (2003).
[2] _____ Is There Love in the Telematic Embrace? In: *Art Journal,* Vol 49, No.3, Computers and Art: Issues of Content pp. 241-247 (1990).
[3] A. Clark. *Being There: Putting Brain, Body, and World Together Again.* Cambridge, London: Bradford Book, MIT Press (1997).
[4] _____ *Natural born-cyborg*. Oxford: Oxford University Press (2003).
[5] S. Gallagher. *How the body shapes the mind.* Oxford, New York: Oxford University Press (2005).
[6] A. Noë. *Action in Perception.* Cambridge: MIT Press (2004).
[7] _____ *Varieties of Presence.*Cambridge, London: Harvard University Press (2012).

The Pleasure of Not Finding Things Out: Dramaturging with Boundary Objects

Freya Vass-Rhee

Abstract. The work of the dramaturg is usually thought of as a practice of helping a director or ensemble to reconcile, refine, and consolidate ideas into a coherent scenic whole. However, in the work of devising dance and theatre, by contrast, neither highly specified task distribution nor acute communicative coordination are necessarily required or even desired. Instead, as in the dramaturgical practice of choreographer William Forsythe and his ensemble, dramaturgy is a distributed phenomenon in which informational sharing is deprioritized in favor of an opening the work of devising to flexibility and change. In this talk, I evoke Star and Griesemer's concept of *boundary objects*—things or concepts which, although jointly deployed by members of a community, are utilized differently by different participants – to describe how The Forsythe Company's dramaturgy, rather than involving an informational "pooling" typically associated with ensemble dramaturgical practice, instead entails a radical and verbally reticent spreading of concepts that unsettles the practice of dramaturgy, while simultaneously calling the dramaturg's function into question. My analysis also reveals how Forsythe's ensemble's practice exemplifies a reversal of the trajectory towards informational coherence that typifies problem-solving, and in doing so, highlights and critiques key aspects of devising and improvisational work in theatre.

The Embodiment of Sound in an Intermedial Performance Space

Dr. Caroline Wilkins[1]

Abstract

Digital technology has merely reinforced the importance of the human body and the physical in live performance. [2]

In this paper I aim to describe the working process of a creative collaboration between electronics composer Oded Ben-Tal[3] and myself as performer, involving interactive audio technology. Methods, tools, terminologies and subjective experience all present some meta-technical issues that will be raised with regard to a project essentially embedded in the medium of sound theatre (a performance concept that draws attention to the phenomenological qualities of sound, music and theatre) and installation.

Coming from a background of theatre, performance and acousmatics, I shall examine the work from the perspective of these disciplines. Documenting the process of exchange at each stage allowed for an ongoing analysis of methods that were used to facilitate communication and developmental procedure within the larger context of a multimedia performance project. As an example of developing performance practice, this took the form of a choreographic installation encompassing dance, video, animation, visual design and virtual worlds, and was entitled *Ukiyo: Moveable Worlds* http://people.brunel.ac.uk/dap/ukiyo.html

I will focus on the use of language and systems as cognitive tools for research, as well as on some phenomenological aspects of performing together with technology, such as acting / reacting, action / sound, 'self / other'. Meta-technical ideas will be explored with regard to the spatial and temporal considerations involved in this kind of process: the acoustic, the three-dimensional, absence/presence of a sound source and its evolving relationship with the visual elements of performance. According to post-human philosophy it is these parameters of technology, belonging to a cognitive system, that have caused our human functionality to expand.[4]

In this case the key sound sources stemmed from a bandoneon (a musical instrument similar to the accordion) and the voice. They were combined with choreographic movement and a wearable costume that incorporated wired and wireless systems of amplification into its design. Sounds of an acoustic nature were thus transformed through the use of technology, into several *extended* instruments in space. This shared artistic space, where audience, performers and sound are considered in a parallel relationship, offers a very different premise for a work's reception and perception when compared to traditional performance practice. What is seen is not necessarily heard (and v. v.) and certainly not experienced in the same way by all.

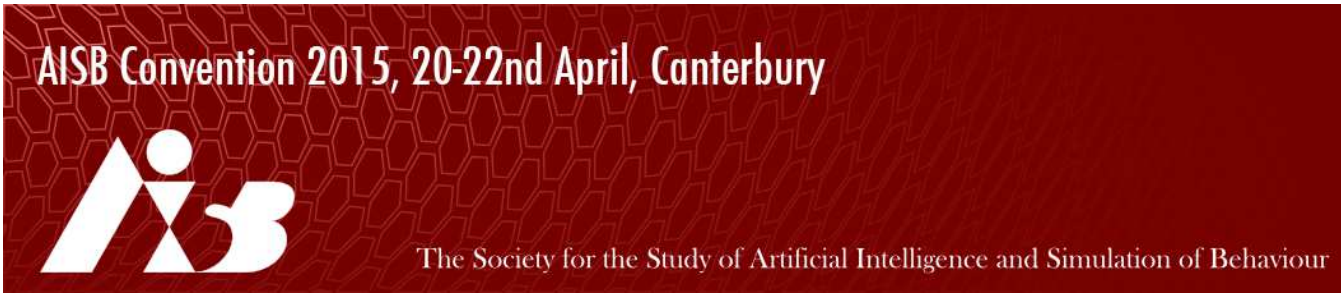
My presentation will include a performance of some of the live sonic material followed by recordings of its electronic transformation into a re-embodied form.

[1] Independent composer/performer/researcher. email: juillet1953@gmail.com

[2] Richards, J. 'Getting the Hands Dirty', *Leonardo Music Journal,* Vol. 18/1, 2008

[3] School of Performance and Screen Studies, Kingston University.

[4] Hutchins, E. (1995) *Cognition in the Wild,* Cambridge: MIT Press.



AISB Convention 2015

University of Kent, Canterbury

Proceedings of the AISB 2015 Symposium on New Frontiers in Human-Robot Interaction

Edited by Maha Salem, Astrid Weiss, Paul Baxter and Kerstin Dautenhahn

ISBN: 978-1-908187-51-2

Introduction to the Convention

The AISB Convention 2015—the latest in a series of events that have been happening since 1964—was held at the University of Kent, Canterbury, UK in April 2015. Over 120 delegates attended and enjoyed three days of interesting talks and discussions covering a wide range of topics across artificial intelligence and the simulation of behaviour. This proceedings volume contains the papers from the *Symposium on New Frontiers in Human-Robot Interaction,* one of eight symposia held as part of the conference. Many thanks to the convention organisers, the AISB committee, convention delegates, and the many Kent staff and students whose hard work went into making this event a success.

—Colin Johnson, Convention Chair

Introduction to the Symposium

The Symposium on “New Frontiers in Human-Robot Interaction (HRI)” is the fourth of a series of symposia held in conjunction with the AISB convention. Its topics cover cutting-edge interdisciplinary research on understanding, designing, and evaluating robotic systems for and with humans. Its main difference to other HRI-related conferences and workshops is its inclusiveness for exploratory research and the amount of time for open discussion. This year’s symposium consists of six sessions covering topics such as verbal and non-verbal interaction, people’s perception of robots, and ethical issues. Moreover, it includes keynote talks by Mark Coeckelbergh and Angelika Peer and a panel on the topic “Robot Perception and Acceptance”.

Introduction

Human-Robot Interaction (HRI) is a quickly growing and very inter- disciplinary research field. Its application areas will have an impact not only economically, but also on the way we live and the kinds of relationships we may develop with machines. Due to its interdisciplinary nature of the research different views and approaches towards HRI need to be nurtured.

In order to help the field to develop, the Symposium on New Frontiers in Human-Robot Interaction encourages submissions in a variety of categories, thus giving this event a unique character. The symposium consists of paper presentations, panels and, importantly, much time for open discussions which distinguishes this event from regular conferences and workshops in the field of HRI.

History

The first symposium on “New Frontiers in Human-Robot Interaction” was held as part of AISB 2009 in Edinburgh, Scotland; the second symposium was run in conjunction with AISB 2010 in Leicester, England; the third symposium took place during AISB 2014 at Goldsmiths, University of London, England. These three previously organised symposia were characterised by excellent presentations as well as extensive and constructive discussions of the research among the participants. Inspired by the great success of the preceding events and the rapidly evolving field of HRI, the continuation of the symposium series aims to provide a platform to present and discuss collaboratively recent findings and challenges in HRI.

Submission Categories

In order to enable a diverse program, the symposium offers a variety of submission categories, which go beyond typical conference formats. The fourth symposium offered the following categories in the call for papers:

**N* Novel research findings resulting from completed empirical studies.* In this category we encourage submissions where a substantial body of findings has been accumulated based on precise research questions or hypotheses. Such studies are expected to fit within a particular experimental framework (e.g. using qualitative or quantitative evaluation techniques) and the reviewing of such papers apply relevant (statistical and other) criteria accordingly. Findings of such studies should provide novel insights into human-robot interaction studies.

**E* Exploratory studies.* Exploratory studies are often necessary to pilot and fine-tune the methodological approach, procedures and measures. In a young research field such as HRI with novel applications and various robotic platforms, exploratory studies are also often required to derive a set of concrete research questions or hypotheses, in partic- ular concerning issues where there is little related theoretical and experimental work. Although care must be taken in the interpretation of findings from such studies, they highlight issues of great interest and relevance to peers.

**S* Case studies.* Due to the nature of many HRI studies, a large-scale quantitative approach is sometimes neither feasible nor desirable. However, case study evaluation provides meaningful findings if presented appropriately. Thus, case studies with only one participant, or a small group of participants, are encouraged if they are carried out and analysed in sufficient depth.

**P* Position papers.* While categories N, E and S required reporting on HRI studies or experiments, position papers can be conceptual or theoretical, providing new interpretations of known results. Also, in this category we consider papers that present new ideas without having a complete study to report on. Papers in this category are judged on the soundness of the argument presented, the significance of the ideas and the interest to the HRI community.

**R* Replication of HRI studies.* To develop as a field, HRI findings obtained by one research group need to be replicated by other groups. Without any additional novel insights, such work is often not publishable. Within this category, authors have the opportunity to report on studies that confirm or disconfirm findings from experiments that have already been reported in the literature. This category includes studies that report on negative findings.

**D* Live HRI Demonstrations.* Contributors have the opportunity to provide live demonstrations (live or via Skype), pending the outcome of negotiations with the local organisation team. The demo should highlight interesting features and insights into HRI. Purely entertaining demonstrations without significant research content are discouraged.

**Y* System Development.* Research in this category included the design and development of new sensors, robot designs and algorithms for socially interactive robots. Extensive user studies are not necessarily required in this category.

Natural Interaction with Social Robotics

The Fourth Symposium on "New Frontiers in Human-Robot Interaction" was organised in conjunction with the Topic Group on Natural Interaction with Social Robotics. This Topic Group was launched within the EU Horizon 2020 funding framework (http://ec.europa.eu/programmes/horizon2020/), with the strategic goal to keep the topic of interaction prominent in the future calls for European projects. An overview on the list of topics and interests of the Topic Group can be found on the website: http://homepages.stca.herts.ac.uk/~comqkd/TG-NaturalInteractionWithSocialRobots.html.

As the symposium offers an ideal opportunity to discuss related research topics that are relevant for the Topic Group, we introduced one new submission category:

**TG* Topic Group Submissions on "Natural Interaction with Social Robots".* Submissions in this category will be discussed in a session dedicated to the euRobotics Topic Group "Natural Interaction with Social Robots". Topics specifically relevant to the TG are e.g. benchmarking of levels of social abilities, multimodal interaction, and human-robot interaction and communication.

Programme Overview

This year's symposium consists of 17 talks, based on submissions in the following categories:

- *N* Novel research findings resulting from completed empirical studies: 5 submissions
- *E* Exploratory studies: 5 submissions
- *P* Position papers : 4 submissions
- *Y* System Development: 2 submissions
- *TG* Topic Group Submissions on "Natural Interaction with Social Robots": 1 submission

The talks are structured in six sessions:

1. Ethical issues in HRI
2. Robots' impact on human performance
3. Verbal interaction
4. Facial expressions & emotions
5. Non-verbal cues & behaviours
6. Robot perception & acceptance

The final session is followed by a panel discussion on the same topic. Two invited keynote talks complete the program:

1. Mark Coeckelbergh: "Human-like Robots and Automated Humans: Socializing and Contextualizing HRI"
2. Angelika Peer: "Towards Remote Medical Diagnosticians"

Conclusion

In summary, the symposium mainly focuses on novel empirical findings on human-robot interaction and their impact on our everyday life. Moreover, also theoretical aspects and ethical issues are dis- cussed. We hope these articles show some future research directions for fellow HRI researchers and stimulate ideas for future European projects on natural interaction with social robots.

Contents

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

General Republics' Opinions on Robot Ethics: Comparison between Japan, the USA, Germany, and France

Tatsuya Nomura [1]

Abstract. Ethical issues on robots need to be investigated based on international comparison because general publics' conceptualizations of and feelings toward robots differ due to different situations with respect to mass media and historical influences of technologies. As a preliminary stage of this international comparison, a questionnaire survey based on openended questions was conducted in Japan, the USA, Germany and France ($N = 100$ from each countries). As a result, it was found that (1) people in Japan tended to react to ethical issues of robotics more seriously than those in the other countries, although those in Germany tended not to connect robotics to ethics, (2) people in France tended to specify unemployment as an ethical issue of robotics in comparison with the other countries, (3) people in Japan tended to argue the restriction of using and developing robots as a solution for the ethical problems, although those in France had the opposite trend.

1 Introduction

The recent development of robotics has begun to introduce robots into our daily lives in our homes, schools, and hospitals. In this situation, some philosophers and scientists have been discussing robot ethics [8, 15, 12, 4, 2]. Asaro [1] argued that robot ethics should discuss the following three things: the ethical systems to be built into robots, the ethics of people who design and use robots, and ethical relationships between humans and robots. Lin [6] proposed the following three broad (and interrelated) areas of ethical and social concerns about robotics:

Safety and errors: including mistakes of recognition by battle robots and security against hacking.
Law and ethics: including codes of ethics to be programed into robots, companionships between humans and robots, responsibility of robot behaviors.
Social impact: including economical and psychological change of the society.

Recently, several researchers have been investigating solutions for these ethical problems. However, the opinions of the general public of different countries have not sufficiently been investigated from the perspective of robot ethics. Some existing studies found the general public's preferences of robot types in the context of domestic use [14], expectation of task types in domestic household robots [11], attitudes regarding robots' suitability for a variety of jobs [17], safety perception of humanoid robots [5], and fear and anxiety [9]. However, these survey studies did not focus on the ethical issues of robots.

Moreover, the ethical issues of robots need to be investigated based on international comparison because general publics' conceptualizations of and feelings toward robots differ due to different situations with respect to mass media and historical influences of technologies. In fact, recent studies [16, 19, 13, 18] show differences of opinions of robots between countries, including attitudes toward robots [3, 20], images of robots [10], and implicit attitudes [7]. In addition, interpretations of the word "ethics" differ between countries because of different social norms. Thus, we should compare the opinions of the general publics of several countries when they face the words "robots" and "ethics" at the same time. This comparison will contribute to preparation of discussion on the international consensus of robotics applications.

As a preliminary stage of the international comparison on robot ethics issues, a questionnaire survey based on open–ended questions was conducted in Japan, the USA, and Europe. To take into account the historical influences of wars into the ethical perspectives of military robotics, the survey in Europe was conducted in Germany and France, which were a defeated country and a victorious country in World War II, respectively. This paper reports the results of the survey and then discusses the implications.

2 Method

2.1 Participants and Data Collection Procedure

The survey was conducted from January to February, 2013. Respondents were recruited by a survey company (Rakuten Research). When the survey was conducted, the numbers of possible respondents registered to the company was about 2,300,000 in Japan, 2,780,000 in the USA, 310,000 in Germany, and 450,000 in France. Among the people randomly selected from these large pools of samples based on gender and age, a total of 100 people of ages ranging from 20's to 60's participated in the survey in each of the four countries. Table 1 shows the sample numbers based on country, gender, and age categories.

[1] Department of Media Informatics, Ryukoku University, Japan, email: nomura@rins.ryukoku.ac.jp

The questionnaire consisting of open–ended items was conducted via Internet homepages in all the countries.

Table 1. Sample Numbers Based on Countries, Gender, and Age Categories

		20's	30's	40's	50–60's	Total
Japan	Male	13	12	13	12	50
	Female	12	13	12	13	50
	Total	25	25	25	25	100
USA	Male	11	13	12	14	50
	Female	11	14	18	7	50
	Total	22	27	30	21	100
Germany	Male	12	11	16	11	50
	Female	10	12	15	13	50
	Total	22	23	31	24	100
France	Male	10	15	12	13	50
	Female	20	8	10	12	50
	Total	30	23	22	25	100
Total		99	98	108	95	400

2.2 Measures

As mentioned in the introduction section, the survey aimed at investigating interpretations of the general publics when they face the words robots and ethics at the same time. To measure and compare their primitive conceptualization between the countries, we did not instruct the definitions of "robots" or "ethics".

The questionnaire solicited information about (1) age, (2) gender, (3) occupation (subject of study if respondents were students), and (4) three questions about ethics and robotics. The questionnaire items about ethics and robotics were open–ended, and designed to elicit a wide variety of responses:

Q1: What would you image when hearing "robots" and "ethics" at the same time?
Q2: What sort of ethical problems would happen when robots widespread in society?
Q3: How should we solve the problems mentioned in item 2?

The questionnaire was conducted in Japanese, English, German, and French languages in Japan, the USA, Germany, and France, respectively. The response sentences in Germany and France were translated into English.

3 Results

3.1 Coding of Open–Ended Responses

For quantitative analyses, the open–ended responses were manually classified into categories based on the contents of the responses. This classification coding was determined by two coders. The first coder dealt with both Japanese and English sentences. The second coder consisted of two people, one for the Japanese sentences and another for the English sentences.

First, coding rules were created for each item. Then, two coders independently conducted the coding of 40% of the responses ($N = 40$ from all the responses of each country), and calculated the κ–coefficients showing the degrees of agreement between the two coded results in order to validate the reliability of the coding rules. The coefficients showed sufficient reliability of the coding rules. Table 2 shows coding rule numbers, examples of sentences in the coding, and κ–coefficients. Furthermore, the two coders interactively discussed the contents of the responses and coding results until they reached a consensus about each coding.

3.2 Q1: Images When Hearing "Robots" and "Ethics" at the Same Time

In Q1, each participant's response was classified into one of the three categories shown in Table 2. Responses assigned L0 showed no concrete image. In the German and French samples, several wrote sentences meaning that the words "robots" and "ethics" clashed with each other. Responses assigned L1 stated images from science fiction contents. Responses assigned L2 included realistic concerns of robotics in society and ambiguous apprehension toward the development of robots.

Table 3 shows the distributions of answer categories based on the countries and the results of a χ^2-test and a residual analysis with $\alpha = .05$. Approximately 60% of the respondents mentioned some apprehension toward robotics. The χ^2–test showed differences between the countries in the category distribution. The residual analysis revealed that in the Japan sample, the frequency of L0 was lower than average and that of L1 was higher than average at statistically significant levels. Moreover, in the German samples, the frequency of L0 was higher than average and that of L2 was lower than average. Furthermore, in the French samples, the frequency of L1 was lower than average and that of L2 was higher than average at statistically significant levels.

To visualize the relationships between countries and images of robots and ethics, a correspondence analysis was performed for the cross-table shown in Table 3. The correspondence analysis allows us to visualize the relationship between categories appearing in a cross–table in two-dimensional space. In this visualization, categories similar to each other are placed at proximate positions. Our analysis using this method aims to clarify the relationship between the countries and respondents' images when hearing "robots" and "ethics" at the same time. We should note that the dimensional axes extracted from the data in the cross–table are specific to the table data and are used to visualize relative distances between categories; that is, they do not correspond to any absolute measure, and so it is difficult to assign realistic meanings to these axes.

Figure 1 shows the results of the analysis. The USA is positioned at the middle point between the three answer categories, and Germany is located at L0. Japan is positioned at the middle point between L1 and L2, and France is near L2. These results can be summarized as follows:

- Compared with the other countries, less German respondents specified images in which robots and ethics appeared at the same time.
- More French respondents specified apprehension toward robotics than did the respondents in the other countries.
- More Japanese respondents specified images from virtual contents in comparison with the respondents in the other countries.

Table 2. Coding Rules of Open–Ended Responses and Reliability

Item	Rule	Label		κ
Q1:	R1:	L0:	Responses that did not image any concrete problems (e.g., "nothing", "don't think ...")	.747
		L1:	Responses that mentioned virtual contents including movies, animations, and comics (e.g., "Robocop", "Blade Runner")	
		L2:	Ones except for the above L0 and L1 (e.g., "What are the ethical rules to apply when using robots?"	
Q2:	R21:	L1:	Responses that mentioned unemployment problems (e.g., "Job losses", "Replacing people with robots so unemployment")	.922
		L0:	Others	
	R22:	L1:	Responses that mentioned crimes or wars (e.g., "People use them to spy", "With battle robots, that will make killing easier and easier")	.717
		L0:	Others	
	R23:	L1:	Responses that mentioned some problems except unemployment, crimes and wars (e.g., "Accidents by robots", "There will be no difference between humans/robots")	.711
		L0:	Others	
Q3:	R3:	L0:	Responses that did not mention any concrete problems in Q2	.647
		L1:	Responses that mentioned restriction of robots' functions, methods of using robots, and areas of robot applications, and legal preparation for the restriction (e.g., "Only use robots in certain situations", "Don't give robots the ability of "think"")	
		L2:	Ones except for the above L0 and L1 (e.g., "I have no idea", "Improvement of human morals", "Keep our manual skills")	

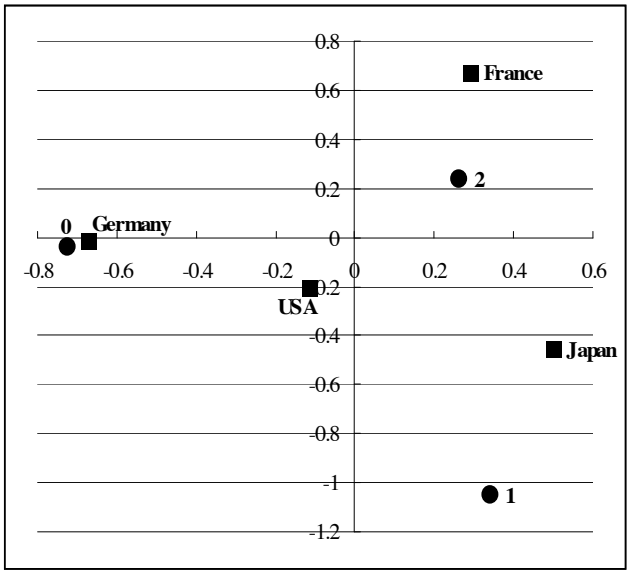


Figure 1. Result of Correspondence Analysis for Table 3

Table 3. Distribution of Answer Categories for Q1 and Results of χ^2–Test and Residual Analysis ($\alpha = .05$)

	Answer Category of R1			
	L0	L1	L2	Total
Japan	$18^{\downarrow}$	$21^{\uparrow}$	61	100
USA	30	15	55	100
Germany	$41^{\uparrow}$	10	$49^{\downarrow}$	100
France	21	$5^{\downarrow}$	$74^{\uparrow}$	100
Total	110 (27.5%)	51 (12.75%)	239 (59.75%)	400 (100%)

$\chi^2(6) = 28.448$, $p < .001$
$^{\uparrow}$: higher than the expected frequency
$^{\downarrow}$: lower than the expected frequency
L0: Responses that did not image any concrete problems
L1: Responses that mentioned virtual contents including movies, animations, and comics
L2: Ones except for the above L0 and L1

3.3 Q2: Ethical Problems in Society

In Q2, one response included several different problems. Thus, each participant's response was assigned multiple labels based on the following rules: (R21) whether it mentioned unemployment problems due to robots, (R22) whether it mentioned the use of robots in crimes and wars, and (R23) whether it mentioned some problems besides unemployment, crimes, and wars. Responses assigned as L1 in R23 included apprehension toward the physical and economical risks of robots, their influences on humans' psychological states, and ambiguous differences between robots and humans.

Table 4 shows the distributions of answer categories based on the countries and the results of the χ^2–test and the residual analysis with $\alpha = .05$. The results can be summarized as follows:

- In the Japan sample, fewer respondents mentioned unem-

Table 4. Distribution of Answer Categories for Q2 and Results of χ^2–Test adn Residual Analysis ($\alpha = .05$)

	R21: Unemployment		R22: Crimes and Wars		R23: Other Problems	
	Not mentioned	Mentioned	Not mentioned	Mentioned	Not mentioned	Mentioned
Japan	$87^{\uparrow}$	$13^{\downarrow}$	$85^{\downarrow}$	$15^{\uparrow}$	$34^{\downarrow}$	$66^{\uparrow}$
USA	77	23	$84^{\downarrow}$	$16^{\uparrow}$	$65^{\uparrow}$	$35^{\downarrow}$
Germany	82	18	$97^{\uparrow}$	$3^{\downarrow}$	47	53
France	$64^{\downarrow}$	$36^{\uparrow}$	$97^{\uparrow}$	$3^{\downarrow}$	$60^{\uparrow}$	$40^{\downarrow}$
Total	310	90	363	37	206	194
	(77.5%)	(22.5%)	(90.75%)	(9.25%)	(51.5%)	(48.5%)
	$\chi^2(3) = 16.803$,	$p < .01$	$\chi^2(3) = 18.673$,	$p < .001$	$\chi^2(3) = 23.261$,	$p < .001$

$^{\uparrow}$: higher than the expected frequency, $^{\downarrow}$: lower than the expected frequency

ployment problems at a statistically significant level in comparison with the other countries.

  - More respondents in the French sample mentioned unemployment.

- The respondents mentioning crimes and wars as ethical problems of robotics in society were in the minority (less than 10%).
  - Nevertheless, more respondents mentioned these problems in the Japan and USA samples than in the German and French samples at statistically significant levels.
- More respondents mentioned problems besides unemployment, crimes, and wars in the Japan samples than in the samples of the other countries.
  - On the other hand, fewer respondents in the USA and French samples mentioned these problems than in the Japan and German samples.

3.4 Q3: Solutions for Ethical Problems of Robotics

In Q3, each participant's response was classified into one of the three categories shown in Table 2. Responses assigned label L0 corresponded to the ones that did not specify anything on the ethical problems of robotics in society in Q2 (that is, participants assigned L0 for R21, R22, and R23). In Q3, responses assigned label L1 mentioned restriction of robots functions, methods of using robots, and areas of robot applications. Some responses classified into this category mentioned the need of legal preparation for the restriction. Responses assigned label L2 included the ones that did not provide any concrete solution or the ones that did show some solutions except restriction of robots.

Table 5 shows the distributions of the answer categories based on the countries and the results of the χ^2–test and the residual analysis with $\alpha = .05$. The χ^2–test showed differences between the countries in the category distribution. The residual analysis revealed that in the Japan sample, the frequency of L0 was lower than average and that of L1 was higher than average at statistically significant levels. About half of them mentioned restriction of robotics usage as a solution to their ethical problems. Moreover, it was found that in the German samples, the frequency of L0 was higher than average. Furthermore, in the French samples, the frequency of L1 was lower than average and that of L2 was higher than average at statistically significant levels.

In the same way as Q1, the correspondence analysis for Q3 in Table 5 was conducted to visualize relationships between countries and solution categories for the ethical problems of robots. Figure 2 shows the result. Japan was positioned far from L0 and L2, near L1. France was positioned far from L0 and L1, near L2. The USA and Germany were positioned at the middle of L0 and L1, far from L2. These results can be summarized by the following comparisons between the countries:

- More respondents in Japan specified ethical problems of robots in society and mentioned restriction of robots in terms of functions and methods of usage as a solution to the problems.
- Fewer French respondents mentioned restriction of robots as the problem solution.
- In the USA and particularly in Germany, many respondents did not specify any problem or solution for the ethical issues of robots in society.

Table 5. Distribution of Answer Categories for Q3 and Results of χ^2–Test adn Residual Analysis ($\alpha = .05$)

	Answer Category of R3			
	L0	L1	L2	Total
Japan	$6^{\downarrow}$	$52^{\uparrow}$	42	100
USA	26	43	31	100
Germany	$27^{\uparrow}$	43	30	100
France	21	$30^{\downarrow}$	$49^{\uparrow}$	100
Total	80	168	152	400
	(20%)	(42%)	(38%)	(100%)

$\chi^2(6) = 26.536$, $p < .001$
$^{\uparrow}$: higher than the expected frequency
$^{\downarrow}$: lower than the expected frequency
L0: Responses that did not mention any concrete problems in Q2
L1: Responses that mentioned restriction of robots' functions, methods of using robots, and areas of robot applications, and legal preparation for the restriction
L2: Ones except for the above L0 and L1

4 Discussion

4.1 Findings

The survey results suggest some characteristics of Japan, the USA, Germany, and France when the general public of each country faces the issues regarding robot ethics.

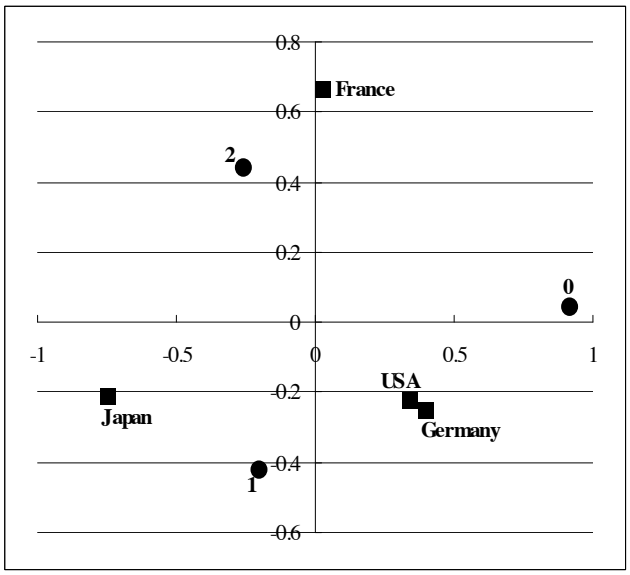


Figure 2. Result of Correspondence Analysis for Table 5

People in Japan tended to react to ethical issues of robotics more seriously than those in the USA, Germany, and France, while they were more influenced by virtual contents such as science fiction movies. In contrast, people in Germany were least likely to connect robotics to ethics. People in France, despite also being in the EU, had a different trend from those in Germany in the sense that they expressed more apprehension toward robotics.

Unemployment as an ethical issue of robotics showed different reactions between these four countries. In particular, Japan and France had opposite trends with respect to this problem. Relationships of robotics with crimes and wars also showed different reactions between the countries. Although a minority of people mentioned this issue as overall, more people tended to specify the issue in Japan and in the USA than in the two European countries.

Consideration of the solutions for the ethical problems of robotics showed opposite trends in Japan and France. Unlike the people in France, the people in Japan tended to argue for restricting the use and development of robots as a solution to ethical problems.

4.2 Implications

The above findings in the survey imply some problems when discussing issues regarding robot ethics at the international level.

First, differences are possible between countries on their general publics awareness of issues regarding robot ethics. Some people may not assume the existence of ethical problems related to robotics. It is implied that the rate of participants in the discussion about robot ethics in society may change depending on the country. Second, it is possible that individual problems have impact on the general public in different ways in different countries. People in one country may participate in discussing an ethical issue and those in another country may not. Such differences in attitudinal biases toward the discussion of robot ethics between countries would make it hard to share problems and solutions internationally. If an ethical problem regarding robots is serious in a country and potentially poses a risk in another country, leaders of the discussion should take into account the differences of awareness of the problem between the countries to establish common assumptions and ways of discussion.

4.3 Limitations

The survey adopted three simple questions and open–ended responses. Thus, the differences of opinions between countries are superficial, and deep factors causing the differences were not explored. It is estimated that these factors include religious beliefs and historical backgrounds in countries, particularly with regard to unemployment and wars. Moreover, the concept of robots may differ between countries [10].

The total number of samples in the survey was not enough to generalize the findings. To clarify more strictly differences in the general publics opinions regarding robot ethics between countries and investigate causes of the differences, we should conduct future surveys using detailed questionnaire items having sufficient validity with a wider area of samples.

ACKNOWLEDGEMENTS

The research was supported in part by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research No. 21118006 and 25280095.

The author deeply thank Ms. Kanako Tomita and Anderson Li in ATR Intelligent Robotics and Communication Laboratories for their cooperation with the analyses of data in the paper.

REFERENCES

[1] P. M. Asaro, 'What should we want from a robot ethic?', *International Review of Information Ethics*, **6**, 9–16, (2006).

[2] W. Barendregt, A. Paiva, A. Kappas, A. Vasalou, C. Heath, S. Serholt, C. Basedow, and A. O. Patrcia, 'Child–robot interaction: Social bonding, learning and ethics', in *Workshop proceedings of Interaction Design and Children Conference IDC14*, (2014).

[3] C. Bartneck, T. Suzuki, T. Kanda, and T. Nomura, 'The influence of people's culture and prior experiences with Aibo on their attitude towards robots', *AI & Society*, **21**(1–2), 217–230, (2007).

[4] G. Briggs and M. Scheutz, 'How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress', *International Journal of Social Robotics*, **6**(3), 343–355, (2014).

[5] H. Kamide, Y. Mae, K. Kawabe, S. Shigemi, and T. Arai, 'Effect of human attributes and type of robots on psychological evaluation of humanoids', in *Proc. IEEE Workshop on Advanced Robotics and its Social Impacts*, pp. 40–45, (2012).

[6] P. Lin, 'Introduction to robot ethics', in *Robot Ethics: The ethical and social implications of robotics*, eds., P. Lin, K. Abney, and G. A. Bekey, 3–15, MIT Press, (2011).

[7] K. F. MacDorman, S. K. Vasudevan, and C-C. Ho, 'Does Japan really have robot mania? comparing attitudes by implicit and explicit measures', *AI & Society*, **23**(4), 485–510, (2009).

[8] A. Moon, P. Danielson, and H.F.M. Van der Loos, 'Survey–based discussions on morally contentious applications of interactive robotics', *International Journal of Social Robotics*, **4**(1), 77–96, (2012).

[9] T. Nomura, K. Sugimoto, D. S. Syrdal, and K. Dautenhahn, 'Social acceptance of humanoid robots in japan: A survey for development of the Frankenstein Syndrome Questionnaire', in *Proc. 12th IEEE–RAS International Conference on Humanoid Robots*, pp. 242–247, (2012).

[10] T. Nomura, T. Suzuki, T. Kanda, J. Han, N. Shin, J. Burke, and K. Kato, 'What people assume about humanoid and animal-type robots: Cross-cultural analysis between Japan, Korea, and the USA', *International Journal of Humanoid Robotics*, **5**(1), 25–46, (2008).

[11] L. Oestreicher and K. S. Eklundh, 'User expectations on human–robot co–operation', in *Proc. IEEE International Symposium on Robot and Human Interactive Communication*, pp. 91–96, (2006).

[12] L. D. Riek and D. Howard. A code of ethics for the human–robot interaction profession. Presented at WeRobot 2014 Conference, March 2014.

[13] L. D. Riek, N. Mavridis, S. Antali, N. Darmaki, Z. Ahmed, M. A. Neyadi, and A. Alketheri, 'Ibn sina steps out: Exploring arabic attitudes toward humanoid robots', in *Proc. 36th Annual Convention of the Society for the Study for Artificial Intelligence and the Simulation of Behaviour (AISB 2010)*, volume 1, pp. 88–94, (2010).

[14] M. Scopelliti, M. V. Giuliani, and F. Fornara, 'Robots in a domestic setting: A psychological approach', *Universal Access in the Information Society*, **4**(2), 146–155, (2005).

[15] A. Sharkey and N. Sharkey, 'Granny and the robots: ethical issues in robot care for the elderly', *Ethics and Information Technology*, **14**(1), 27–40, (2012).

[16] T. Shibata, K. Wada, Y. Ikeda, and S. Sabanovic, 'Cross–cultural studies on subjective evaluation of a seal robot', *Advanced Robotics*, **23**(4), 443–458, (2009).

[17] L. Takayama, W. Ju, and C Nass, 'Beyond dirty, dangerous and dull: What everyday people think robots should do', in *Proc. 3rd ACM/IEEE International Conference on Human-Robot Interaction*, pp. 25–32, (2008).

[18] G. Trovato, M. Zecca, S. Sessa, L. Jamone, J. Ham, K. Hashimoto, and A. Takanishi, 'Cross–cultural study on human-robot greeting interaction: acceptance and discomfort by egyptians and japanese', *Paladyn, Journal of Behavioral Robotics*, **4**(2), 83–93, (2013).

[19] S. Šabanović, 'Robots in society, society in robots', *International Journal of Social Robotics*, **2**(4), 439–45–, (2010).

[20] L. Wang, P-L. P. Rau, V. Evers, B. Krisper, and P. Hinds, 'When in Rome: The role of culture & context in adherence to robot recommendations', in *Proc. 5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 359–366, (2010).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Differences on Social Acceptance of Humanoid Robots between Japan and the UK

Tatsuya Nomura[1], **Dag Sverre Syrdal**[2], **and Kerstin Dautenhahn**[2]

Abstract. To validate a questionnaire for measuring people's acceptance of humanoid robots in cross-cultural research (the Frankenstein Syndrome Questionnaire: FSQ), an online survey was conducted in both the UK and Japan including items on perceptions of the relation to the family and commitment to religions, and negative attitudes toward robots (the NARS). The results suggested that 1) the correlations between the FSQ subscale scores and NARS were sufficient, 2) the UK people felt more negative toward humanoid robots than did the Japanese people, 3) young UK people had more expectation for humanoid robots, 4) relationships between social acceptance of humanoid robots and negative attitudes toward robots in general were different between the nations and generations, and 5) there were no correlations between the FSQ subscale scores, and perception of the relation to the family and commitment to religions.

1 INTRODUCTION

In recent years, several studies have revealed the influences of human cultures into feelings and behaviors toward robots [1, 2, 3, 4, 5, 6], and some of them focused on social acceptance of robots. Evers, et al. [1] revealed differences between the US and Chinese people on their attitudes toward and the extent to which they accepted choices made by a robot. Li, et al. [2] found an interaction effect between human cultures (Chinese, Korean and German) and robots' tasks (teaching, guide, entertainment and security guard) on their engagement with the robots. Yueh and Lin [5] showed differences on preferences of home service robots between Taiwanese and Japanese people.

The research group also have been developing a questionnaire to measure and compare humans' acceptance of humanoid robots between nations, and explore factors influencing social acceptance of humanoids including cultural ones [7, 8]. The questionnaire, called "Frankenstein Syndrome Questionnaire" (FSQ), aims at clarification of differences on social acceptance of humanoid robots between the Westerners and Japanese based on Kaplan's idea [9] reflecting the concept of "Frankenstein Syndrome" originated from genetic engineering [10]. The surveys using this questionnaire suggested age differences on acceptance of humanoid robots in Japan [11], and some differences between the UK and Japan [8].

However, the previous studies had some problems on sampling in the sense that data from an online survey and that based on a normal paper-and-pencil method were mixed in one nation sample. As a result, the factor structure extracted from the sample was not stable [12]. Moreover, the previous survey did not take into account verification of criterion-related validity of the questionnaire.

To overcome the above problems, an online survey was conducted in both the UK and Japan under more strict control of sampling. The survey included another psychological scale of which validity had already been supported, the Negative Attitudes toward Robots Scale [13]. The scale was used to verify correlations between social acceptance of humanoid robots and attitudes toward robots in general, to investigate the criterion-related validity of the Frankenstein Syndrome Questionnaire.

As well as cultures, the survey aimed at exploring other factors related to social acceptance of humanoid robots. As factors to be explored, the survey firstly focused on age. In the survey conducted in Japan about ten years ago, our research group found that persons in their 40s had positive opinions of robots in comparison with other generations [14]. Thus, the survey aimed at comparing one group of persons in their 50s with another in their 20s to clarify age differences. Moreover, a survey conducted in Japan and Sweden adopted perceptions of the relation to the family and commitment to religions as indices reflecting differences between these different nations [15]. Thus, the survey also included these two factors "the relation to the family" and "commitment to religions".

The paper reports the results of the survey, and discusses the implications from the perspective of development of humanoid robots.

2 Method

2.1 Date and Participants:

The survey was conducted from January to February 2014. 100 Japanese and 100 UK respondents were recruited by a survey company at which about one million and six hundred thousand Japanese and one million and one hundred thousand UK persons have registered. Respondents in each nation were limited to people who were born and had been living only in the corresponding nation. The respondents consisted of fifty persons in their 20s (male: 25, female: 25) and fifty persons in their 50s (male: 25, female: 25) in each of the nations.

The homepage of the online survey had been open for these participants during the above period. The questionnaire of the online survey was conducted with the native language for the respondents in each of the nations.

2.2 Survey Design:

The questionnaire did not give the explicit definition of robots, or include any photo and image of robots, except for the instruction on humanoid robots just before conducting the Frankenstein Syndrome Questionnaire. The scale on attitudes

[1] Dept. of Media Informatics, Ryukoku Univ., Shiga 520-2194, Japan. Email: nomura@rins.ryukoku.ac.jp.
[2] Adaptive Systems Research Group, Univ. of Hertfordshire, AL10 9AB, UK. Email: {D.S.Syrdal, K.Dautenhahn}@herts.ac.uk.

toward robots in general was firstly conducted, and then the Frankenstein Syndrome Questionnaire was conducted since the reverse order had a possibility that envisions of humanoids evoked by the conduction of the FSQ affected the measurement of attitudes toward robots in general. The concrete items and scales in the survey were as follows:

Perception of the Relation to the Family and Commitment to Religions:

The following two items, which were used in the comparison survey between Japan and the Northern Europe by Otsuka et al. [15], were presented on the face sheet measure participants' degrees of perception of the relation to the family and commitment to religions:

- Do you think you relate to your family members? (five-graded answer from "1. I completely agree" to "5. I completely disagree")
- Does such notion as "I have nothing to do with religion or faith" apply to you? (five-graded answer from "1. It strongly applies to me" to "5. It does not apply to me at all.")

Negative Attitudes toward Robots Scale (NARS):

To measure participants' attitudes toward robots in general, the NARS [13] was adopted in the survey. The scale consists of 14 items classified into three subscales. The first subscale (S1, six items) measures negative attitude toward interaction with robots (e.g., "I would feel paranoid talking with a robot."). The second subscale (S2, 5 items) measures negative attitude toward the social influence of robots (e.g., "Something bad might happen if robots developed into living beings."). The third subscale (S3, 3 items) measures negative attitude toward emotional interaction with robots (e.g., "I feel comforted being with robots that have emotions.").

Each item is scored on a five-point scale: 1) strongly disagree; 2) disagree; 3) undecided; 4) agree; 5) strongly agree, and an individual's score on each subscale is calculated by adding the scores of all items included in the subscale, with some items reverse coded.

Frankenstein Syndrome Questionnaire (FSQ):

The questionnaire was developed to measure acceptance of humanoid robots including expectations and anxieties toward this technology in the general public [8,11]. It consists of 30 items shown in Table 1. Each questionnaire item was assigned with a seven-choice answer (1: "Strongly disagree", 2: "Disagree", 3: "Disagree a little", 4: "Not decidable", 5: "Agree a little", 6: "Agree", 7: "Strongly agree".).

Just before conducting the FSQ, the definition of "humanoids robots" was instructed only with texts as follows:

"Humanoid robots are robots that roughly look like humans, that have two arms, legs, a head, etc. These robots may be very human-like in appearance (including details such as hair, artificial skin etc.), but can also have machine-like features (such as wheels, a metal skin etc)."

3 RESULTS

3.1 Subscales of the FSQ and Reliability:

Although previous studies had explored the factor structures in the FSQ [8,13], they were sufficiently not stable to be replicated across studies [12]. To extract the subscales of the FSQ again, a factor analysis with maximum likelihood method and Promax rotation was conducted for the 30 items. Although the analysis found five factors having eigen values more than 1, the scree plot showed that the difference on the eigen values between the fourth and fifth factors was small. Thus, the factor analysis was conducted based on four-factor structure. The cumulative contribution of these four factors was 52.8%.

After removing items having factor loadings more than .3 on more than one item, item analysis using Cronbach's α-coefficients and I-T correlations was performed for each factor in turn to select items in the corresponding subscale. Table 1 shows the results of these analyses.

The subscale corresponding to the first factor consisted of 9 items representing negative feelings toward social impacts of humanoid robots such as "Humanoid robots may make us even lazier." Thus, the subscale was interpreted as "negative feelings toward humanoid robots." The subscale corresponding to the second factor consisted of 8 items representing positive expectation of humanoid robots in the society such as "Humanoid robots can be very useful for teaching young kids." Thus, the subscale was interpreted as "expectation for humanoid robots". The subscale corresponding to the third factor consisted of 3 items representing negative feelings toward humanoid robots at religious and philosophical levels such as "The development of humanoid robots is blasphemous." Thus, the subscale was interpreted as "root anxiety toward humanoid robots". The fourth factor was removed in the analysis since it consisted of only two items.

Cronbach's reliability coefficients α, showing the internal consistencies of the subscales, were .899 for "negative feelings toward humanoid robots," .861 for "expectation for humanoid robots," and .859 for "root anxiety toward humanoid robots." These values showed sufficient internal consistencies for all three subscales. The score of each subscale was calculated as the sum of the scores of all items included in the subscale ("negative feelings toward humanoid robots": max 63, min 9, "expectation for humanoid robots": max 56, min 8, and "root anxiety toward humanoid robots": max 21, min 3).

3.2 Comparison between Nations and Generations:

FSQ Subscale Scores:

Three-way ANOVAs with gender by nation (Japan vs. UK) by generation (20's vs. 50's) were conducted for the subscale scores of the FSQ. Table 2 shows the results. For "negative feelings toward humanoid robots," the main effects of gender and nations were at statistically significant levels although the effect size on gender was small. For "expectation for humanoid robots," only the first order interaction effect between nations and generations was at a statistically significant level.

Figure 1 shows the means and standard deviations of the subscale scores of "negative feelings toward humanoid robots" and "expectation for humanoid robots". Bonfferoni Post Hoc tests revealed that the UK respondents in their 20s had higher expectation for humanoid robots than the UK respondents in

Item No.	Item Sentences	Factor I	Factor II	Factor III	Factor IV
30	Widespread use of humanoid robots would take away jobs from people.	**.929**	.076	-.098	-.212
4	Humanoid robots may make us even lazier.	**.766**	.037	-.057	-.077
12	If humanoid robots cause accidents or trouble, persons and organizations related to development of them should give sufficient compensation to the victims.	**.705**	.113	-.285	.132
8	I am afraid that humanoid robots will encourage less interaction between humans.	**.697**	.026	.167	-.015
20	I feel that if we become over-dependent on humanoid robots, something bad might happen.	**.681**	-.071	-.011	.245
17	I would hate the idea of robots or artificial intelligences making judgments about things.	**.655**	-.132	.279	-.045
11	I would feel uneasy if humanoid robots really had emotions or independent thoughts.	**.548**	-.055	-.004	.178
27	Something bad might happen if humanoid robots developed into human beings.	**.512**	-.048	.191	.193
23	*Humanoid robots should perform dangerous tasks, for example in disaster areas, deep sea, and space.*	***.493***	***.346***	*-.242*	*.055*
16	I am concerned that humanoid robots would be a bad influence on children.	**.491**	-.171	.245	.148
24	*Many humanoid robots in society will make it less warm.*	***.452***	*.009*	***.396***	*.144*
13	I can trust persons and organizations related to development of humanoid robots.	-.147	**.777**	.256	-.018
15	Humanoid robots can be very useful for teaching young kids.	-.225	**.737**	.262	.077
10	I don't know why, but I like the idea of humanoid robots.	-.259	**.733**	.044	.295
25	*I trust persons and organizations related to the development of humanoid robots to disclose sufficient information to the public, including negative information.*	*-.015*	***.720***	***.314***	*-.210*
19	Humanoid robots can make our lives easier.	.204	**.672**	-.282	.118
3	Persons and organizations related to development of humanoid robots are well-meaning.	.103	**.672**	-.018	-.054
18	Humanoid robots are a natural product of our civilization.	-.072	**.660**	.083	-.111
28	Persons and organizations related to development of humanoid robots will consider the needs, thoughts and feelings of their users.	.303	**.547**	-.119	.022
5	Humanoid robots can be very useful for caring the elderly and disabled.	.054	**.544**	-.184	.144
6	Humanoid robots should perform repetitive and boring routine tasks instead of leaving them to people.	.123	**.524**	-.053	.200
29	The development of humanoid robots is blasphemous.	-.032	.013	**.892**	.001
9	The development of humanoid robots is a blasphemy against nature.	-.038	.000	**.863**	.077
26	*Technologies needed for the development of humanoid robots belong to scientific fields that humans should not study.*	*-.072*	*.203*	***.663***	*.058*
21	I don't know why, but humanoid robots scare me.	.297	-.205	**.567**	.006
22	*I feel that in the future, society will be dominated by humanoid robots.*	***.314***	***.331***	***.403***	*-.186*
1	*I am afraid that humanoid robots will make us forget what it is like to be human.*	*.234*	*-.097*	***.379***	***.323***
7	People interacting with humanoid robots could sometimes lead to problems in relationships between people.	.240	.049	.292	**.547**
2	*Humanoid robots can create new forms of interactions both between humans and between humans and machines.*	*.010*	***.433***	*-.112*	***.474***
14	Widespread use of humanoid robots would mean that it would be costly for us to maintain them.	.248	.099	.037	**.452**

(Items shown with Italic: reduced based on the criterion of factor loadings more than .3 on more than one item and item analysis)

Table 1. Items of the Frankenstein Syndrome Questionnaire and Results of Factor Analysis

their 50s ($p < .0.13$) and the Japan participants in their 20's ($p < .0.55$). There were neither main effects nor any interactions for "root anxiety toward humanoid robots" (mean = 9.9, *SD* = 4.1).

Correlations with the NARS, Perception of the Relation to the Family, and Commitment to Religions:

The Cronbach's α-coefficients for the NARS subscales were .854, .779, and .842 for S1, S2, and S3, respectively. These values showed that these subscales had sufficient internal consistency.

Table 3 shows Pearson's correlation coefficients between the FSQ subscale scores, the NARS subscale scores, and item scores of relation to family and religious commitment based on the nations and generations. Tests of equality on correlation coefficients found statistically significant differences between the four respondents groups, suggesting the following trends:

		Main Effect			First Order Interaction			Second Order Interaction
		Gender	Nation	Generation	Gender X Nation	Gender X Generation	Nation and Generation	
I. Negative Feelings toward Humanoid Robots	F	6.121	24.630	.406	.027	.444	2.420	.985
	p	.014	< .001	.525	.871	.506	.121	.322
	η^2	.027	.108	.002	.000	.002	.011	.004
II. Expectation for Humanoid Robots	F	2.281	.376	2.013	.185	3.186	4.548	.855
	p	.133	.540	.158	.668	.076	.034	.356
	η^2	.011	.002	.010	.001	.016	.022	.004
III. Root Anxiety toward Humanoid Robots	F	1.877	.676	2.702	1.606	1.437	.264	.019
	p	.172	.412	.102	.207	.232	.608	.891
	η^2	.009	.003	.013	.008	.007	.001	.000

Table 2. Results of ANOVAs for the FSQ Subscale Scores

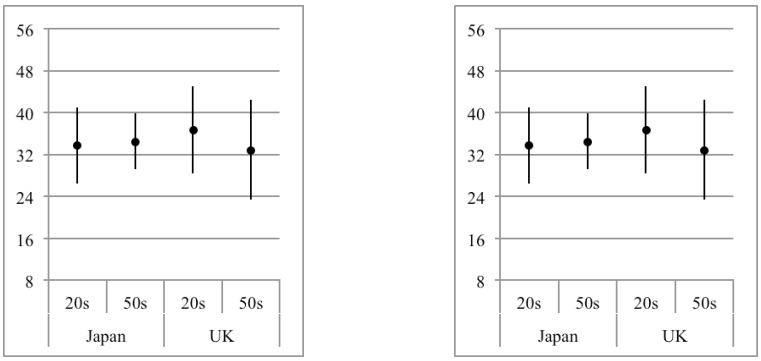


Figure 1. Means and Standard Deviations of Scores of Negative Feelings toward and Expectation for Humanoid Robots

- Between "negative feelings toward humanoid robots" and "expectation for humanoid robots" ($\chi^2(3) = 19.677$, $p < .001$): positive correlation in the Japan respondents in their 20s, and negative correlation in the UK respondents in their 50s,
- Between "negative feelings toward humanoid robots" and "negative attitude toward social influences of robots" ($\chi^2(3) = 11.091$, $p < .05$): moderate levels of correlations in the respondents in their 20s, and strong correlations in the respondents in their 50s,
- Between "negative feelings toward humanoid robots" and "negative attitude toward emotional interaction with robots" ($\chi^2(3) = 14.468$, $p < .01$): moderate levels of positive correlations only in the respondents in their 50s,
- Between "expectation for humanoid robots" and "root anxiety toward humanoid robots" ($\chi^2(3) = 12.840$, $p < .01$): a moderate level of negative correlation only in the UK respondents in their 50s,
- Between "expectation for humanoid robots" and "negative attitude toward social influences of robots" ($\chi^2(3) = 13.715$, $p < .01$): moderate levels of negative correlations only in the respondents in their 50s,
- Between "root anxiety toward humanoid robots" and "expectation for humanoid robots" ($\chi^2(3) = 11.770$, $p < .01$): strong correlation in the Japan respondents in their 20's, and moderate levels of correlations in the other respondents,
- Between "root anxiety toward humanoid robots" and "negative attitude toward emotional interaction with robots" ($\chi^2(3) = 8.279$, $p < .05$): a moderate level of positive correlation only in the UK respondents in their 50s.

On the other hand, there were moderate levels of positive correlations between "negative feelings toward humanoid robots" and "root anxiety toward humanoid robots", between "negative feelings toward humanoid robots" and "negative attitude toward interaction with robots", and between "root anxiety toward humanoid robots" and "negative attitude toward interaction with robots". Moreover, there was a moderate level of negative correlation between "expectation for humanoid robots" and "negative attitude toward social influences of robots".

There were no correlations between the FSQ subscale scores, and perception of the relation to the family and commitment to religions, although only the UK participants in 50's showed statistically significant correlations between these scores and perception of the relation to the family.

4. DISCUSSION

4.1 Findings:

The survey results suggest sufficient correlations between the FSQ subscale scores and NARS. It supports the criterion-related validity of the FSQ. Negative attitude toward interaction with

		FSQII	FSQIII	NARSS1	NARSS2	NARSS3	Religion	Family
FSQI	Whole	-.059	.472**	.426**	.664**	.139	.012	-.081
	Jp 20s	.381**	.534**	.316*	.605**	-.117	.001	-.179
	Jp 50s	-.234	.617**	.431**	.744**	.411**	.143	.196
	UK 20s	.149	.474**	.446**	.478**	-.049	-.133	-.147
	UK 50s	-.402**	.431**	.516**	.820**	.461**	.121	.223
FSQII	Whole		-.208**	-.076	-.169*	-.554**	-.095	-.182**
	Jp 20s		.125	.008	.186	-.383**	.047	-.155
	Jp 50s		-.182	-.159	-.307*	-.473**	-.022	-.157
	UK 20s		-.195	-.037	-.064	-.698**	-.247	-.007
	UK 50s		-.544**	-.261	-.487**	-.584**	-.079	-.317*
FSQIII	Whole			.620**	.526**	.089	.034	.054
	Jp 20s			.734**	.757**	-.113	-.113	-.101
	Jp 50s			.604**	.391**	.191	.034	.233
	UK 20s			.588**	.345*	.020	.124	-.070
	UK 50s			.562**	.593**	.420**	.138	.308*

FSQI: Negative Feelings toward Humanoid Robots, FSQII: Expectation for Humanoid Robots,
FSQIII: Root Anxiety toward Humanoid Robots,
NARSS1: Negative Attitude toward Interaction with Robots, NARSS2: Negative Attitude toward Social Influences of Robots,
NARSS3: Negative Attitude toward Emotional Interaction with Robots,
Religion: Religious Commitment, Family:Relation to Family

Table 3. Pearson's Correlation Coefficients between FSQ and NARS Subscale Scores, and Item Scores of Relation to Family and Religious Commitment

robots in general was related to negative feelings and root anxiety toward humanoid robots in both the UK and Japan.

The survey results also suggest some differences on social acceptance of humanoid robots between the two countries. The UK participants felt more negative towards humanoid robots than their Japanese counterparts. In addition, the UK participants in their 20s had more positive expectations for humanoid robots than any other group..

These results suggest some differences dependent on generation, on relationships between social acceptance of humanoid robots and negative attitudes toward robots in general. The correlation between negative attitudes toward emotional interaction with robots and negative feelings toward humanoids was at a moderate level only in 50s people. The correlation between negative attitude toward social influences of robots and expectation for humanoids also had the similar trend. The correlation between negative attitude toward emotional interaction with robots and root anxiety toward humanoids was at a moderate level only in UK participants in their 50s.

4.2 Implications:

The results in the survey imply that people in the UK have more negative feelings toward humanoid robots than those in Japan. This however, depends on the generation of the participants. Likewise, relationships between feelings toward humanoid robots and attitudes toward robots in general also depend on the generation of respondent. This suggests that changing attitudes toward some particular types of robots may not lead to acceptance of other types of robots, nor robots in general.

In order to further social acceptance of humanoid robots across cultures, designers of robots need to consider individual, generational, and cultural factors in their potential users.

4.3 Limitations and Future Works:

The survey did not take into account concrete attitudes toward the relation to family and religious commitment. It may lead to non-correlation between these factors and social acceptance of robots. On the other hand, previous research has found correlations between these factors and negative attitudes toward robots [16]. It suggests that religious and family factors may indirectly influence social acceptance of humanoid robots. Future surveys need to include this indirect influence in the survey design.

Moreover, the survey did not adopt any image stimulus of robots in order to avoid influences of images of specific types of robots. Future surveys should include more sophisticated items while exploring dominant images of robots in the corresponding nations.

In addition, the survey did not consider possible differences between human attitudes toward humanoid robots measured in questionnaires and live interactions with them, such as dealt with by Wang, et al. [17]. We need to conduct experiments to investigate how psychological constructs measured by the FSQ affect human behaviors toward humanoid robots in real situations.

REFERENCES

[1] V. Evers, H. Maldonado, T. Brodecki, P. Hinds, 'Relational vs. group self-construal: Untangling the role of national culture in HRI', in Proc. ACM/IEEE International Conference on Human-Robot Interaction, pp.255-262, (2009).

[2] D. Li, P. L. P. Rau, and Y. Li. 'A Cross-cultural Study: Effect of Robot Appearance and Task', International Journal of Social Robotics, 2(2), 175-186, (2010).

[3] G. Eresha, M. Haring, B. Endrass, E. Andre, and M. Obaid, 'Investigating the influence of culture on proxemic behaviors for humanoid robots', in Proc. IEEE International Symposium on Robot and Human Interactive Communication, pp.430-435, (2013).

[4] M. Makatchev, R. Simmons, M. Sakr, and M. Ziadee, 'Expressing ethnicity through behaviors of a robot character', in Proc. ACM/IEEE International Conference on Human-Robot Interaction, pp.357-364, (2013).
[5] H-P. Yueh, and W. Lin, 'The Interaction between Human and the Home Service Robot on a Daily Life Cycle', in Proc. International Conference on Cross-Cultural Design, pp.175-181, (2013).
[6] Y. Ho, E. Sato-Shimokawara, T. Yamaguchi, and N. Tagawa, 'Interaction robot system considering culture differences', in Proc. IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), pp.170-174, (2013).
[7] D. S. Syrdal, T. Nomura, H. Hirai, and K. Dautenhahn, 'Examining the Frankenstein Syndrome: An Open-Ended Cross-Cultural Survey', in Social Robotics: Third International Conference, ICSR 2011, Proceedings (B. Mutlu, et al. (Eds)), pp.125-134, Springer, (2011).
[8] D. S. Syrdal, T. Nomura, and K. Dautenhahn, 'The Frankenstein Syndrome Questionnaire - Results from a Quantitative Cross-Cultural Survey', in Social Robotics: 5th International Conference, ICSR 2013 (G. Herrmann, et al. (Eds)), pp.270-279, Springer, (2013).
[9] F. Kaplan, 'Who is afraid of the humanoid?: Investigating cultural differences in the acceptance of robots', International Journal of Humanoid Robot., 1(3), 465–480, (2004).
[10] B. E. Rollin, The Frankenstein Syndrome: Ethical and Social Issues in the Genetic Engineering of Animals, Cambridge University Press, (1995).
[11] T. Nomura, K. Sugimoto, D. S. Syrdal, and K. Dautenhahn, 'Social Acceptance of Humanoid Robots in Japan: A Survey for Development of the Frankenstein Syndrome Questionnaire', in Proc. IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012), pp.242-247, (2012).
[12] T. Nomura, D. S. Syrdal, and K. Dautenhahn, 'Cultural Difference on Factor Structure in a Questionnaire about Humanoid Robots', presented in Workshop on Culture Aware Robotics, May, (2014).
[13] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, 'Measurement of Negative Attitudes toward Robots', Interaction Studies, 7(3), 437-454, (2006).
[14] T. Nomura, T. Tasaki, T. Kanda, M. Shiomi, H. Ishiguro, and N. Hagita, 'Questionnaire-Based Social Research on Opinions of Japanese Visitors for Communication Robots at an Exhibition', AI and Society, 21(1-2), 167-183, (2007).
[15] M. Otsuka, M. Akiyama, K. Mori, and H. Hoshino, 'Comparative Study of Values, Work Ethics, and Lifestyles in Japan and Sweden: An Initial Report', Bulletin of Human Science, 33, 105-119, (2011) (in Japanese).
[16] T. Nomura, 'Comparison on Negative Attitude toward Robots and Related Factors between Japan and the UK', ACM International Conference on Collaboration Across Boundaries: Culture, Distance & Technology, pp.87-90, (2014).
[17] L. Wang, P-L. P. Rau, V. Evers, B. Krisper, and P. Hinds, 'When in Rome: The role of culture & context in adherence to robot recommendations', in Proc. ACM/IEEE International Conference on Human-Robot Interaction, pp. 359–366, (2010).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Presence of Life-Like Robot Expressions Influences Children's Enjoyment of Human-Robot Interactions in the Field

David Cameron[1], Samuel Fernando[2], Emily Collins[1], Abigail Millings[1], Roger Moore[2], Amanda Sharkey[2], Vanessa Evers[3], and Tony Prescott[1]

Abstract. Emotions, and emotional expression, have a broad influence on the interactions we have with others and are thus a key factor to consider in developing social robots. As part of a collaborative EU project, this study examined the impact of life-like affective facial expressions, in the humanoid robot Zeno, on children's behavior and attitudes towards the robot. Results indicate that robot expressions have mixed effects depending on the gender of the participant. Male participants showed a positive affective response, and indicated greater liking towards the robot, when it made positive and negative affective facial expressions during an interactive game, when compared to the same robot with a neutral expression. Female participants showed no marked difference across two conditions. This is the first study to demonstrate an effect of life-like emotional expression on children's behavior in the field. We discuss the broader implications of these findings in terms of gender differences in HRI, noting the importance of the gender appearance of the robot (in this case, male) and in relation to the overall strategy of the project to advance the understanding of how interactions with expressive robots could lead to task-appropriate symbiotic relationships.

1 INTRODUCTION

A key challenge in human robot interaction (HRI) is the development of robots that can successfully engage with people. Effective social engagement requires robots to present engaging personalities [1] and to dynamically respond to and shape their interactions to meet human user needs [2].

The current project seeks to develop a biologically grounded [3] robotic system capable of meeting these requirements in the form of a socially-engaging *Synthetic Tutoring Assistant* (STA). In developing the STA, we aim to further the understanding of human-robot symbiotic interaction where symbiosis is defined as the capacity of the robot, and the person, to mutually influence each other in a positive way. Symbiosis, in a social context, requires that the robot can interpret, and be responsive to, the behavior and state of the person, and adapt its own actions appropriately. By applying methods from social psychology we aim to uncover key factors in robot personality, behavior, and appearance that can promote symbiosis. We hope that this work will also contribute to a broader theory of human-robot bonding that we are developing drawing on comparisons with our psychological understanding of human-human, human-animal and human-object bonds [4].

A key factor in social interaction is the experience of emotions [5]. Emotions provide important information and context to social events and dynamically influence how interactions unfold over time [6]. Emotions can promote cooperative and collaborative behavior and can exist as shared experiences, bringing individuals closer together [7]. Communication of emotion can be thought of as a request for others to acknowledge and respond to our concerns and to shape their behaviors to align with our motives [8]. Thus emotional expression can be important to dyadic interactions, such as that between a teacher and student, where there is a need to align goals.

Research with a range of robot platforms has demonstrated the willingness of humans to interpret robot expressive behavior – gesture [9], posture [10], and facial expression [1] – as affective communication. The extent to which robot expression will promote symbiosis will depend, however, on how well the use of expression is tuned to the ongoing interaction. Inappropriate use of affective expression could disrupt communication and be detrimental to symbiosis. Good timing and sending clear signals is obviously important.

Facial expression is a fundamental component of human emotional communication [11]. Emotion expressed through the face is also considered to be especially important as a means for communicating evaluations and appraisals [12]. Given the importance of facial expressions to the communication of human affect, they should also have significant potential as a communication means for robots [13]. This intuition has lead to the development of many robot platforms with the capacity to produce human-like facial expression, ranging from the more iconic/cartoon-like [e.g., 14, 15] to the more natural/realistic [e.g., 16, 17, 18].

Given the need to communicate clearly it has been argued that, for facial expression, iconic/cartoon-like expressive robots may be more appropriate for some HRI applications, for instance, where the goal is to communicate/engage with children [16, 15]. Nevertheless, as the technology for constructing robot faces has become more sophisticated, robots are emerging with richly-expressive life-like faces [16, 17, 18], with potential for use in a range of real-world applications including use with children. The current study arose out of a desire to evaluate one side of this symbiotic interaction – exploring the value of life-like facial expression in synthetic tutoring assistants for children. Whilst it is clear that people can distinguish robot expressions almost as well

[1] Dept. of Psychology, University of Sheffield, S10 2TN, UK.
Email: {d.s.cameron, e.c,.collins, a.millings, t.j.prescott}@sheffield.ac.uk
[2] Dept. of Computer Science, University of Sheffield S10 2TN, UK.
Email: {s.fernando, r.k.moore, a.sharkey}@sheffield.ac.uk
[3] Dept. of Electrical Engineering, Mathematics and Computer Science, University of Twente, NL. Email: v.evers@utwente.nl

as human ones [16, 18], there is little direct evidence to show a positive benefit of life-like expression on social interaction or bonding. Although children playing with an expressive robot are more expressive than those playing alone [19], this finding could be a result of the robot's social presence [20] and not simply due to its use of expression. A useful step toward improving our understanding would be the controlled use of emotional expression in a setting in which other factors, such as the presence of the robot and its physical and behavioral design, are strictly controlled.

In the current study the primary manipulation was to turn on or off the presence of appropriate positive and negative facial expressions during a game-playing interaction, with other features such as the nature and duration of the game, and the robot's bodily and verbal expression held constant. As our platform we employed a Hanson Robokind Zeno R50 [21] which has a realistic silicon rubber ("flubber") face, that can be reconfigured, by multiple concealed motors, to display a range of reasonably life-like facial expressions in real-time (Figure 1).

Figure 1. The Hanson Robokind Zeno R50 Robot with example facial expressions

By recording participants (with parental consent), and through questionnaires, we obtained measures of proximity, human emotional facial expression, and reported affect. We hypothesized that children would respond to the presence of facial expression by (a) reducing their distance from the robot, b) showing greater positive facial expression themselves during the interaction, and c) reporting greater enjoyment of the interaction compared to peers who interacted with the same robot but in the absence of facial expression. Previous studies have shown some influence of demographics such as age and gender on HRI [22, 23, 24]. In our study, a gender difference could also arise due to the visual appearance of the Zeno robot as similar to a male child, which could prompt different responses in male and female children. We therefore considered these other factors as potential moderators of children's responses to the presence or absence of robot emotional expression.

2 METHOD

2.1 Design

Due to the potential of repeated robot exposure prejudicing participants' affective responses, we employed a between-subjects design, such that participants were allocated to either the experimental condition – interaction with a facially expressive robot, or to the control condition of a non-facially-expressive robot. Allocation to condition was not random, but determined by logistics due to the real-world setting of the research. The study took place as part of a two-day special exhibit demonstrating modern robotics at a museum in the UK. Robot expressiveness was manipulated between the two consecutive days, such that visitors who participated in the study on the first day were allocated to the expressive condition, and visitors who participated in the study on the second day were allocated to the non-expressive condition.

2.2 Participants

Children visiting the exhibit were invited to participate in the study by playing a game with Zeno. Sixty children took part in the study in total (37 male and 23 female; M age = 7.57, SD = 2.80). Data were trimmed by age to ensure sufficient cognitive capacity (those aged < 5 were excluded[4]) and interest in the game (those aged >11 were excluded) leaving 46 children (28 male and 18 Female; M age = 8.04, SD = 1.93).

2.3 Measures

Our primary dependent variables were interpersonal responses to Zeno measured through two objective measures: affective expressions and interpersonal distance. Additional measures comprised of a self-report questionnaire, completed by participating children, with help from their parent/carer if required, and an observer's questionnaire, completed by parents/carers.

2.3.1 Objective Measures

Interpersonal distance between the child and the robot over the duration of the game was recorded, using a Microsoft Kinect sensor, and mean interpersonal distance during the game calculated. Participant expressions were recorded throughout the game and automatically coded for discrete facial expressions: Neutral, Happy, Sad, Angry, Surprised, Scared, and Disgusted, using Noldus FaceReader version 5. Mean intensity of the seven facial expressions across the duration of the game were calculated. Participants' game performances (final scores) were also recorded. FaceReader offers automated coding of expressions at an accuracy comparable to trained raters of expression [25].

2.3.2 Questionnaires

Participants completed a brief questionnaire on their enjoyment of the game and their beliefs about the extent to which they thought that the robot liked them. Enjoyment of playing Simon Says with Zeno was recorded using a single-item, four-point measure, ranging from 'I definitely did not enjoy it' to 'I really enjoyed it'. Participants' perceptions of the extent to which Zeno liked them single-item on a thermometer scale, ranging from 'I do not think he liked me very much' to 'I think he liked me a lot'. They were also asked if they would like to play the game again. Parents and

[4] Additional reasons for excluding children below the age of 5 were questionable levels of understanding when completing the self-report questionnaires, and low reliability in FaceReader's detection of expressions in young children.

carers completed a brief questionnaire on their perceptions of their child's enjoyment and engagement with the game on single-item thermometer scales, ranging from 'Did not enjoy the game at all' to 'Enjoyed the game very much and 'Not at all engaged' to 'Completely engaged'.

2.4 Procedure

The experiment took place in a publicly accessible lab and prospective participants could view games already underway. Brief information concerning the experiment was provided to parents or carers and informed consent was obtained from parents or carers prior to participation.

During the game, children were free to position themselves relative to Zeno within a 'play zone' boundary marked on the floor by a mat (to delineate the area in which the system would correctly detect movements) and could leave the game at their choosing. The designated play zone was marked by three foam .62msq mats. The closest edge of the play zone was 1.80m from the robot and the play zone extended to 3.66m away. These limits approximate the 'social distance' classification [26]. This range was chosen for 2 reasons i) Participants would likely expect the game used to occur within social rather than public- or personal-distance ii) This enabled reliable recordings of movement by the Kinect sensor. The mean overall distance for the participants from the robot fell well within social-distance boundaries (2.48m).

At the end of the game, participants completed the self-report questionnaire, while parents completed the observer's questionnaire. Participant-experimenter interaction consistency was maintained over the two days by using the same experimenter on all occasions for all tasks.

Interaction with the robot took the form of the widely known *Simon Says* game (Figure 2). This game was chosen for several reasons: children's familiarity with the game, its uncluttered structure allows autonomous instruction and feedback delivery by Zeno, and its record of successful use in a prior field study [27].

The experiment began with autonomous instructions delivered by Zeno as soon as children stepped into the designated play zone in front of the Kinect sensor. Zeno introduced the game by saying, "*Hello. Are you ready to play with me? Let's play Simon Says. If I say Simon Says you must do the action. Otherwise you must keep still*." The robot would then play ten rounds of the game or play until the child chose to leave the designated play zone. In each round, Zeno gave one of three simple action instructions: 'Wave your hands', 'Put your hands up' or 'Jump up and down'. Each instruction was given either with the prefix of 'Simon says' or no prefix.

Figure 2. A child playing Simon Says with Zeno

The OpenNI/Kinect skeleton tracking system was used to determine if the child had performed the correct action in three seconds following instruction. For the 'Wave your hands' action, our system monitored the speed of the hands moving. If sufficient movement for the arms were detected following instruction then the movement was marked as a wave. For the 'Jump up and down' action the vertical velocity of the head was monitored, again with a threshold to determine if a jump had taken place. Finally for the 'Put your hands up' action, our system monitored the positions of the hands relative to the waist. If the hands were found to be above the waist for more than half of the three seconds following the instruction then the action was judged to have been executed. The thresholds for the action detection were determined by previous trial and error during pilot testing in a university laboratory. The resulting methods of action detection were found to be over 98% accurate in our study. In the rare cases where the child did the correct action and the system judged incorrectly then the experimenters would step in and say "*Sorry, the robot made a mistake there, you got it right*".

If children followed the action instruction after hearing 'Simon says' the robot would say, "*Well done, you got that right*". If the child remained still when the prefix was not given, Zeno would congratulate them on their correct action with "*Well done, I did not say Simon Says and you kept still*". Conversely, if the child did not complete the requested movement when the prefix was given Zeno would say, "*Oh dear, I said Simon Says, you should have waved your hands*". If they completed the requested movement in the absence of the prefix, Zeno would inform them of their mistake with, "*Oh dear, I did not say Simon Says, you should have kept still*". Zeno gave children feedback of a running total of their score at the end of each round (the number of correct turns completed).

If the child left the play zone before ten rounds were played, the robot would say, "*Are you going? You can play up to ten rounds. Stay on the mat to keep playing*". The system would then wait three seconds before announcing, "*Goodbye. Your final score was (score)*". This short buffer was to prevent the game ending abruptly if the child accidentally left the play zone for a few seconds.

At the end of the ten rounds, the robot would say, "*All right, we had ten goes. I had fun playing with you, but it is time for me to play with someone else now. Goodbye*."

The sole experimental manipulation coincided with Zeno's spoken feedback to the children after each turn. In the expressive robot condition, Zeno responded with appropriate 'happiness' or 'sadness' expressions, following children's correct or incorrect responses. These expressions were prebuilt animations, provided with the Zeno robot, named 'victory' and 'disappointment' respectively. These animations were edited to remove gestures so only facial expression were present. In contrast, in the non-expressive robot condition, Zeno's expressions remained in a neutral state regardless of child performance. Previous work indicates that children can recognize these facial expression representations by the Zeno robot with a good degree of accuracy [28].

3 RESULTS

A preliminary check was run to ensure even distribution of participants to expressive and non-expressive conditions. There were 9 female and 16 male participants in the expressive

condition and 9 female and 12 male participants in the non-expressive condition. A chi square test was run before analysis to check for even gender distribution across conditions indicates no significant difference (X2 (1,48) = 2.25, p = .635).

3.1 Objective Measures

Overall, we did not observe any significant main effects of Zeno's expressiveness on objective measures of interpersonal distance or facial expressions between conditions. However, there were significant interaction effects, when gender was included as a variable.

There was a significant interaction of experimental condition and child's gender on average child's expressions of happiness $F(1,39) = 4.75$, $p = .038$. While male participants showed greater average happiness in the expressive robot condition in comparison to those in the non-expressive condition (19.1%, SE 3.3% versus 5.3%, SE 4.1%), female participants did not differ between conditions (7.4%, SE 4.3% versus 12.6%, SE 4.6%). Simple effects tests (with Bonferroni correction) indicated that the observed differences between conditions for male participants was significant (p = .012).

A contrasting interaction was found for average expressions of surprise $F(1,39) = 5.16$, $p = .029$. Male participants in the expressive robot condition showed less surprise than those in the non-expressive condition (6.1%, SE 3.2% versus 19.6%, SE 4.0%), whereas female participant expressions for surprise did not differ between conditions (11.9%, SE 4.2% versus 7.1%, SE 4.5%). There were no further significant interactions for any of the remaining expressions.

There was a near significant interaction for experimental condition and child's gender for interpersonal distance $F(1,41) = 2.81$, $p = .10$ (Figure 3). Male participants interacting with the expressive robot tended to stand closer (M = 2.28m, SE .10m) than did those interacting with the non-expressive robot (M = 2.57m, SE .13m), whereas female participants interacting with the expressive robot tended to stand further away (M = 2.59m, SE .14m) than those interacting with the non-expressive robot (M = 2.45m, SE .14m). A follow-up simple effect test indicates that the difference between conditions for male participants was also near significant (p = .086).

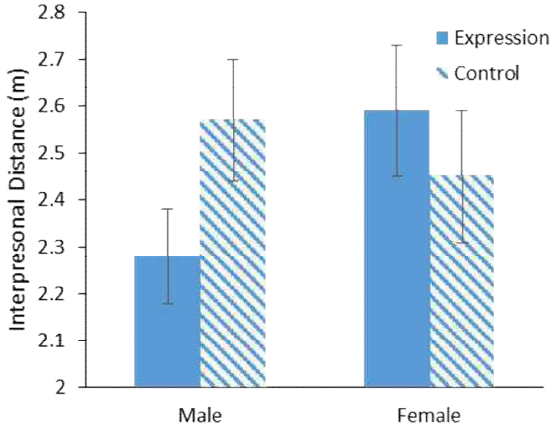


Figure 3. Mean interpersonal distance during game

Controlling for participant age or success/failure in the game made no material difference to any of the objective measures findings.

3.2 Questionnaires

No significant main effects of condition were seen for self-reported measures or observer reported measures. However, there were significant gender effects, and significant gender X condition effects. Gender had a main effect on children's beliefs about the extent to which the robot liked them $F(1,38) = 5.53$, $p = 0.03$. Female participants reported significantly lower ratings (M = 3.08, SE .34) than did male participants (M = 4.17, SE .31).

We observed a significant interaction of gender and experimental condition for participants' enjoyment in interacting with Zeno $F(1,38) = 4.64$, $p = .04$. Male participants interacting with the expressive Zeno reported greater enjoyment of the interaction than those who interacted with the non-expressive Zeno (M = 3.40, SE .18 versus M = 3.00, SE .23), whereas female participants interacting with the expressive Zeno reported less enjoyment than those interacting with the non-expressive Zeno (M = 3.22, SE .23 versus M = 3.78, SE .23). Simple effects tests did not indicate that the difference found between conditions were significant for either male participants ($p > .10$) or female participants ($p > .10$).

Results from the observer reports generated by the participants' parents or carers showed the same trends as those from the self-report results but did not show significant main or interaction effects. Controlling for participant age or success/failure in the game made no material difference to any of the questionnaire data findings.

4 DISCUSSION

The results provide new evidence that life-like facial expressions in humanoid robots can impact on children's experience and enjoyment of HRI. Moreover, our results are consistent across multiple modalities of measurement. The presence of expressions could be seen to cause differences in approach behaviors, positive expression, and self-reports of enjoyment. However, the findings are not universal as boys showed more favorable behaviors and views towards the expressive robot compared to the non-expressive robot, whereas girls tended to show the opposite.

Sex differences towards facially expressive robots during HRI could have profound impact on the design and development of future robots; it is important to replicate these experimental conditions and explore these results in more depth in order to identify why these results arise. At this stage, the mechanisms underpinning these differences still remain to be determined. We outline two potential processes that could explain our results.

The current results could be due to children's same-sex preferences for friends and playmates typically exhibited at the ages range tested (ages five to ten) [29]. Zeno is nominally a 'boy' robot and expressions may be emphasizing cues seen on the face to encourage user perceptions of it as a boy. As a result, children may be acting in accordance with existing preferences for play partners [30]. If this is the case, it would be anticipated that replication of the current study with a 'girl' robot counterpart would produce results contrasting with the current findings.

Alternatively, results could be due to the robot's expressions emphasizing the existing social situation experienced by the children. The current study took place in a publically accessible space, with participants in the company of museum visitors, other volunteers, and the children's parents or carer. Results from the current study could represent children's behavior towards the robot based on existing gender driven behavioral attitudes. Girls

may have felt more uncomfortable than boys when in front of their parents whilst engaging in explorative play [20] with a strange person (in the form of their perceived proximity to the experimenter) and an unfamiliar object (the robot). Social cues from an expressive robot, absent in a neutral robot, may reinforce these differences through heightening the social nature of the experiment.

Behavioral gender differences in children engaging in public or explorative play are well established, and the link between these gender differences and the influence of direct parents/carers differential socialization of their children dependent upon the sex [31,32], is a further established link of developmental study. To better explore the gender difference observed in our study we must take into consideration existing observed behavioral patterns in children engaging in explorative play around their parents. Replication in a familiar environment away from an audience including children's parents may then impact on apparent sex differences observed in the current HRI study.

The current study is a small-sample field experiment. As with the nature of field studies, maintaining an exacting control over experimental conditions is prohibitively difficult. Along with possible confounds from the public testing space, the primary experimenter knew the condition each child was assigned to; despite best efforts in maintaining impartiality, the current study design cannot rule out potential unconscious experimenter influence on children's behaviors. In studies concerning emotion and expression, potential contagion effects of expression and emotion [33] could impact on participant's expressions and reported emotions. The current results therefore offer a strong indication of the areas to be further explored under stricter experimental conditions.

We aim to repeat the current study in a more controlled experimental environment. Children will complete the same Simon-says game in the familiar environment of their school, this time without an audience. Rather than allocation by day to condition, the study protocol will be modified to randomly allocate children to conditions, and the study will be conducted by an experimenter naïve to conditions. Testing at local schools offers better controls over participant sample demographics as children can be recruited based on age and having similar educational and social backgrounds. The environment of this study also removes any direct influence by the presence of parents/carers. Thus, a repeat of the current study under stricter conditions also offers opportunity to further test the proposed hypotheses for the observed sex differences in enjoyment in interacting with a facially expressive robot.

We have previously proposed that human-robot bonds could be analyzed in terms of their similarities to different types of existing bond with other human, animals, and objects [4]. Our relationships with robots that are lacking in human-like faces may have interesting similarities to human-animal bonds which can be simpler than those with other people—expectations are clearer, demands are lower, and loyalty is less prone to change. Robots with more human-like faces and behavior, on the other hand, may prompt responses from users that include more of the social complexities of human-human interaction. Thus, aspects of appearance that indicate gender can become more important, subtleties of facial and vocal expression may be subjected to greater scrutiny and interpretation. Overall, as we progress towards more realistic human-like robots we should bear in mind that whilst the potential is there for a richer expressive vocabulary, the bar may also be higher for getting the communication right.

5 CONCLUSION

This paper offers further steps towards developing a theoretical understanding of symbiotic interactions between humans and robots. The production of emulated emotional communication through facial expression by robots is identified as a central factor in shaping human attitudes and behaviors during HRI. Results from both self-repot and objective measures of behavior point towards possible sex differences in responses to facially expressive robots; follow-up work to examine these is identified. These findings highlight important considerations to be made in the future development of a socially engaging robot.

6 ACKNOWLEDGMENTS

This work is supported by the European Union Seventh Framework Programme (FP7-ICT-2013-10) under grant agreement no. 611971. We wish to acknowledge the contribution of all project partners to the ideas investigated in this study.

7 REFERENCES

[1] Breazeal, C., & Scassellati, B. How to build robots that make friends and influence people. In *Intelligent Robots and Systems, 1999. IROS'99. Proceedings*. 1999 IEEE/RSJ International Conference on (Vol. 2, pp. 858-863). IEEE.

[2] Pitsch, K., Kuzuoka, H., Suzuki, Y., Sussenbach, L., Luff, P., & Heath, C. "The first five seconds": Contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. *In Robot and Human Interactive Communication*, 2009. RO-MAN 2009. The 18th IEEE International Symposium on (Toyama, Japan, Sept 27, 2009) IEEE 985-991 DOI:10.1109/ROMAN.2009.5326167

[3] Verschure, P. F. Distributed adaptive control: a theory of the mind, brain, body nexus. *Biologically Inspired Cognitive Architectures*, **1**, 55-72, (2012). DOI:10.1016/j.bica.2012.04.005

[4] Collins, E. C., Millings, A., Prescott, T. J. 2013. Attachment in assistive technology: A new conceptualisation. In *Assistive Technology: From Research to Practice*, Encarnação, P., Azevedo,L., Gelderblom, G. J., Newell, A., & Mathiassen N. IOS Press, 823-828. DOI:10.3233/978-1-61499-304-9-823

[5] Van Kleef, G. A. How emotions regulate social life the emotions as social information (EASI) model. *Current Directions in Psychological Science*, **18**, 184-188, (2009). DOI:10.1111/j.1467-8721.2009.01633.x

[6] Hareli, S., & Rafaeli, A. Emotion cycles: On the social influence of emotion in organizations. *Research in organizational behavior*, **28**, 35-59, (2008). DOI:10.1016/j.riob.2008.04.007

[7] Kelly, J. R., & Barsade, S. G. Mood and emotions in small groups and work teams. *Organizational behavior and human decision processes*, **86**, 99-130, (2001). DOI:10.1006/obhd.2001.2974

[8] Parkinson, B. Do facial movements express emotions or communicate motives. *Personality and Social Psychology Review*, **9**, 278-311, (2005). DOI = 10.1207/s15327957pspr0904_1

[9] Tielman, M., Neerincx, M., Meyer, J., & Looije, R. Adaptive emotional expression in robot-child interaction. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, (Bielefeld, Germany, Mar. 03 – 06, 2014) ACM, New York, NY, 407-414. DOI:10.1145/2559636.2559663.

[10] Beck, A., Cañamero, L., Damiano, L., Sommavilla, G., Tesser, F., & Cosi, P. Children interpretation of emotional body language displayed by a robot. *Social Robotics*, 62–70. (2011) Springer, Berlin Heidelberg.

[11] Buck, R. W., Savin, V. J., Miller, R. E., & Caul, W. F. Communication of affect through facial expressions in humans.

Journal of Personality and Social Psychology, **23**, 362-371, (1972). DOI:10.1037/h0033171

[12] Parkinson, B. Emotions are social. *British Journal of Psychology*, **87**, 663-683, (1996). DOI:10.1111/j.2044-8295.1996.tb02615.x

[13] Nitsch, V., & Popp, M. Emotions in robot psychology. *Biological cybernetics*, 1-9, (2014). DOI: 10.1007/s00422-014-0594-6

[14] Breazeal, C. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, **59**, 119-155, (2003). DOI:10.1016/S1071-5819(03)00018-1

[15] Espinoza, R.R , Nalin, M., Wood, R., Baxter, P., Looije, R., Demiris, Y., ... & Pozzi, C. Child-robot interaction in the wild: advice to the aspiring experimenter. In *Proceedings of the 13th international conference on multimodal interfaces* (Alicante, Spain, Nov. 14 – 18, 2011) ACM, New York, NY, 335-342. DOI:10.1145/2070481.2070545

[16] Becker-Asano, C., & Ishiguro, H. Evaluating facial displays of emotion for the android robot Geminoid F. In *Affective Computational Intelligence* (WACI), 2011 IEEE Workshop on (Paris, France, April, 11-15, 2011) IEEE 1-8 DOI:10.1109/WACI.2011.5953147

[17] Fagot, B. I. The influence of sex of child on parental reactions to toddler children. *Child development*, **2**, 459-465, (1978). DOI:jstor.org/stable/1128711

[18] Mazzei, D., Lazzeri, N., Hanson, D., & De Rossi, D. HEFES: An Hybrid Engine for Facial Expressions Synthesis to control human-like androids and avatars. In *Biomedical Robotics and Biomechatronics* (BioRob), 2012 4th IEEE RAS & EMBS International Conference on (Rome, Italy, Jun. 24 -27, 2012) IEEE 195-200 DOI:10.1109/BioRob.2012.6290687

[19] Shahid, S., Krahmer, E., & Swerts, M. Child–robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend? *Computers in Human behaviour*, **40**, 86-100, (2014). DOI:10.1016/j.chb.2014.07.043.

[20] Kraut, R. E., & Johnston, R. E. Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology*, **37**, 1539-1553, (1979). DOI:10.1037/0022-3514.37.9.1539

[21] Hanson, D., Baurmann, S., Riccio, T., Margolin, R., Dockins, T., Tavares, M., & Carpenter, K.. Zeno: A cognitive character. *AI Magazine*, 9-11, (2009).

[22] Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, **19**, 61-84, (2004). DOI:10.1207/s15327051hci1901&2_4

[23] Kuo, I. H., Rabindran, J. M., Broadbent, E., Lee, Y. I., Kerse, N., Stafford, R. M. Q., & MacDonald, B. A. Age and gender factors in user acceptance of healthcare robots. In *Robot and Human Interactive Communication*, 2009. RO-MAN 2009. The 18th IEEE International Symposium on (Toyama, Japan, Sept. 27- Oct. 2, 2009) IEEE. 214-219 DOI:10.1109/ROMAN.2009.5326292

[24] Shahid, S., Krahmer, E., Swerts, M., & Mubin, O. Child-robot interaction during collaborative game play: Effects of age and gender on emotion and experience. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (Brisbane, Australia. November 22-26, 2010) ACM, New York, USA. 332-335 DOI:10.1145/1952222.1952294

[25] Lewinski, P., den Uyl, T. M., & Butler, C. Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics*, **7** 227-236, (2014). DOI:10.1037/npe0000028

[26] Burgess, J. Interpersonal spacing behavior between surrounding nearest neighbors reflects both familiarity and environmental density. *Ethology and sociobiology*, **4**, 11-17, (1983). doi:10.1016/0162-3095(83)90003-1

[27] Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Kose-Bagci, H., Mirza, N. A., & Blow, M. KASPAR–a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics*, **6**, 369-397, (2009). DOI:10.1080/11762320903123567

[28] Costa, S., Soares, F., & Santos, C. Facial Expressions and Gestures to Convey Emotions with a Humanoid Robot. *In Social Robotics* 542-551). Springer International Publishing. (2013). DOI:10.1007/978-3-319-02675-6_54

[29] Martin, C. L., & Fabes, R. A. The stability and consequences of young children's same-sex peer interactions. *Developmental psychology*, **37**, 431-446, (2001). DOI:10.1037/0012-1649.37.3.431.

[30] Lindsey, E. W. Physical activity play and preschool children's peer acceptance: Distinctions between rough-and-tumble and exercise play. *Early Education and Development*, **25**, 277-294, (2014). DOI:10.1080/10409289.2014.890854

[31] Gonzalez, A. M. *Parenting Preschoolers with Disruptive Behavior Disorders: Does Child Gender Matter*? Dissertation. Washington (2013) University in St. Louis, St Louis, Missouri, USA.

[32] Kim, H. J., Arnold, D. H., Fisher, P. H., & Zeljo, A. Parenting and pre-schoolers' symptoms as a function of child gender and SES. *Child & Family Behavior Therapy*, **27**, 23-41, (2005). DOI: 10.1300/J019v27n02_03

[33] Hatfield, E., Cacioppo, J. T. & Rapson, R. L. *Emotional Contagion.* Cambridge university press, Cambridge, UK, 1994.

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

The Paro robot seal as a social mediator for healthy users

Natalie Wood[1] and **Amanda Sharkey**[2] and **Gail Mountain**[3] and **Abigail Millings**[4]

Abstract. Robots are being designed to provide companionship, but there is some concern that they could lead to a reduction in human contact for vulnerable populations. However, some field data suggests that robots may have a social mediation effect in human-human interactions. This study examined social mediation effects in a controlled laboratory setting. In this study 114 unacquainted female volunteers were put in pairs and randomised to interact together with an active Paro, an inactive Paro, or a dinosaur toy robot. Each pair was invited to evaluate and interact with the robot together during a ten minute session. Post-interaction questionnaires measured the quality of dyadic interaction between participants during the session. Our results indicate that the strongest social mediation effect was from the active Paro.

1 INTRODUCTION

Over the last decade robots have been developed as an alternative to companion animals for older-aged adults and people with dementia in care homes. These companion robots are designed to improve the physical and psychological health of users by calming them, providing companionship, and have the potential to help reduce loneliness and improve the well-being of their users [11, 2].

Despite the benefits these assistive robots bring, there are objections to their use with vulnerable populations. Sparrow and Sparrow [15] raise one main concern as the loss of human contact had by these populations as their human carers are replaced with robotic counterparts. They argue that robotic technology is not currently capable of meeting the social and emotional needs of their users. As the amount of human-human contact between patients and their carers decreases, this could lead to a reduction in the number and quality of their social relationships, and therefore their quality of life.

This concern is supported by Sharkey and Sharkey [13], who consider the negative effects of reduced social contact on the physical and psychological well-being of the elderly. They propose that access to human social contact must be considered before robotic technology is brought into elder-care.

However, a recent developing area of research has shown that robotics can have a role in improving human-human relationships. This small but growing body of field data suggests that a companion robot, the Paro robot seal, can be used to encourage social interaction between individuals, in addition to providing human-robot companionship.

The majority of these studies examined the social mediation effect of Paro using samples of people with cognitive impairment in care home settings.

This paper aims to contribute to this research by investigating whether the social mediation effect is present in healthy populations and under controlled conditions. Animals have been found to act as a social catalyst for healthy individuals as well as for people with dementia and older adults [5][9]. We propose that the same could be true of animal-like robots. Our study looks at the ability of Paro to mediate social interaction between strangers by providing an ice breaker effect in a controlled laboratory setting.

Section 1.1 of this paper introduces the existing work on social mediation with Paro. Section 2 details our hypotheses. This is followed by the methodology used for the study in section 3. Our analytic strategy and results are discussed in section 4. We discuss our findings and limitations of the work in section 5. Finally section 6 concludes the paper.

1.1 Background

Previous studies conducted in care homes have reported the ability of Paro as a social mediator. A randomised controlled trial by Robinson, Macdonald, Kerse, and Broadbent [12] showed a significant decrease in the loneliness reported by 17 residents of a retirement home after 12 weeks of regular activity with Paro. They also found an increase in social interaction between residents when they engaged in activity with Paro compared to during normal activities with and without the resident dog.

Wada and Shibata [19] found that the social network of 12 elderly residents in a care home increased after Paro was available in an open public space for two months.

In an ethnographic case study, Giusti and Marti [4] found that not only did the amount of social interaction increase, but the social dynamic between three residents of a nursing home changed from primarily one-to-one social interactions to group interaction involving all three during interactions with Paro.

Kidd, Taggart and Turkle [7] investigated the effect that a small number of interactions with Paro had on social activity in the nursing home setting. They found that the 23 residents reported more social interaction with others when they were with active Paro than when it was turned off. They also found that presence of more people, including caregivers and experimenters, improved the amount of social engagement.

These findings were supported in another nursing home where Šabanović et al. [18] observed that the social interactions increased between seven residents, including those who were not directly interacting with Paro, during robot-assisted therapy sessions.

[1] Sheffield Robotics, University of Sheffield, S1 3JD, UK. Email: natalie.wood@shef.ac.uk.
[2] Department of Computer Science, University of Sheffield, S1 4DP, UK. Email: a.sharkey@dcs.shef.ac.uk.
[3] School of Health and Related Research, University of Sheffield, S1 4DA, UK. Email: g.a.mountain@sheffield.ac.uk
[4] Centre for Assistive Technology & Connected Healthcare, and Department of Psychology, University of Sheffield, S10 2TP, UK. Email: a.millings@sheffield.ac.uk.

Although the results of these studies show support for Paro as a social mediator in the nursing home setting, they are limited by small sample sizes. In addition, the majority of these studies lack control conditions, such that the social meditation effect cannot be attributed specifically to the Paro. It is unclear whether any novel, robotic stimuli would produce the effects observed. In the current study, we examine the social mediation effect of an active Paro which is turned on and interactive, compared to that of an inactive Paro which is turned off and resembling a cuddly toy, and another interactive robotic toy, Pleo the dinosaur.

2 HYPOTHESES

This study aims to answer the following questions: Can the social mediation effect of Paro apply to a healthy population? Can the effect be measured under a controlled laboratory setting?

To investigate the social mediation effect of Paro we invited pairs of strangers to interact for the first time together, along with an active Paro, an inactive Paro, or a Pleo.

We anticipate that the social mediation effect of Paro when active will lead to participants enjoying interacting with the other participant more and having a better experience when interacting together, than with an inactive Paro and the Pleo. We also anticipate that interacting together with an active Paro will lead to a more positive opinion of the other participant compared to the other two conditions.

Secondary to this we also expect the Pleo to be a more effective social mediator than an inactive Paro. This leads to our hypotheses: Primary hypotheses:

- H1: Compared to the Pleo and inactive Paro conditions, the participants in the active Paro condition will report a:
  - (a): higher quality of interaction.
  - (b): higher opinion of the other participant.

Secondary hypotheses:

- H2: Compared to the inactive Paro condition, the participants in the Pleo condition will report a:
  - (a): higher quality of interaction.
  - (b): higher opinion of the other participant.

3 METHODOLOGY

3.1 Participants

Participants were recruited using a number of methods. Firstly, undergraduate psychology students were invited to participate through the University's research participation scheme in exchange for course credit. Secondly, an email was sent using volunteer mailing lists for University of Sheffield staff and students, inviting volunteers to participate in exchange for entry into a prize draw for one of two £30 Amazon vouchers. Female participants were chosen due to the availability of volunteers at the university which were predominantly female at the time.

In total 114 participants were recruited, aged from 15 to 59 ($M = 23.94, SD = 8.38$), and were paired according to availability. Pairs of participants were randomly allocated into conditions with 21 participant pairs in the active Paro condition, 19 participant pairs in the inactive Paro condition, and 17 participants pairs in the Pleo condition.

3.2 Materials

3.2.1 Paro

The Paro was developed in Japan by Shibata [21] as a therapeutic tool for use with people with dementia. It is a pet-like robot based on a harp seal pup and its body is covered in soft, white, and antibacterial fur. It uses a number of sensors for touch and sound to detect interaction. The robot responds to the stimulation of interaction by making noises and moving.

3.2.2 Pleo dinosaur robot

The Pleo [1] is a commercially available pet dinosaur toy which was designed to have a lifelike appearance and adaptive behaviours. The 2008 model used in the experiment has a number of touch sensors on its head, chin, shoulders, back and feet, and audio and light sensors in its head. A range of actuators means it can respond to different types of interaction in different ways. The Pleo is covered with plastic which feels rubbery to touch.

3.2.3 Measures

All measures except the pen-and-paper evaluation form were administered via an online questionnaire on a tablet.

Quality of interaction with the other This was measured using items about how the participant felt during the interaction with the other person, and how the participant perceived the interaction itself:

Participants reported feelings experienced during the interaction by rating eight items from Leary, Kowalski, & Bergen [8] on a 7-point Likert scale from 1 (*not at all*) to 7 (*very much*). Factor analysis [5] reduced these items to two composite measures: *'relaxed', 'awkward', 'nervous'*, and *'confident'* loaded highly onto a factor of 'Confidence' during the interaction ($\alpha = .81$). *'Accepted', 'respected', 'disrespected'*, and *'rejected'* loaded onto a factor of ' Feeling Acceptance' during the interaction ($\alpha = .76$).

How the interaction was perceived was measured using 16 items adapted from Berry and Hansen[3], rated on a 7-point Likert scale from 1 (*not at all*) to 7 (*very much*). Factor analysis reduced these 16 items to four composite measures. First *'relaxed', 'smooth',* and *'natural'* loaded onto how 'Comfortable' the interaction felt ($\alpha = .84$). Secondly *'enjoyable', 'fun', 'pleasant', satisfying', 'intimate'*, and *'boring'* loaded onto a factor of the interaction 'Feeling Positive' ($\alpha = .86$). The third factor had loadings of *'upsetting', 'unpleasant'*, and *'annoying'* on a factor of the interaction 'Feeling Negative' ($\alpha = .65$). Finally *'forced', 'awkward', 'reserved'*, and *'strained'* loaded onto a factor of 'Difficulty' of the interaction ($\alpha = .86$).

Opinion of the other participant Participants answered the following questions adapted from Sprecher, Treger, Wondra, Hilaire, and Wallpe[16] about the interaction with the other participant and about the other participant on a 7-point Likert scale from 1 (*not at all*) to 7 (*very much*).

Liking of the other was measured with three items: '*How much did you like the other participant?*', '*How much would you like to interact with the other participant again?*', and '*How likeable did you find the other participant?*' ($\alpha = .86$)

Closeness to the other was measured with a single item: '*How close do you feel toward the other participant*?'

[5] Factor analysis for the purpose of dimension reduction was conducted using principal component analysis using oblimin rotation with each scale to create composite measures.

Perceived similarity was measured with two items: '*How much do you think you have in common with the other participant?*', and '*How similar do you think you and the other participant are likely to be?*'($\alpha = .86$)

Enjoyment of the interaction: This was measured with a single item:'*How much did you enjoy the interaction with the other participant*?'

Evaluation form The evaluation form consisted of a 10-item questionnaire about the robot which participants completed as a dyad. Five of the items were from Shibata, Wada, Ikeda, and Šabanović[14] and asked participants to indicate on a 7-point Likert scale how much they felt the words *'friendly', 'lively', 'expressive', 'natural',* and *'relaxing'* applied to the robot. The other five items were adapted from Wada, Shibata, Musha, and Kimura [20] and asked participants to answer on 7-point Likert scales the questions '*How cute/ugly do you find the robot?*', '*How much do you like the robot?*', '*How fun/boring is interacting with the robot?*', '*How much more would you want to interact with the robot?*' and '*How much do you want to touch the robot?*'.

3.3 Recording and coding behaviour

The interaction between the participants and the robot was covertly recorded in the experiment room with two Replay digital action cameras. Observed behavioural data will not be reported in this paper but will be detailed elsewhere.

3.4 Procedure

All participants were told that the study aimed to investigate people's opinions of different types of interactive robots, and that they would be asked to interact with and evaluate a robot. Participants were tested in dyads by a female experimenter. On arrival each participant was taken to a separate location to read the information sheet and provide consent to participate. Participants were told that they would meet another participant with whom they would evaluate a robot.

Both participants were first asked to complete a questionnaire (data not included in the current study). At this point the dyad was randomly assigned into either the active Paro, inactive Paro or Pleo conditions. Once both participants had completed the questionnaire, they were introduced to each other (as'the other participant you'll be evaluating the robot with') and together given an explanation of the robot evaluation task they were to undertake.

Participants were told that there would be a robot on the table in the room and were asked to interact with the robot together, in any way they wanted to, but to keep the robot off the floor. In the inactive Paro condition, participants were told that the robot would remain off for the duration of the task and that they would have the opportunity to see it turned on at the end of the session during individual debriefings. All participants were then told that there was an evaluation form on the table and were asked to complete the form together. The participants were told that they would be left and given 10 minutes to complete the task, after which the experimenter would knock on the door to the room and enter to take them to finish the experiment. The experimenter then took them into the room and before leaving, told them they could take a seat at the table.

Participants were given 10 minutes, which would provide sufficient time to complete the task and enable them to interact together beyond the scope of the evaluation. After the 10 minutes the experimenter entered the room and told the participants that the evaluation task was over. The participants were then taken to separate locations to complete a questionnaire to measure the quality of the interaction with the other and their opinion of the other participant. Subsequently the participants were individually thanked, debriefed, and informed of the covert recording which took place before providing their consent for use of the video data. In the inactive Paro condition participants were finally offered the opportunity to have a short interaction with the active Paro.

4 RESULTS

In this paper we report the quantitative data from the post-interaction questionnaire.

Table 1. Multilevel model of robot condition on quality of initial interactions and liking of other. $(*)$ indicates significance ($p < 0.05$), $(+)$ indicates a trend ($p < 0.1$)

	b	SE_b	p	95% CI
Feelings during interaction				
Confidence				
Active Paro vs Inactive Paro	0.26	0.26	0.335	-0.28,0.80
Active Paro vs Pleo	0.33	0.28	0.237	-0.22,0.89
Pleo vs Inactive Paro	-0.07	0.28	0.807	-0.64,0.50
Accepted				
Active Paro vs Inactive Paro	0.17	0.15	0.248	-0.12,0.47
Active Paro vs Pleo	0.18	0.15	0.247	-0.13,0.48
Pleo vs Inactive Paro	-0.01	0.15	0.970	-0.31,0.30
Perception of interaction				
Comfortable				
Active Paro vs Inactive Paro	0.16	0.29	0.585	-0.43,0.75
Active Paro vs Pleo	0.28	0.30	0.358	-0.33,0.89
Pleo vs Inactive Paro	-0.12	0.31	0.700	-0.74,0.50
Positive				
Active Paro vs Inactive Paro	0.46	0.23	0.049 $(*)$	0.00,0.92
Active Paro vs Pleo	0.42	0.24	0.083 $(+)$	-0.06,0.89
Pleo vs Inactive Paro	0.04	0.24	0.855	-0.44,0.53
Negative				
Active Paro vs Inactive Paro	-0.01	0.16	0.965	-0.33,0.31
Active Paro vs Pleo	-0.05	0.16	0.768	-0.38,0.28
Pleo vs Inactive Paro	0.04	0.17	0.804	-0.29,0.38
Difficult				
Active Paro vs Inactive Paro	-0.43	0.31	0.175	-1.05,0.20
Active Paro vs Pleo	-0.35	0.32	0.281	-0.99,0.29
Pleo vs Inactive Paro	-0.08	0.33	0.809	-0.73,0.58
Opinion of other				
Liking				
Active Paro vs Inactive Paro	0.33	0.22	0.135	-0.11,0.77
Active Paro vs Pleo	0.32	0.22	0.165	-0.13,0.76
Pleo vs Inactive Paro	0.01	0.23	0.948	-0.44,0.47
Closeness				
Active Paro vs Inactive Paro	-0.15	0.33	0.658	-0.81,0.52
Active Paro vs Pleo	0.36	0.34	0.297	-0.32,1.04
Pleo vs Inactive Paro	-0.51	0.35	0.150	-1.20,0.19
Similarity				
Active Paro vs Inactive Paro	0.00	0.31	0.992	-0.63,0.63
Active Paro vs Pleo	0.67	0.32	0.044 $(*)$	0.02,1.31
Pleo vs Inactive Paro	-0.66	0.33	0.049 $(*)$	-1.32,-0.00
Enjoyment of interacting				
Active Paro vs Inactive Paro	0.34	0.26	0.203	-0.19,0.86
Active Paro vs Pleo	0.60	0.67	0.031 $(*)$	0.61, 0.14
Pleo vs Inactive Paro	-0.26	0.27	0.350	-0.81,0.29

Dyadic analysis was required to account for the non-independence inherent in dyadic data [6]. This is due to the hierarchical structure

of the data, with individuals nested into dyads. We used multilevel modelling in SPSS with the three robotic interaction conditions as predictors of the quality of interaction and liking of the other. The results are reported in table 1.

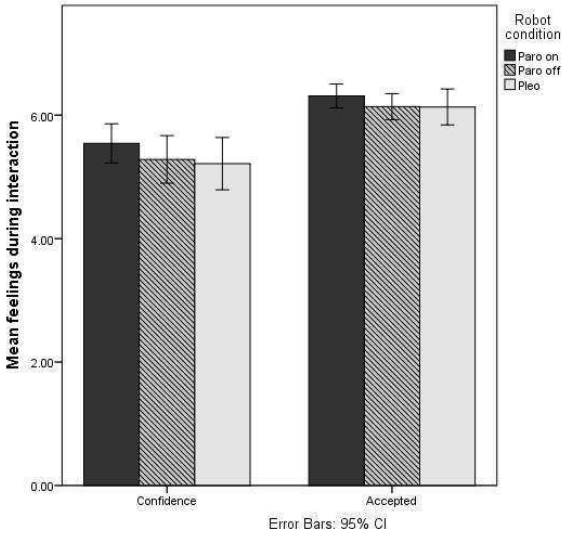


Figure 1. Feelings experienced by participants during the interaction for each robot condition

For the two factors measuring how participants felt during the interaction, no statistically significant differences between conditions were found, as seen in figure 1.

We found a significant difference between the active Paro and inactive Paro conditions for one quality of interaction factor, how positive the interaction felt. Participants in the active Paro condition had a significantly higher rating for positivity than those in the inactive Paro condition, $(b = 0.46, t(57.09) = 2.01, p = 0.049)$. In addition there was a positive trend toward significance for how positive the interaction felt for participants in the active Paro condition compared to those in the Pleo condition, $(b = 0.42, t(57.05) = 1.76, p = 0.083)$. There were no significant differences for how comfortable the interaction felt, how negative the interaction felt, and the difficulty of interaction. Figure 2 illustrates these results.

From the factors measuring participants' opinions of the other in Figure 3, perceived similarity to the other participant was significantly higher in the active Paro condition than in the Pleo condition $(b = 0.67, t(56.78) = 2.06, p = 0.044)$ but was significantly lower than the inactive Paro condition $(b = -0.66, t(56.16) = -2.01, p = 0.049)$. Participants in the active Paro condition had a significantly higher rating of enjoying interacting with the other than those in the Pleo condition, $(b = 0.60, t(56.89) = 2.21, p = 0.031)$

5 DISCUSSION

The results from this study suggest that participants found the interaction with their partner more positive and had a higher opinion of their partner when interacting together with the active Paro, than with the inactive Paro or with the Pleo. This supports the hypotheses H1a and H1b.

However no results were found to support the hypotheses H2a or H2b, that participants who interact with the Pleo would have a stronger social mediation effect than the inactive Paro.

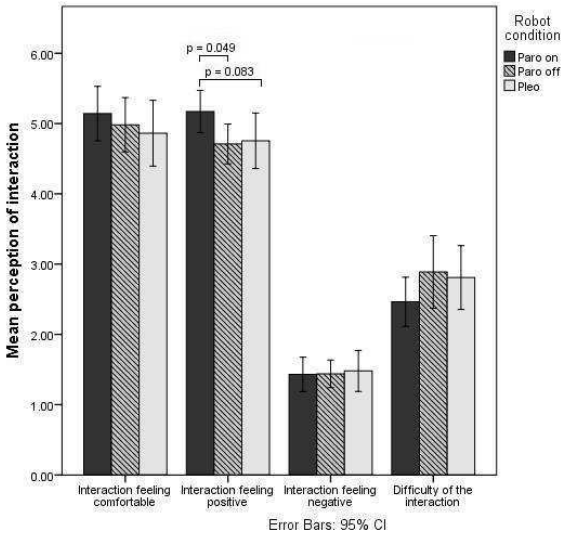


Figure 2. How the interaction was perceived for each robot condition

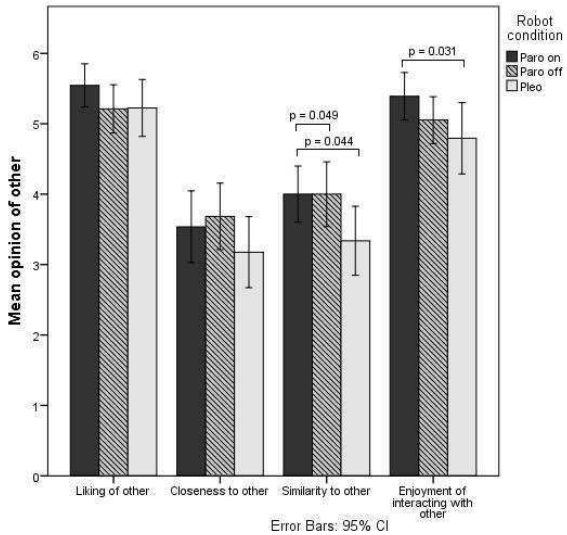


Figure 3. Participants' opinion of the other participant for each robot condition

Of the hypotheses in H1, we found a significant result to partially support hypothesis 1a which concerns the quality of the interaction. The results show that participants who interacted with the active Paro had a greater generally positive feeling about the interaction with their partner than those who interacted with the inactive Paro. The trend between the active Paro and the Pleo, while still positive, was only near significant. A possible explanation for this is that when the Paro is active and interactive it is much more stimulating for both participants than when it was inactive, and provides a stronger focus for their interaction. The interactive Pleo may have been less effective due to the different appearance and texture, which is less cuddly and tactile and therefore less engaging.

Of the four factors to measure participants' opinions of the other two factors, similarity and enjoyment of interacting with the other person, show a significant effect. The significant effect was found between the active Paro condition and the inactive Paro and Pleo conditions which supports hypothesis 1b.

It is known that perceived similarity predicts interpersonal attraction [10], and has been found to predict long term attraction and the development of relationships in newly acquainted dyads[17]. Because interacting with the Paro, when active or inactive, has a larger impact on perceived similarity within pairs in this study, they may be judged as more likely to go on to form relationships than those with the Pleo. We suggest that this is because the Pleo has a more polarising effect than Paro, in which some people dislike it whereas others find it appealing, and is more likely to divide opinions during the interaction.

The higher ratings for the enjoyment of interacting with their partner for participants in the active Paro condition show that the experience of interacting together was improved by the presence of active Paro compared to the Pleo and inactive Paro.

In accordance with our primary hypothesis, these results show that the Paro, when active, is more effective as a social mediator and an ice-breaker for first-time interactions that the Pleo or inactive Paro. The lack of significant differences between the Pleo and inactive Paro conditions show that the second hypothesis is unsupported, and there is no difference between them as social mediators. This research suggests that the interactivity and the tactile texture are important factors of Paro which make it an and engaging and appealing object for individuals to interact over for the first time.

5.1 Limitations

A number of limitations need to be acknowledged in this study: the sample size did not provide the power to verify the findings with confidence. A number of results displayed the trend we hypothesised, and it is possible that larger numbers of participants would affect the significance values of these results.

The current study has only examined the social mediation effect of Paro with female participants and these results cannot be extended to male-male or female-male dyads. The response of males participants must be investigated as due to gender role norms, it is possible that males may respond more positively towards a robot which resembles a dinosaur to one resembling a seal.

One of the questions we posed was'Can the social mediation effect of the Paro be measured under laboratory conditions?' and these results show that some effect is measurable. However, while conducting the study under laboratory conditions allows a more controlled examination of the social mediation effect, the findings cannot be generalised to all social situations, and must be replicated in different situations to understand the possible applications of this effect.

Further work could include measures of personality and attachment in order to statistically control for individual differences in forming relationships. It would also be interesting to compare this study which used unacquainted dyads to one which uses people who already know each other.

6 CONCLUSIONS AND FURTHER WORK

The present study was designed to investigate the social mediation effect of Paro under controlled conditions. This research adds to the limited evidence which shows that robotic technologies can support social interaction between people. Our results suggest that when people interact together with Paro it helps provide a context in which to form a good first impression of their partner, and have a positive experience with them.

The findings of this study demonstrate that robotic technologies can support human-human interactions by encouraging social interaction and assist in the formation of relationships. More research is needed to fully understand this potential role for the further development of robot companions.

As the quantitative data in this study comes from self-report measures in the questionnaire, we expect the observed behavioural data from the covert video recording might highlight differences between interactions in robot conditions more clearly. The next stage of this study will be to examine the content of the interactions with the video data. Further research is needed to examine the social mediation effect of the Paro with its target users; older-aged adults, including those who are healthy and those with dementia. One application of the social mediation effect of Paro which has not been evaluated to date is its use in visits to care homes from family and friends. It would be valuable to investigate the role of Paro during these visits, and whether it leads to an increase in quality of the visitation time.

REFERENCES

[1] PLEO, 2010.

[2] R. Bemelmans, G. J. Gelderblom, P. Jonker, and L. De Witte. Socially assistive robots in elderly care: a systematic review into effects and effectiveness. *Journal of the American Medical Directors Association*, 13(2):114 – 120.e1, Feb. 2012.

[3] D. S. Berry and J. S. Hansen. Personality, Nonverbal Behavior, and Interaction Quality in Female Dyads. *Personality and Social Psychology Bulletin*, 26(3):278–292, Mar. 2000.

[4] L. Giusti and P. Marti. Robots as social mediators: a study in the wild. *Gerontechnology*, 7(2), Apr. 2008.

[5] S. J. Hunt, L. A. Hart, and R. Gomulkiewicz. Role of Small Animals in Social Interactions between Strangers. *The Journal of Social Psychology*, 132(2):245–256, Apr. 1992.

[6] D. A. Kenny, D. A. Kashy, and W. L. Cook. *Dyadic Data Analysis*. 2006.

[7] C. D. Kidd, W. Taggart, and S. Turkle. A sociable robot to encourage social interaction among the elderly. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, number May, pages 3972–3976. IEEE, 2006.

[8] M. R. Leary, R. M. Kowalski, and D. J. Bergen. Interpersonal Information Acquisition and Confidence in First Encounters. *Personality and Social Psychology Bulletin*, 14(1):68–77, Mar. 1988.

[9] J. McNicholas and G. M. Collis. Dogs as catalysts for social interactions: Robustness of the effect. *British Journal of Psychology*, 91(1):61–70, Feb. 2000.

[10] R. M. Montoya, R. S. Horton, and J. Kirchner. *Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity*, volume 25. Dec. 2008.

[11] E. Mordoch, A. Osterreicher, L. Guse, K. Roger, and G. Thompson. Use of social commitment robots in the care of elderly people with dementia: A literature review. *Maturitas*, 74(1):14–20, 2013.

[12] H. Robinson, B. Macdonald, N. Kerse, and E. Broadbent. The psychosocial effects of a companion robot: a randomized controlled trial. *Journal of the American Medical Directors Association*, 14(9):661–7, Sept. 2013.

[13] A. Sharkey and N. Sharkey. Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40, July 2010.

[14] T. Shibata, K. Wada, Y. Ikeda, and S. Sabanovic. Cross-Cultural Studies on Subjective Evaluation of a Seal Robot. *Advanced Robotics*, 23(4):443–458, Jan. 2009.

[15] R. Sparrow and L. Sparrow. In the hands of machines? The future of aged care. *Minds and Machines*, 16(2):141–161, Aug. 2006.

[16] S. Sprecher, S. Treger, J. D. Wondra, N. Hilaire, and K. Wallpe. Taking turns: Reciprocal self-disclosure promotes liking in initial interactions. *Journal of Experimental Social Psychology*, 49(5):860–866, Sept. 2013.

[17] M. Sunnafrank and A. Ramirez. At First Sight: Persistent Relational Effects of Get-Acquainted Conversations. *Journal of Social and Personal Relationships*, 21(3):361–379, June 2004.

[18] S. Šabanović, C. C. Bennett, W.-L. Chang, and L. Huber. PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. *IEEE ... International Conference on Rehabilitation Robotics : [proceedings]*, 2013:1–6, June 2013.

[19] K. Wada and T. Shibata. Robot therapy in a care house - Change of relationship among the residents and seal robot during a 2-month long study. In *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, volume 23 of *16th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN*, pages 1250–1255, IEEE, Japan, Oct. 2007.

[20] K. Wada, T. Shibata, T. Musha, and S. Kimura. Robot therapy for elders affected by dementia. *IEEE Engineering in Medicine and Biology Magazine*, 27(4):53–60, July 2008.

[21] K. Wada, T. Shibata, T. Saito, and K. Tanie. Psychological and social effects of robot assisted activity to elderly people who stay at a health service facility for the aged. In *Proceedings - IEEE International Conference on Robotics and Automation*, volume 3 of *2003 IEEE International Conference on Robotics and Automation*, pages 3996–4001, Intelligent Systems Institute, AIST, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan, 2003.

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Can Less be More? The Impact of Robot Social Behaviour on Human Learning

James Kennedy[1] and **Paul Baxter**[1] and **Tony Belpaeme**[1]

Abstract. In a large number of human-robot interaction (HRI) studies, the aim is often to improve the social behaviour of a robot in order to provide a better interaction experience. Increasingly, companion robots are not being used merely as interaction partners, but to also help achieve a goal. One such goal is education, which encompasses many other factors such as behaviour change and motivation. In this paper we question whether robot social behaviour helps or hinders in this context, and challenge an often underlying assumption that robot social behaviour and task outcomes are only positively related. Drawing on both human-human interaction and human-robot interaction studies we hypothesise a curvilinear relationship between social robot behaviour and human task performance in the short-term, highlighting a possible trade-off between social cues and learning. However, we posit that this relationship is likely to change over time, with longer interaction periods favouring more social robots.

1 INTRODUCTION

Social human-robot interaction (HRI) commonly focuses on the experience and perception of human users when interacting with robots, for example [2]. The aim is often to improve the quality of the social interaction which takes place between humans and robots. Companion robots increasingly aim not just to merely interact with humans, but to also achieve some goal. These goals can include, for example, imparting knowledge [11], eliciting behaviour change [17] or collaborating on a task [3, 13]. Studies with these goal-oriented aims often still apply the same principles for social behaviour as those without goals - that of maximising human interaction and positive perception towards the robot. The implicit assumption is often that if the interaction is improved, or the human perception of the robot is improved, then the chance of goal attainment will be increased as well.

In this paper, we focus on learning. In this context, we take learning to be the acquisition and retention of novel information, and its reuse in a new situation. This definition covers 3 areas from each of the 'Cognitive Process' (remember, understand, apply) and 'Knowledge' (factual, conceptual, procedural) dimensions of learning according to the revised version of Bloom's taxonomy [14]. Learning outcomes can depend on many different elements of behaviour, such as motivation [20] and engagement [4], which will also be considered here.

The remainder of this paper is structured as follows. First, studies in which social robots assist humans in learning will be reviewed, with the intention of showing the complex variety of results obtained when relating learning to the social behaviour of the robot (Section 2). Human-human interactions are then considered and are used as a basis to create a hypothesis about the relationship of robot social behaviour and human performance in tasks over both the long and short-term (Section 3). This leads to a discussion of the implications for HRI design in such contexts (Section 4).

[1] Centre for Robotics and Neural Systems, Cognition Institute, Plymouth University, United Kingdom
email: {james.kennedy, paul.baxter, tony.belpaeme}@plymouth.ac.uk

2 MIXED LEARNING RESULTS IN HRI

One area of great potential in HRI is in using robots for education. However, mixed results are often found when using social robots to teach or tutor humans. Despite regular reports of liking robots more than virtual avatars, or preferring more socially contingent robots over those with less social capability, the human performance in learning tasks doesn't always reflect these positive perceptions [11, 12, 17, 22]. Conversely, significant cognitive gains have been found when comparing robots to virtual avatars, with varied amounts of contingent behaviour [15, 16]. Similar effects have been seen in compliance when comparing agents of differing embodiments [1]. Whilst the varied context and content to be learned between these studies could account for many of the differences in results, we suggest that the relationship between social behaviour and learning performance may be more complex than typically assumed.

Commonly, when behavioural manipulations are carried out on one or two cues, such as in a study by Szafir *et al.* varying the gestures and vocal volume that a robot uses, there are clear benefits to the human in terms of performance in learning tasks [26]. However, these positive benefits may be lost, or even reversed when larger manipulations to the social behaviour of the robot are applied, as in [12]. While it may be reasonably assumed that the effect of multiple individual cues is additive, this does not seem to be in accordance with the empirical evidence. Indeed, the proposition that social cues are perceived by humans as a single percept [29] considers individual social cues as providing the context for the interpretation of other social cues (recursively), leading to non-trivial interactions and consequences when multiple social cues are applied. There is thus the possibility that making large manipulations in social behaviour by varying multiple social cues simultaneously does not elicit the benefits that varying each of these cues individually would, as suggested by the data.

Human expectations of sociality will play a large role in an interaction with a robot. It has been suggested that a discrepancy between categorical expectations and perceptual stimuli could account for negative cognitive reactions [19]. We posit that humans don't necessarily expect to interact with a robot exhibiting social behaviours and that the discrepancy between their expectation and the reality of the interaction could create a cognitive reaction which impedes learning. This might explain some results showing a lack of improvement when social presence of an agent is increased (such as when going from a virtual avatar to a robot, as in [10, 17]), or when social behaviour

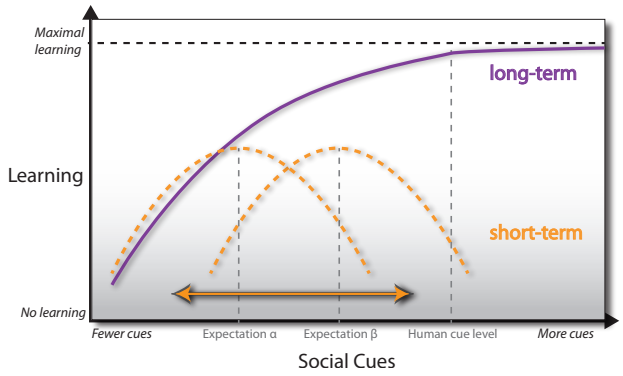


Figure 1: Hypothesised relationship between social behaviour (characterised by *immediacy* for example) as exhibited by a robot and its impact on the learning of a human in both the short and long-term. The position of the short-term curve is dependent on the humans' prior expectations of social behaviour (e.g. α is the expectation of fewer social cues from the robot than expectation β). Over time, these expectations normalise with reality, with increased use of social cues tending to lead to improved learning performance for the human interactant.

becomes more contingent, as in [12]. Expectation discrepancy would consequently lead to changes in the cognitive reaction over time as expectations change, and vary based on individuals, contexts, and so on; this is reflected in Figure 1 and will be expanded upon in Section 3.

Although there are many questions regarding learning in the context of HRI that remain unexplored, it would be useful to try and first create a testable hypothesis to attempt to explain why the results gathered so far are so varied. Whether this lies in social presence differences between virtual and physical robots, or in social behaviour manipulation between robot conditions, the main variable in all of the studies considered in this section is sociality. As such, we now consider how social behaviour might influence learning.

3 SOCIAL BEHAVIOUR AND LEARNING

In order to understand more about the nature of the relationship between social behaviour and learning, literature from human-human interaction (HHI) studies will now be introduced. Learning in the context of HHI has been under study for far longer than HRI, so longer-term research programmes have been carried out, and more data is consequently available.

When exploring the connection between learning and social behaviour in HHI literature, one behavioural measure repeatedly found to correlate with learning is 'immediacy'. Particularly applied to educational contexts, this concept has been long-established and validated across many cultures [18, 24] and age ranges [21]. Immediacy provides a single value definition of the social behaviour of a human in an interaction by characterising conduct in a range of verbal and non-verbal behavioural dimensions [23]. Immediacy could therefore prove a useful means of characterising robot social behaviour in HRI (as in [26]). Further, it has been shown that more immediate behaviours on the part of a human tutor increases cognitive learning gains [28]. However, the exact nature of the relationship between immediacy and cognitive learning gain is debated [5, 28].

Many HRI studies seem to implicitly assume a linear relationship between an increase in the number of social cues used or in social behaviour contingency and learning gains (or gains in related measures such as engagement, compliance, etc). Upon reviewing the literature concerning immediacy between humans, this has sometimes found to be the case [5], but more recent work has shown that this relationship may in fact be curvilinear [6]. A curvilinear relationship could go some way to explaining the mixed results found so far in HRI studies considering task performance with respect to robot social behaviour; it is possible that some studies make the behaviour *too social* and fall into an area of negative returns.

It is hypothesised that the curvilinear nature of immediacy may have been the effect observed in the study by Kennedy *et al.* in which a 'social' robot led to less learning than a robot which was actively breaking social expectations [12]. Over the short term, the novelty of social behaviour displayed by a robot may cause this kind of curvilinear relationship as has been observed in relation to immediacy [6]. As alluded to in Section 2, humans have a set of expectations for the sociality of the robot in an interaction. We would suggest that the greater the discrepancy between these expectations and the actual robot behaviour, the more detrimental the effect on learning. Individuals will have varied expectations, which is manifested in different short-term curves (Figure 1): the short-term curve shifts such that its apex (translating to the greatest possible amount of learning in the time-frame) is at the point where the expected and actual level of social cues is most closely matched. Prior interactions and the range of expectations created could also change the shape of the short-term curve, making the apex flatter or more pronounced depending on the variety of previous experiences.

However, when considering the interaction over the longer-term, such novelty effects wear off as the human adapts to the robot and their expectations change [7, 8, 25]. In this case we suggest that substantial learning gains could be made as the robot behaviour approaches a 'human' level of social cues; having attained a reasonable matching of expectation to reality, the robot can leverage the advantages that social behaviour confers in interactions, as previously suggested [9, 26]. Beyond this level, improvement would still be found by adding more cues, but the rate of increase is much smaller as the cues will require more conscious effort to learn and interpret. These concepts are visualised in the long-term curve seen in Figure 1.

4 PERSPECTIVES

So far, we have challenged the assumption that social behaviour has a simple linear relationship with learning by providing conflicting examples from HRI literature and also by tying concepts of social behaviour to the measure of immediacy from HHI literature. Given the regular use of HHI behaviour in generating HRI hypotheses, the non-linear relationship between immediacy and learning is used to hypothesise a non-linear relationship for HRI, particularly in the short-term (Figure 1).

A series of controlled studies would be needed to verify whether these hypothesised curves are correct. One particular challenge with this is the measuring of social behaviour. It is unclear what it is to be 'more' or 'less' social, and how this should be measured. This is where we propose that *immediacy* could be used as a reasonable approximation. All factors in immediacy are judgements of different aspects of social behaviour, which are combined to provide a single number representing the overall 'immediacy' (i.e. sociality of social behaviour) of the interactant. This makes the testing of such a hypothesis possible as the social behaviour then becomes a single dimension for consideration.

Of course, there are many other issues (such as robotic platform and age of human) which would need to be explored in this context,

but with a single measure approximating sociality this would at least be possible. Providing an immediacy measure for robot behaviour makes it much easier to compare results between studies, allowing improved analysis of the impact of things such as task content and context, which are currently very difficult to disentangle when comparing results between studies. Literature from the field of Intelligent Tutoring Systems may be a useful starting point for future work to investigate specific aspects of learning activities due to their proven effectiveness across many contexts [27].

It should be noted that the aim of this paper is to highlight the potential directionality of the relationships involved between social cues and learning. There is not enough data available to represent the shape of the curves presented in Figure 1 with any great accuracy. The curves have been devised based on the few data points available from the literature, and following from concepts of immediacy and discrepancies of expectation, as explored in Sections 2 and 3.

5 CONCLUSION

We suggest that immediacy could be taken from the HHI literature to be validated and applied to HRI more extensively as it presents itself as an ideal means to facilitate comparison of highly varied social behaviour between studies. The large volume of immediacy literature in relation to learning and other contexts could also provide a firm theoretical basis for the generation and testing of hypotheses for HRI.

In this position paper we have shown through examples from HHI and HRI literature that the relationship between social behaviour and task outcome, specifically learning in the present work, for humans cannot be assumed to be linear. We hypothesise a model in which social behaviour not only has a non-linear relationship with learning, but also a relationship which changes over interaction time. Following the hypothesised model, we suggest that although in the short-term there may be some disadvantages for a robot to be maximally socially contingent, the benefits conferred by social behaviour as proposed by prior work will be seen in the long-term.

ACKNOWLEDGEMENTS

This work was partially funded by the School of Computing and Mathematics, Plymouth University, U.K., and the EU FP7 DREAM (FP7-ICT-611391) and ALIZ-E (FP7-ICT-248116) projects.

REFERENCES

[1] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati, 'The effect of presence on human-robot interaction', in *Proc. RO-MAN'08*, pp. 701–706. IEEE, (2008).

[2] Mike Blow, Kerstin Dautenhahn, Andrew Appleby, Chrystopher L Nehaniv, and David C Lee, 'Perception of robot smiles and dimensions for human-robot interaction design', in *Proc. RO-MAN'06*, pp. 469–474. IEEE, (2006).

[3] Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung, 'Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment', *Journal of Human-Robot Interaction*, **2**(1), 82–111, (2013).

[4] Robert M Carini, George D Kuh, and Stephen P Klein, 'Student engagement and student learning: Testing the linkages*', *Research in Higher Education*, **47**(1), 1–32, (2006).

[5] Laura J Christensen and Kent E Menzel, 'The linear relationship between student reports of teacher immediacy behaviors and perceptions of state motivation, and of cognitive, affective, and behavioral learning', *Communication Education*, **47**(1), 82–90, (1998).

[6] Jamie Comstock, Elisa Rowell, and John Waite Bowers, 'Food for thought: Teacher nonverbal immediacy, student learning, and curvilinearity', *Communication Education*, **44**(3), 251–266, (1995).

[7] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C Schultz, et al., 'Designing robots for long-term social interaction', in *Proc. IROS'05*, pp. 1338–1343. IEEE, (2005).

[8] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro, 'Interactive robots as social partners and peer tutors for children: A field trial', *Human-Computer Interaction*, **19**(1), 61–84, (2004).

[9] A. Kendon, 'Some functions of gaze-direction in social interaction.', *Acta Psychol (Amst)*, **26**(1), 22–63, (1967).

[10] James Kennedy, Paul Baxter, and Tony Belpaeme, 'Children comply with a robot's indirect requests', in *Proc. HRI'14*, (2014).

[11] James Kennedy, Paul Baxter, and Tony Belpaeme, 'Comparing robot embodiments in a guided discovery learning interaction with children', *International Journal of Social Robotics*, (2014).

[12] James Kennedy, Paul Baxter, and Tony Belpaeme, 'The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning', in *Proc. HRI'15*, (2015).

[13] Hatice Kose-Bagci, Ester Ferrari, Kerstin Dautenhahn, Dag Sverre Syrdal, and Chrystopher L Nehaniv, 'Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot', *Advanced Robotics*, **23**(14), 1951–1996, (2009).

[14] David R Krathwohl, 'A revision of Bloom's taxonomy: An overview', *Theory into practice*, **41**(4), 212–218, (2002).

[15] Daniel Leyzberg, Sam Spaulding, and Brian Scassellati, 'Personalizing robot tutors to individual learning differences', in *Proc. HRI'14*, pp. 423–430, (2014).

[16] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati, 'The physical presence of a robot tutor increases cognitive learning gains', in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, CogSci 2012, pp. 1882–1887, (2012).

[17] Rosemarijn Looije, Anna van der Zalm, Mark A Neerincx, and Robbert-Jan Beun, 'Help, I need some body: The effect of embodiment on playful learning', in *Proc. RO-MAN'12*, pp. 718–724. IEEE, (2012).

[18] James C McCroskey, Aino Sallinen, Joan M Fayer, Virginia P Richmond, and Robert A Barraclough, 'Nonverbal immediacy and cognitive learning: A cross-cultural investigation', *Communication Education*, **45**(3), 200–211, (1996).

[19] Roger K Moore, 'A Bayesian explanation of the 'uncanny valley' effect and related psychological phenomena', *Nature Scientific Reports*, **2**(864), (2012).

[20] Paul R Pintrich and Elisabeth van de Groot, 'Motivational and self-regulated learning components of classroom academic performance.', *Journal of educational psychology*, **82**(1), 33, (1990).

[21] Timothy G Plax, Patricia Kearney, James C McCroskey, and Virginia P Richmond, 'Power in the classroom VI: Verbal control strategies, nonverbal immediacy and affective learning', *Communication Education*, **35**(1), 43–55, (1986).

[22] Aaron Powers, Sara Kiesler, Susan Fussell, and Cristen Torrey, 'Comparing a computer agent with a humanoid robot', in *Proc. HRI'07*, pp. 145–152. IEEE, (2007).

[23] Virginia P Richmond, James C McCroskey, and Aaron D Johnson, 'Development of the Nonverbal Immediacy Scale (NIS): Measures of self- and other-perceived nonverbal immediacy', *Communication Quarterly*, **51**(4), 504–517, (2003).

[24] Vincent Santilli and Ann Neville Miller, 'The effects of gender and power distance on nonverbal immediacy in symmetrical and asymmetrical power conditions: A cross-cultural study of classrooms and friendships', *Journal of International and Intercultural Communication*, **4**(1), 3–22, (2011).

[25] Ja-Young Sung, Henrik I Christensen, and Rebecca E Grinter, 'Robots in the Wild: Understanding long-term use', in *Proc. HRI'09*, pp. 45–52. IEEE, (2009).

[26] Daniel Szafir and Bilge Mutlu, 'Pay attention!: Designing adaptive agents that monitor and improve user engagement', in *Proc. CHI'12*, pp. 11–20, New York, NY, USA, (2012). ACM.

[27] Kurt VanLehn, 'The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems', *Educational Psychologist*, **46**(4), 197–221, (2011).

[28] Paul L Witt, Lawrence R Wheeless, and Mike Allen, 'A meta-analytical review of the relationship between teacher immediacy and student learning', *Communication Monographs*, **71**(2), 184–207, (2004).

[29] Jamil Zaki, 'Cue integration a common framework for social cognition and physical perception', *Perspectives on Psychological Science*, **8**(3), 296–312, (2013).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Robots Guiding Small Groups: The Effect of Appearance Change on the User Experience

Michiel Joosse, Robin Knuppe, Geert Pingen, Rutger Varkevisser, Josip Vukoja, Manja Lohse and Vanessa Evers [1]

Abstract. In this paper we present an exploratory user study in which a robot guided small groups of two to three people. We manipulated the appearance of the robot in terms of the position of a tablet providing information (facing the group that was guided or the walking direction) and the type of information displayed (eyes or route information). Our results indicate that users preferred eyes on a display that faced the walking direction and route information on a display that faced them. The study gave us strong indication to believe that people are not in favor of eyes looking at them during the guiding.

1 Introduction

Social robots are designed to interact with humans in human environments in a socially meaningful way [3]. As a logical consequence, the design of robots often includes human-like features, e.g., heads or arms in order to generate social responses. It has been found that by using such anthropomorphic cues, people automatically have expectations of the robot's behavior [4].

However, the capabilities of robots differ from those of humans which allows them to use the anthropomorphic cues in different ways. For example, robot eyes can face the user while walking because the robot has other means (e.g., laser range finders) to detect the way to go. Thus, robots can walk backward. As eye contact has been shown to impact our image of others, and whether positive or negative, this being a sign of potential social interaction [6], robots facing users while guiding might actually be beneficial. On the other hand, literature indicates that people use a combination of head and eye movement to non-verbally indicate their direction [1] and users might expect robots to do the same.

Robots can also use non-anthropomorphic cues in different ways than humans, e.g. in the guiding context they can display route information rather than eyes. Related work found that visitors in historic places prefer a guide, as they would not have to worry about the route, or carry a map [2]. Therefore this could be beneficial for robots as well.

In the FP7-project SPENCER[2] we aim at developing a guide robot for a public place (airport) which will have a head and a screen. In this context, the questions arise which direction the head and screen should face when guiding a small group and what content should be displayed on the screen.

In related work, Shiomi et al. [5] conducted an experiment with the Robovie robot that drove either forward or backward while guiding participants in a mall (over a short distance). The overall finding in this experiment was that more bystanders joined when the robot moved backwards compared with frontwards, and that more people were inclined to follow the robot the entire time when moving backwards. In our work we are not so much interested in attracting people, but more in guiding people over a longer distance. Thus the question we pose here is how these design decisions impact the user experience in the process of guiding.

In this paper we present an exploratory study, in which we asked participants to follow a guide robot through a public lab space. This robot was equipped with a tablet (facing forwards, or facing the user) providing information to the participants. We were specifically interested in finding out which combination of tablet direction and type of information provided (eyes or route information) would yield the most positive user experience.

2 Method

In order to answer our research question, we designed an exploratory user study in which small groups of two to three participants were given a short guided tour by a robot.

2.1 Robot platform

For this study we attached a shell on top of a remote-controlled Robotino robot platform[3]. The height of the robot was 170cm and it drove at a speed of approximately 0.7 m/s. For purposes of this exploratory study, it was not deemed necessary to have the robot drive the path autonomously. Furthermore, the location of obstacles in the DesignLab changed from time to time (e.g. couches, chairs). As we were primarily interested in user experience ratings, the robot was remotely operated by an experimenter. Participants were not made aware of this before participating in the experiment.

2.2 Manipulations

We manipulated the direction of the tablet mounted on top of the robot and the information displayed on the tablet (Figure 1 and Table 1). In conditions A (Figure 1a) and B (Figure 1c) a set of blinking eyes was displayed on the tablet either facing the participants or the walking direction. In condition C we programmed the tablet to display route information, i.e., the remaining distance to the target (Figure 1e). A condition having the tablet mounted on the front of the robot, while displaying route information was deemed unnecessary as this would neither provide information for the participants following the robot, nor for other people present in the laboratory.

[1] Human Media Interaction group, University of Twente, the Netherlands, email: {r.a.knuppe, g.l.j.pingen, r.a.varkevisser, j.vukoja}@student.utwente.nl, {m.p.joosse, m.lohse, v.evers}@utwente.nl
[2] http://www.spencer.eu
[3] http://www.festo-didactic.com/int-en/learning-systems/education-and-research-robots-robotino/

(a) Condition A front (b) Condition A back (c) Condition B front (d) Condition B back (e) Condition C front (f) Condition C back

Figure 1: The appearance of the robot in the three conditions, showing the front and back side of the robot

Table 1: Overview of study conditions and number of participants

Condition	A	B	C
Tablet direction	Front	Back	Back
Tablet display	Eyes	Eyes	Time to destination
N	9	8	8
Group distribution	3x 3-person	1x 2-person	1x 2-person
		2x 3-person	2x 3-person

2.3 Measures

In the post-experiment questionnaire user experience was assessed using a variety of measures.

All questions (except demographic- and open questions) were formulated as 5-point Likert-scaled items. General experience was assessed with eleven questions measuring among others if participants trusted that the robot knew where it was going, if it was clear where the robot was going and whether or not the robot was helpful in guiding someone. In this set of questions also the speed of the robot and volume of the audio messages were evaluated.

Five questions related to the physical appearance assessed the design, and specifically the height of the robot. Usability questions included questions related to users' expectancies of system capabilities and whether or not they were satisfied with the overall performance of the robot. Depending on the condition, this section included 5 (condition A), 6 (condition B), or 7 (condition C) questions.

Eight questions were included related to demographic information (age, gender, educational background) and familiarity with robots, social robots, and the premises where the test was conducted. A control question about the position of the tablet was included, and finally, we were interested in knowing whether or not the instructions provided were clear. Overall, this resulted in 30-32 questions

2.4 Procedure

Small groups of participants were recruited to participate in a guided tour of the DesignLab, a recently-opened lab of the University of Twente. Participants were given a briefing, after which they were given a tour of about five minutes through the lab. Participants were requested to follow the robot. No specific instructions were provided regarding the distance they should keep to the robot (Figure 4). The tour went past two points of interest (Figure 2, point B and C) where the robot provided a brief statement about the purpose using a text-to-speech engine. For example, when arriving at waypoint A, participants would see a tray with kinetic sand, and the robot would state that "The kinetic sand is made up of 98 percent sand, and 2 percent polyminethyl siloxane which gives it its elastic properties."

Afterwards the robot returned to the starting position where participants were requested to fill out the post-experiment questionnaire (Figure 2 point A). Following debriefing, participants were provided some candy as reward for their participation.

2.5 Participants

A total of 25 participants (14 males, 11 females) participated in the user study, with ages ranging from 17 to 40 (M=23.76, sd=5.93). All participants were students and staff from the University of Twente, primarily of Dutch (68%), German (8%) and Greek (8%) nationality. Participants had average experience with robots in general (M=2.84, sd=.90) and little experience with social robots (M=2.12, sd=1.09).

2.6 Data analysis

We calalculated means for all items. To compare between conditions, the data were first tested for normality. In case of normally distributed data, we report ANOVA's and T-tests in the results section, otherwise Kruskal-Wallis and post-hoc Mann-Whitney tests are reported.

3 Results

Overall, participants indicated they were quite satisfied with the robot: they believed the robot was helpful (M=4.47, sd=0.78), it

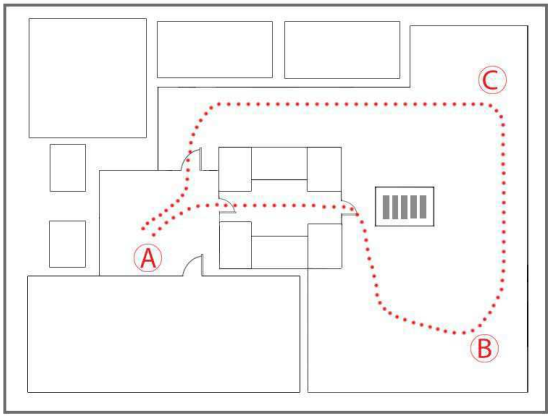


Figure 2: Layout of the laboratory showing start/end position (A) and two points of interest (B and C)

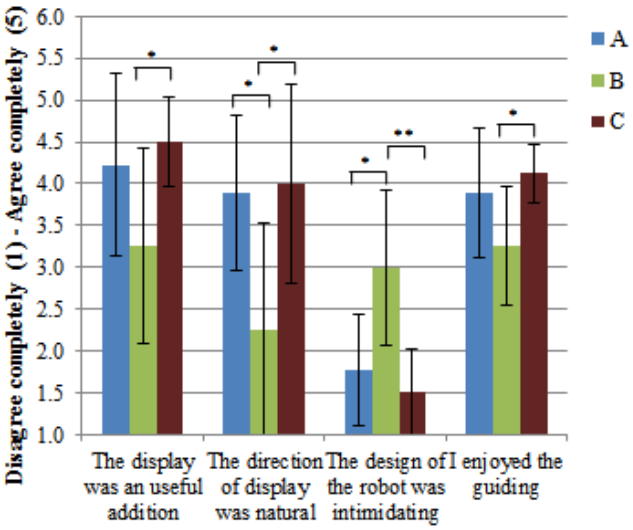


Figure 3: User experience ratings in the conditions; * indicates significance at the 0.05 level, ** at the 0.01 level

moved at a comfortable speed (M=3.12, sd=1.37), and participants trusted that the robot knew where it was going to (M=4.47, sd=0.78). These ratings did not differ significantly between conditions. Participants were moderately positive about the usability of the system: they felt comfortable using it (M=3.67, sd=1.05) and were satisfied by its performance (M=3.56, sd=0.77). No main effects or correlations were found including gender, age, robot experience and/or educational background.

Between conditions, Kruskal-Wallis tests indicated there were significant differences which were mostly due to the location of the tablet, thus between conditions A and C, versus condition B where the tablet was mounted on the front of the robot.

Post-hoc Mann-Whitney's indicated participants felt the direction of the screen was more appropriate in condition A (M=3.89, sd=.928) compared with B (M=2.25, sd=1.28), U=11.5; Z=-2.459, $p<0.05$. A similar effect was found between conditions B and C (M=4.0, sd=1.20), U=10.0, Z=-2.36, $p<0.05$ Furthermore, the design in condition B was more intimidating (M=3.00, sd=.97) compared with condition A (M=1.78, sd=.68), U=11.5, Z=-2.51, $p<0.05$ and condition C (M=1.50, sd=.54), U=6.00, Z=-2.885, $p<0.01$. Participants in condition C enjoyed the guiding more (M=4.13, sd=.35) compared with those in condition B (M=3.25, sd=.71), U=10.5, Z=-2.62, $p<0.05$.

With respect to the robot's appearance, participants felt that the body design matches the robot's function (M=2.71, sd=0.94). One of the interesting findings was that participants indicated the height was appropriate (M=4.21, sd=0.82). Informal sessions with participants indicated the robot would be too tall for a guiding robot, but in the end this was not the case. One of the reasons for this could be that participants' own average height was 177cm (sd=8.5cm), thus, most of them being taller than the robot.

4 Discussion & Conclusion

In this paper we presented an exploratory study into the effect of a robot's physical appearance on usability and user experience. Small groups of people were provided a short tour by a guide robot. Our results indicate that the location of the screen can be either forward

Figure 4: A small group of participants being guided by the robot

or backward, depending on the information displayed. In the case of eyes facing participants, our results showed that this was considered to be very unnatural and intimidating. On the other hand, when the tablet faced participants and route information was provided this was again evaluated as more useful. This might seem to be in contrast with the results of Shiomi et al. [5] who found that eyes facing participants are more effective to attract bystanders. However, we think this could be explained because in our setup the participants had already been introduced to the robot and asked to follow it.

Neither gender, age or experience with robots influenced the evaluation of the robots significantly, which could be due to small sample size.

Our future work will include a more interactive setup (e.g. provide participants some choices) during the tour. A second area of interest would be robot speed, and to investigate whether or not the speed of a guiding robot could be slower when guiding small groups compared with individual people. To conclude: the appearance of a guide robot can greatly influence user experience, something subtle as two eyes facing participants significantly decreases a robot's evaluation. Hence, more research is needed to even better understand how to design acceptable guide robots.

ACKNOWLEDGEMENTS

This research has been partly supported by the European Commission under contract number FP7-ICT-600877 (SPENCER).

REFERENCES

[1] Mark A Hollands, AE Patla, and JN Vickers, 'look where youre going!: gaze behaviour associated with maintaining and changing the direction of locomotion', *Experimental Brain Research*, **143**(2), 221–230, (2002).

[2] Daphne E Karreman, Elisabeth MAG van Dijk, and Vanessa Evers, 'Using the visitor experiences for mapping the possibilities of implementing a robotic guide in outdoor sites', in *RO-MAN, 2012 IEEE*, pp. 1059–1065. IEEE, (2012).

[3] Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan, 'Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction', *Journal of communication*, **56**(4), 754–772, (2006).

[4] Manja Lohse, 'The role of expectations and situations in human-robot interaction', *New Frontiers in Human-Robot Interaction*, 35–56, (2011).

[5] Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita, 'A larger audience, please!: encouraging people to listen to a guide robot', in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pp. 31–38. IEEE Press, (2010).

[6] Michael von Grünau and Christina Anston, 'The detection of gaze direction: A stare-in-the-crowd effect', *Perception*, **24**(11), 1297–1313, (1995).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

How can a tour guide robot's orientation influence visitors' orientation and formations?

Daphne E. Karreman[1], Geke D.S. Ludden[2], Elisabeth M.A.G. van Dijk[1], Vanessa Evers[1]

Abstract. In this paper, we describe a field study with a tour guide robot that guided visitors through a historical site. Our focus was to determine how a robot's orientation behaviour influenced visitors' orientation and the formations groups of visitors formed around the robot. During the study a remote-controlled robot gave short guided tours and explained some points of interest in the hall of Festivities in the Royal Alcázar in Seville (Spain). To get insight into visitors' reactions to the robot's non-verbal orientation behaviour, two orientations of the robot were tested; either the robot was oriented with its front towards the visitors, or the robot was oriented with its front towards the point of interest. From the study we learned that people reacted strongly to the orientation of the robot. We found that visitors tended to follow the robot tour guide from a greater distance (more than 3 meters away from the robot) more frequently when the robot was oriented towards the visitors than when it was oriented towards the point of interest. Further, when the robot was oriented towards the point of interest, people knew where to look and walked towards the robot more often. On the other hand, people also lost interest in the robot more often when it was oriented towards the point of interest. The analysis of visitors' orientation and formations led to design guidelines for effective robot guide behaviour.

1 INTRODUCTION

Several robots have been developed to give guided tours in a museum-like setting (some examples are described in [1]–[4]). These previously developed robotic tour guides did good jobs in their navigation and localization tasks, such as avoiding collisions with visitors or objects, and showing they were aware of the visitors' presence. While giving the tours, these robots captured the attention of visitors, had interactions with visitors and guided the visitors through smaller or larger parts of exhibitions. Studies reported some information about the visitors' reactions to the robot's actions which has led to knowledge on specific reactions of people to the modalities of these robots and behaviour shown by these robot designs.

Within the EU FP7 FROG project we were, among other innovations and application areas, interested in effective tour guide behaviour and personality for a robot guide. To find effective behaviours we started to examine the effect of single modalities on robot behaviour and visitor reactions to those behaviours. The question we wanted to answer with this study is: how does the robot orientation behaviour influence the orientations of the visitors, as well as the type of formations that (groups of) visitors form around the robot? The findings of the study we present in this paper led to guidelines to design behaviours (for FROG and other robots) that will influence visitors' reactions, such as orientation and group formations.

One way of creating robot behaviour is to copy human behaviour to a robot. A limitation of copying human tour guide behaviour to robots is that robots in general, and the FROG robot specifically, do not have the same modalities to perform actions that human tour guides perform. On the other hand, robots might have modalities to perform actions that human tour guides cannot perform. Therefore, we need to carefully study how and which robot modalities can effectively be used in interaction.

In previous studies, the reactions of the visitors were assumed to be similar to visitor reactions to human tour guides, but it turned out that these were different. For example, people often crowded around the robots [1], [2], [4], [5] or started to search for its boundaries by blocking the path [1] or pushing the emergency button [2], [6]. On the other hand, people often used their known human-human interaction rules to interact with the robots [2], even if the robots were not humanoid and people were informed that not all cues could be understood by the robot. Similar to robots that have been used in other studies, our FROG robot is not humanoid. We know that human tour guides influence visitor reactions of a group of visitors by using gaze behaviour and orientation [7]. Therefore, we are interested in visitors' reactions to a basic tour guide robot with limited interaction modalities. Also, we wanted to find out whether these reactions are similar to or different from visitor reactions to a human tour guide.

In this paper we will focus on the formation and orientation of visitors as a reaction to the robot orientation behaviour. We use the term formation to indicate the group structure, distance and orientation of the visitors who showed interest in the robot and/or the point of interest the robot described. In human guided tours, people generally stand in a common-focus gathering, a formation in which people give each other space to focus on the same point of interest, often a semi-circle [8]. For robot guided tours, we expected to find similar formations. However, from previous research we learned that single persons or pairs of visitors also joined the tour [9], [2]. Therefore, we considered the combination of distance and orientation of these individuals or pairs as formations as well. We assumed that people would be engaged with the robot or the explanation, when they were oriented towards the robot or the point of interest for a longer period of time. Hence, we also use the terms formation, orientation and engagement separately from each other in order to be specific in the description of the results.

[1] Human Media Interaction, Faculty of Electrical Engineering, Mathematics and Computing Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: {**d.e.karreman**, **e.m.a.g.vandijk, v.evers**}**@utwente.nl**

[2] Product Design, Faculty of Industrial Design Engineering, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: **g.d.s.ludden@utwente.nl**

In this paper, we first will discuss the related work on effects of robot body orientation, gaze behaviour and the use of several modalities in tour guide robots. Then we will present a field study where we aimed to find how robot orientation behaviour influences the group formations and orientations of the visitors. Next, we present will the results and discuss them. Finally, we will present design guidelines for non-verbal robot behaviour. The paper will end with a conclusion, in which we give directions for future research.

2 RELATED WORK

A tour guide robot for instance engages visitors and directs their attention to points of interest. This is similar to what human tour guides do intuitively. Human tour guides use their (body) orientation and give gaze cues to direct visitors' attention. However, most important are their subtle reactions to visitors' actions [7]. Kuzuoka et al. showed that a robot could effectively reshape the orientation of a visitor by changing its own orientation with a full body movement [10]. Also, human-like gaze cues can be successfully copied to robots, as shown by Yamazaki et al. They found that visitors showed higher engagement to a robot tour guide that used human-like gaze cues and its story than when the robot was not using these human-like gaze cues [11]. Sidner et al. found that head movements (and thus gaze cues) of the robot helped to keep people engaged during interaction [12]. Subtle gaze cues of robots can also be understood by people, as was shown by Mutlu et al. who let a robot describe an object among several other objects that were placed on a table. When the robot was "gazing" at the object it described, people found it easier to select the corresponding item [13].

The previously described body of work has focussed on copying two important types of cues that human guides use. However, robots are often able to apply a more diverse set of cues than body orientation and gaze cues alone. Different types of robots can use alternative modalities to give cues about their intentions. For example, if a robot uses a screen to convey information, visitors will stand close and orient themselves so that they can see the screen. However, when a robot uses arms to point and has no screen, visitors will probably orient themselves so they can easily see the robot and the exhibit the robot is pointing at.

Researchers have tried different modalities for museum robots to communicate intentions to their users. In the next paragraphs some examples of behaviour will be given to illustrate the effects of specific behaviours. The robot Rhino as developed by Burgard et al. blew a horn to ask visitors to get out of the way, which often had the opposite effect and made visitors stand in front of the robot until the horn sounded again [1]. Thrun et al. developed Minerva, the successor of Rhino. This robot did not have the problem that people clustered around when it wanted to pass, because it used several emotions and moods using its face and tone of voice. First, the robot asked in a happy and friendly state to get out of the way and if people did not react, the robot became angry after a while. With this behaviour Minerva was able to indicate its intentions and internal states successfully to the visitors [2]. However, the design of emotions and moods should be done carefully, as Nourbakhsh et al. found in the development of their robots. The robots Chips and Sweetlips showed moods based on their experiences that day. Visitors who only had a short interaction timeframe with the robots did not always understand these moods [4]. Touch screens and buttons have also been used for interaction purposes. These were found to make people stand closer to the robot, inviting them to interact with the buttons. This was for example found for the eleven Robox at the Expo.02 that were developed by Siegwart et al. [3]. However, buttons also can ruin the intended interaction. For example, Nourbakhsh et al. found that for the robot Sage [14] and Graf et al. for their robots in the museum of Kommunikation in Berlin (Germany) [6], people liked to push the emergency stop button and unintentionally stopped the robot from functioning.

All robots mentioned so far, had some interactive and social behaviour. However, specific guide behaviours - to engage multiple visitors and give information about exhibits - have still received little attention. To make a guided tour given by a robot a success, a smooth interaction between the robot guide and the visitors is essential, and therefore, interaction cues should be designed carefully.

Another challenge for museum robots is that they often have to interact with groups of people rather than with just one person. Research on group dynamics and behaviour of visitors gathering around a (dynamic) object in a museum setting or following a tour guide has revealed that visitors often stand in a specific formation (so-called F-formation) and react to each other and the (dynamic) exhibit (e.g. [7], [8], [15], [16]). For example, when a small group gathers around one person giving them information, they usually form a sort of (semi-) circle. In that way all group members can listen to the person who has the word [15]. Of course, the type of formation depends on the size of the group. However, the previously described formation is also recognizable when a human tour guide is guiding a (small) group of visitors and when people gather around a point of interest to all have the chance to see it [7]. When gathering around a museum object there are differences between gathering around interactive objects and static objects. When gathering around static objects, a lot of visitors get a chance to see the object at the same time. However, when gathering around interactive objects (often including a screen), fewer people can see the object at the same time [16], because people tend to stand closer to see the details shown on the screen or to directly interact with the (touch) screen. Museum exhibit designers tend to make the exhibits more interactive in order to keep the attention of the visitors, which also is effective for tour guide robots to attract visitors [4]. While these exhibits introduce more interactivity to the exhibition, it decreases the social interactions and collaborations between visitors [16]. Therefore, interactivity of robots should be designed for a larger group and other modalities than a screen/buttons should be used to shape the visitors' orientations and formations.

Our question is, can we design robots that have robot specific and intuitively understandable behaviour? To answer this question, robot designers have often resorted to directly copying human behaviour. In the design of other product categories, designers have often used anthropomorphism, (copying human forms and/or behaviour) in an abstract way rather than by directly copying. Subtly copying human forms or behaviour might likewise give cues about a product's intention and help people to understand the function of a product intuitively [17]. For robots, this implies that a robot does not have to directly resemble a human being, while it can still be capable of clearly

communicating its intentions. Creating a robot with some anthropomorphic features does not necessarily mean that the robot needs to be human-like. However, to smooth the interaction human-like cues or features can be used in the design of robots [18]. Another question is, what should be designed first; the behaviour or the appearance of the robot. In most research on robots and their behaviour, the visual design for the robot was made first, and afterwards accompanying behaviour was designed. We decided to start from the other end. In this study, we used a very basic robot that showed some anthropomorphic behaviour in its body orientation. We were interested to find if and how people react to this behaviour while the appearance of the robot is far from human-like. In this way we expected to find some general guidelines for robot behaviour to influence people's reactions to the robot, while the options for the design of the robot are still multiple.

3 STUDY DESIGN

The goal of this study was to determine how orientation behaviour of a very basic robot influenced visitors' orientation and the formations groups of visitors formed around the robot. The orientation behaviour of the robot was manipulated, while other interaction features were limited. To evaluate how visitors reacted to the robot, we performed a study in the Royal Alcázar in Seville (Spain). The robot gave short tours with four stops in the Hall of Festivities of in the Royal Alcázar.

Participants

Participants of the study were visitors of the Royal Alcázar. At both entrances of the room, all visitors were informed with signs that a study was going on. By entering the room, visitors gave consent to participate in the study. It was up to them if they wanted to join the short tour given by the robot or not. Approximately 500 people (alone or in groups ranging from 2 to 7 visitors) interacted with the robot during the study.

Robot

The robot used for the field study was a four-wheeled data collection platform (see Figure 1). The body of the robot was covered with black fabric to hide the computers inside. A bumblebee stereo camera was visible at the top of the robot, as well as a Kinect below the bumblebee camera. The robot was remotely operated. The operator was present in the room, but he was not in the area where the robot gave tours. The robot was operated using a laptop. The laptop screen was used to check the status of the robot, while the keyboard was used to actually steer the robot. The interaction modalities of the robot were limited; the robot was able to drive through the hall, change its orientation and play pre-recorded utterances. The instruction "follow me" was visible on the front of the robot, and signs informing people about the research (in English and Spanish) were fixed to the sides of the robot.

The robot used for this study was very basic. We chose this particular robot to be able to determine the effects of body orientation on visitors' reactions without being influenced by other factors in robot design and behaviour (such as aesthetics of the robot, pointing mechanisms, visualisations on a (touch-) screen or active face modifications).

During the study we used a user-centred iterative design approach [19] for the behaviour of the robot. When the robot charged in between sessions, we discussed robot behaviours that had the intended effect and behaviours that did not work well. During the study we modified the explanation of the robot after session one, because it became clear that visitors did not understand where to look. A total of three iterations were performed. In all iterations only changes to the explanation of the robot were made, however the content about the points of interest remained the same.

Procedure

The tour given by the robot took about 3 minutes and 10 seconds. The points of interest chosen were all visible on the walls of the room (no exhibits were placed anywhere in the room), however the position of the points of interest on the walls differed in height. During a tour the distance to drive in between the points of interest also differed, from approximately two meters up to approximately five meters. This was done so we could see if there were different visitor behaviours when following the robot. However in this paper we will not focus on the results on following the robot.

When visitors entered in the Hall of Festivities, the robot stood at the starting place (1) (see Figure 2) and began the tour by welcoming the visitors and giving some general information about the room. When the robot finished this explanation, it drove to the next stop (about 3.5 meters away), asking the visitors to follow. At the next stop (2) the robot told the visitors about the design of the figures on the wall that were all made with tiles, after which it drove the short distance (about 2 meters) to the next exhibit. At the third stop (3) the robot told the visitors about the banner that hung high above an open door. At the end of this story the robot asked the visitors to follow after which it drove the long distance to the last stop (about 5 meters). Here (at point 4) it gave information about the faces visible on the tiles on the wall. Before ending the tour the robot drove back to the starting point (about 3.5 meters), informed the visitors the tour had finished and wished them a nice day.

After a while, when new visitors had entered the room, the robot started the tour again. During the study the robot tried to persuade visitors to follow it with the sentences "please follow me" and "don't be afraid", when visitors were hesitant. In all cases it was up to the visitors to decide whether they followed the robot or not. Visitors were never instructed to follow the robot by researchers present in the room.

As the study was performed in a real life setting, with uninformed naïve visitors, we sometimes had to deviate a bit from the procedure. The robot had defined places for stops. However, sometimes the robot had to stop close to the defined place, because people walked or stood in front of the robot.

Figure 1. Impression of the robot and visitors in the site

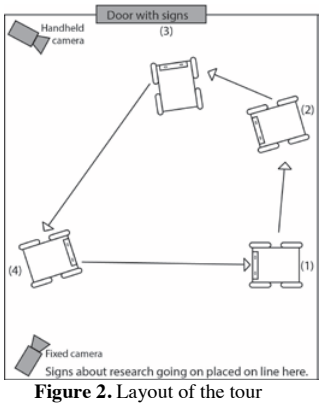


Figure 2. Layout of the tour

Another reason to deviate was when the robot lost the attention of all people who were following the tour. Then, it drove back to the starting place and started over again. If some visitors lost interest and left, but other visitors remained listening to the robot, it continued the tour.

When all visitors left the hall, or did not show any attention towards the robot, the trail was aborted, and restarted when new visitors entered the hall. Therefore the number of times the robot was presenting at each of the four exhibits was decreasing. The robot started the tour 87 times at the first exhibit, continued 70 times at the second exhibit. At the third exhibit the robot started its presentation 63 times and it finished the story only 58 times at the fourth exhibit. A total of 278 complete explanations at points of interest were performed (see table 1 for a specification of the actions per point of interest).

Manipulations

During the study, we manipulated the robot's orientation behaviour. Either the robot was orientated towards the point of interest or the robot was orientated towards the visitors. When it was orientated towards the point of interest, the front of the robot was in the direction of the point of interest. The points of interest were all located a few meters apart from each other. When the robot was oriented towards the visitors, its front was directed towards a single visitor or towards the middle of the group of visitors. See table 1 for a specification of the orientation of the robot per iteration and per point of interest.

In between the three iterations, some changes were made to the explanation by the robot. The explanations for the robot were developed in such a way that they could be used for both orientations of the robot. During the first iteration we observed that these explanation worked fine when the robot was oriented towards the points of interest. However, we found that it seemed unclear where to look when the robot was oriented towards the visitors. Therefore, for the second iteration, the explanations of the robot when oriented towards the visitors at points of interest two, three and four were modified. Information about where visitors had to look exactly to find the point of interest the robot explained about was added. As a result, the robot explained more clearly to the visitors "to look behind it" when it was orientated to the visitors and "to look here" when it was oriented towards the point of interest. Also, the sentences "please follow me" and "don't be afraid" were added to try to convince people to follow the robot to the next point.

Table 1: **Specification of manipulations**

	Robot actions	Point 1	Point 2	Point 3	Point 4
Iteration 1	**109**	**38**	**26**	**23**	**22**
To exhibit	*66*	*4*	*23*	*20*	*19*
To people	*27*	*27*	*0*	*0*	*0*
Excluded	*16*	*7*	*3*	*3*	*3*
Iteration 2	**90**	**25**	**24**	**22**	**19**
To exhibit	*42*	*0*	*10*	*17*	*15*
To people	*35*	*20*	*11*	*0*	*4*
Excluded	*13*	*5*	*3*	*5*	*0*
Iteration 3	**79**	**24**	**20**	**18**	**17**
To exhibit	*1*	*0*	*0*	*1*	*0*
To people	*65*	*16*	*18*	*16*	*15*
Excluded	*13*	*8*	*2*	*1*	*2*
Total	**278**	**87**	**70**	**63**	**58**
To exhibit	*109*	*4*	*33*	*38*	*34*
To people	*127*	*63*	*29*	*16*	*19*
Excluded	*42*	*20*	*8*	*9*	*5*

In the third iteration another modification was made to the explanation of the robot when it was oriented towards the visitors. The sentences were ordered in such a way that the robot would capture the attention of the visitors with something trivial, so people would not miss important parts of the explanations. All iterative sessions took about 1 hour and 40 minutes.

Data collection

During the study, the visitors were recorded with two cameras: a fixed camera that recorded the whole tour and a handheld camera that was used to record the facial expressions of the visitors close to the robot. Also, several visitors who followed (a part of) the tour were interviewed about their experiences. The interviews were sound recorded.

For this study only the data collected with the fixed camera was used, because the data from this camera gave a good overview of the room and the actions, orientation and formations of the visitors. We decided to not to use recordings from the cameras that were fixed on the robot, because their angle of view was limited to only the front of the robot. Using these recordings would not give us opportunities to study the behaviour of visitors who were next to or behind the robot (for example when the robot was oriented towards the exhibit), which in this study would lead to the loss of a lot of information on visitor orientation and formations. The proximity of the visitors was measured based on the number of tiles they stood away from the robot. Data collected through the short interviews was also not used in this analysis, because in this case we were only interested in how robot orientation influenced the actual orientation of visitors and their formations and less in their experience with the robot.

Data analysis

For the analysis, 236 robot actions of a total of 278 robot actions were used. Forty-two cases were excluded from analysis because no visitors were in the room or no robot was visible, because it was out of the angle of view of the camera, or the view was blocked by large numbers of visitors (for example a group with a human tour guide that did not show any interest in the robot). This resulted in 236 robot actions (278-42=236) in 3 iterations that were left for the analysis. The robot was oriented towards

the exhibit while it explained 127 times, and the robot was oriented towards the visitors while it presented 109 times.

We were interested in the reactions of the visitors that might be influenced by the robot orientation during each if these 278 complete explanations at the points of interest. However, exact visitor behaviour to search for was not defined before the study. We performed a content analysis of the recordings from the fixed camera. We isolated robot actions -the moments that the robot stood close to a point of interest and presented about it- in the data for coding purposes. Coding of the data was done by using a Grounded Theory Method [20] and use of an affinity diagram [21] for the open coding stage. No exact codes were defined before the start of the analysis. We defined the codes based on the actions of the visitors found in the video recordings. Some examples of found codes are: "standing very close to the robot and oriented towards each other," "visitors standing in a semi-circle and robot oriented towards the exhibit," "visitors losing interest during the robot story and robot oriented towards the visitors," "visitors walking towards the robot and robot oriented towards exhibit." We used a count method to compare the reactions of the visitors during the robot actions between the different robot orientations and the different points of interest.

10 % of the data was double coded and we found an overall inter-rater reliability of a=0.662 (Cohen's Kappa), which indicates a substantial agreement between the coders. Hence, one coder finished the coding of the dataset that was used for analysis.

4 RESULTS

We found that visitors stood far away more often when the robot was oriented towards the visitors (31 times, 24.4% of all cases in this condition) than when the robot was oriented towards the point of interest (17 times, 15.6% of all cases in this condition). Further, no differences were found in formations of the visitors between both conditions. However, when the robot was oriented towards the visitors, just 18 times (14.2% of all cases in this condition) visitors walked towards the robot, while when the robot was oriented towards the point of interest visitors walked towards the robot 25 times (22.9 % of all cases in this condition). In both conditions and at all stops, a lot of people (78% of all cases) were just walking by, showing no attention for the robot at all. However, most of the time one or a few visitors had already joined the robot by then. A few times we observed that visitors waited until the robot was free again and then followed the tour. Also, when some of the visitors left the robot, others stayed to hear the rest of the explanation about the point of interest.

We found more differences between visitor formations when we focussed our analysis on the interactions in stops two, three and four, while excluding stop one. We decided to exclude stop one from our analysis, because at that stop the robot was always oriented towards the visitors and it was not explaining about a specific point in the room. We found that when the robot provided information about points of interest two, three and four, more people lost interest when the robot was oriented towards the point of interest (22 times, 21% of all cases in this condition) than when the robot was oriented towards the visitors (8 times, 12.5 % of all cases in this condition). Also, 6 times (9.4 % of all cases in this condition) visitors did not have a clue where to look when oriented towards the visitors. This was never the case (0% of all cases in this condition) when the robot was oriented towards the point of interest.

The number of visitors standing close to the robot was comparable between both conditions (5 times, 3.9% of all cases with orientation towards the visitors and 6 times, 5.5% of all cases with orientation towards the exhibit). However a difference was found between the exhibits. Only at stops one and two, did visitors stand really close to the robot when the robot was oriented towards the visitors. However, in the condition where the robot was oriented towards the point of interest people stood close to the robot at all stops. From reviewing the video, we observed that when people stood very close to the robot and the robot was oriented towards them, visitors only seemed to focus on the robot, while visitors focussed on the point of interest when the robot was oriented towards the point of interest.

Also we found some differences in visitor reactions between the different stops. Fewest visitors walked towards the robot at stop three (5 times; 9.3% of the cases in this condition), most did at stop four (16 times, 30.2% of the cases in this condition). Visitors lost interest in the story and the robot most often at stop three (14 times; 25.9% of all cases in this condition) and least often in stop four (6 times; 11.3% of all cases in this condition).

Looking only at the differences between the stops over both conditions, we found that many more single visitors and pairs joined the robot for at least one stop (86 times, 36.4% of all cases) than that people gathered around the robot in any group formation (38 times, 16.1% of all cases). We found that during 11 robot actions (4.7% of all cases) visitors stood less than 30 cm away from the robot. During 48 robot actions (20.3% of all cases) people stood more than 3 meters away from the robot. In 131 robot actions (55.5% of all cases) visitors stood between the 30 cm and 3 meters from the robot. Note that these cases can overlap, because there could be more than one visitor at the same time. In the rest of the cases no visitors or no robot were in the field of view or the visitors did not join the robot tour.

5 DISCUSSION

Influences of robot orientation

We found that visitors stood far away from the robot more often when the robot was oriented towards the visitors than when it was oriented towards the point of interest. Furthermore, we found that visitors tended to walk towards the robot more often when the robot was oriented towards the point of interest than when the robot was oriented towards the visitors. One possible explanation for this visitor reaction might be that visitors could not hear the robot well enough. However, we do not consider this a valid explanation in all cases, since people generally in both conditions followed the robot from a distance and they were able to hear the explanations of the robot. Therefore, we argue that it might be that the visitors felt that a distance was created by this specific orientation of the robot. This may have caused that people felt safer to approach the robot when it was oriented towards the point of interest. Perhaps, the robot kept people at a distance with its "eyes" when it was oriented towards the visitors. This finding is in line with findings from other studies that people walked closer to a robot that was not following them with gaze than when the robot was following them with gaze, as shown by Mumm and Mutlu [22]. Remarkable was that more people lost interest when the robot was oriented towards the point of interest than when the robot was oriented towards the

visitors. As we argued before, the orientation of the robot towards the point of interest might have felt safer for people, at the same time, it might also have given them the feeling of being excluded, which made them leave the robot.

In stops one and two, several people were walking towards the robot, because the robot captured their attention and they were curious to see what it was for. Fewest visitors walked towards the robot at stop three, most did at stop four. Visitors probably did not have to walk to the robot in stop three, because it was really close to stop two. From stop three to stop four was the longest walk. Visitors who walked towards the robot in stop four were probably a bit reserved following the robot and therefore just walked to the robot when it had already started the next explanation. Apart from that, stop three was close to an open door, the entrance to the next room, therefore people who lost interest could easily walk away from the robot into the next room. When visitors followed to stop four, the last stop of the tour, they were likely to follow the robot the whole tour. We assume these visitors liked to hear the explanations of the robot and stayed with the robot until the final explanation, therefore fewer of them left the robot in stop four.

Visitor actions that were coded with "losing interest" showed that most of the time not all visitors lost their interest at the same moment. When one visitor of a pair or group walked away, the other(s) either followed the leaving person directly, stayed until the end of the explanation at that point or stayed until the end of the tour. This indicates that visitors of pairs or groups gave each other the time to do what they liked and that they did not have to leave together at the same moment. An advantage was that for most people it was clear that the robot just gave a short tour, so the people who left did not have to wait for a long time if the others stayed. In some cases we observed visitors discussing if they would follow the robot and in the end they decided that one would follow the tour, and that the other would wait outside the research area. It was important for the robot that when one visitor lost interest, most of the time the robot had other visitors (either close or far) who were still interested in the robot and the story, so it went on with the story.

We found a difference in the distance people kept from the robot and the orientation of the robot. Only at stops one and two, did visitors stand really close to the robot when the robot was oriented towards the visitors. However, when the robot was oriented towards the point of interest, visitors stood very close in all four stops. It seemed that when visitors stood very close to the robot and the robot was oriented towards them, visitors only had interest in the robot as an object and they tried to make contact with the robot (by waving at the robot or bringing their eyes on the same height as the lenses of the camera of the robot). We think this visitor behaviour mainly occurred at points one and two, because at these moments the robot captured people's attention. In stop three and four only visitors who were already following the tour seemed to be present and people who were only interested in the robot as an object did not disturb the robot guide and its visitors in these points. When visitors stood close and the robot was oriented towards the point of interest, the visitors probably could not hear the voice of the robot well enough to follow the story in the crowded area, while they were interested in the point of interest the robot presented about and wanted to hear the explanation.

Visitors who were interacting with the robot oriented towards them, sometimes appeared to have no clue where to look. This indicates that visitors were sensitive for the orientation of the robot. More verbal cues were added to the explanation of the robot in iterations 2 and 3. However, during these iterations, we still observed that when the robot was oriented towards them visitors got the clue where to look later than they expected. So, even though we changed the explanation of the robot to make more clear where to look and started with something trivial, just as human tour guides do [23], visitors did not readily understand where to look. This might be due to the length of the explanations of the robot. These were much shorter than explanations given by a human tour guide at a point of interest usually are. So, in general visitors had less time to focus again before they would miss something. The robot orientation towards the point of interest avoided this problem.

Visitor reactions to the "eyes" of the robot

Our observations showed that visitors were aware of the lenses of the camera on the robot and responded to them as if they were the eyes of the robot. This can for example be seen from the observation that some visitors waved at the camera when they arrived or when they left the robot. People also stood in front of the camera when they wanted to make contact with the robot. The observation that people are sensitive to the camera of a robot and orient in front of it was also made by Walters et al. [24]. These examples make clear that visitors react to the orientation of the robot and probably see the lenses of the camera as the eyes of the robot. Another observation that strengthens these conclusions is that visitors most often lost their interest in stop three. In this stop the explanation was difficult to understand because the story was about a banner that hung high in the room, above an open door. When the robot was oriented towards the exhibit, it seemed as if it was "looking" at a point in the other room because it was not able to tilt its orientation upwards. This confused the visitors, even when the robot was clear in its explanation about where to look.

Differences between robot guide and human tour guide

We found that visitors reacted differently to the robot tour guide than we would expect from observed reactions to a human tour guide. First of all fewer groups and more individual visitors or pairs of visitors joined the robot tour guide. Also, visitors seemed not prone to join strangers, but rather waited till the tour was finished and they could join a new tour.

Most visitors stood between 30 cm and 3 meters from the robot. When there were visitors standing very close or far away from the robot, there also could be visitors who stood at average distance (between 30 cm and 3 m) from the robot. While most visitors stood at an average distance, standing really close or staying at a distance differs from visitor behaviour shown when they follow a human tour guide. Most of the time visitors of a group of a human tour guide does not show that large difference in proxemics to a guide and often stand in a semi-circle to give everyone a chance to see the guide [7]. Also, Walters et al. [25] and Joosse et al. [26] showed in controlled experiments that people allowed different approach distances and appropriate proxemics for a robot than they allow for confederates. This leads to the conclusion that we cannot assume that people react the same to robot tour guides as to human tour guides.

Implications of study set-up

The study was performed in the wild which influenced the execution of the study and the manner of analysis. One disadvantage was that the situations of guiding could not be controlled. Also, less information of the visitors could be obtained. For example, we could not have extended questionnaires because people did not want to spend their time to filling these in.

We performed the study in several iterations in which we modified the explanation of the robot. Without these modifications to the explanations, we would not have been able to perform the manipulation of the orientation of the robot, because with the original explanation visitors did not seem to know where to find the point of interest when the robot was oriented towards them. This led to the following differences between the iterations. In iteration one the robot was mainly oriented towards the point of interest. In iteration two the modification of the explanation seemed insufficient, so the robot was mainly oriented towards the points of interest. In iteration three the robot was mainly oriented towards the visitors.

An advantage of the in-the-wild set-up of this study was that we observed the reactions of the visitors the way they would probably be if an autonomous tour guide robot were to be installed in the Royal Alcázar. The findings of this research were an important step for the development of FROG, because with in-the-lab studies with small groups of users, it would be difficult to create a similar environment including people who are acquaintances and strangers. Probably, we would also not have found how people react when the robot is already occupied by strangers, while in this set-up we did find interesting reactions of visitors in the real-world context.

Also, we used a very basic robot with limited interaction modalities. Nevertheless, the influence of body orientation and was largely observable in the visitor reactions. We expect that these factors will keep influencing visitor reactions when more robot modalities (such as arms to point, or a screen to show information) are added to the robot.

6 DESIGN IMPLICATIONS FOR ROBOT BEHAVIOUR

Findings described in the previous section led to the following set of design guidelines for the design of the non-verbal behaviour of a tour guide robot, that can be used irrespective of the visual design of the robot.

1) *Check for visitors standing far away when people close-by leave the robot during the explanation.*

The robot did not only catch the attention of people who were standing close such as we would expect with human tour guides. Visitors who chose to stay at a distance also followed the robot tour. Although these visitors were interested in the story and the robot, they did not want to be close. The tour guide robot should therefore not only focus on visitors nearby, but scan the surrounding once in a while and go on with the story or tour if it detects visitors who are not standing close, but show an orientation towards the robot and stay there during explanation. This behaviour of scanning the environment is even more important when visitors who are standing close all leave. Also, the robot should not rely solely on its detection of visitors by gaze (cameras directed to the front-side of the robot) to determine whether it should go on or stop the explanations, because in some situations the visitors tend to stand next to or behind the robot, while they are still interested in its story. The robot should be aware of these visitors and continue the explanation at the exhibit.

2) *Define behaviour of people standing close-by to decide whether to stop or to continue the story.*

For visitors who are standing close, the robot should make a distinction between people standing very close that are following the tour and people standing very close that show interest in the robot only. When people are still following the story, the robot should go on giving information. However, when people only show interest in the robot, the robot can decide to play with them a bit and show it is aware of the visitors being there. Possibly the robot can catch their attention for the story and change the playful or disturbing interaction to a guide-visitors interaction.

3) *Ask people to join the tour when they are hesitant to join strangers.*

The robot mainly attracted individuals and pairs who did not join other people who had started following the tour before them. People preferred to wait until others had left before they decided to join the tour. In other cases they just followed the tour from a distance, when other people were already close. This fits the purpose of the robot, however it would be nice if the small groups joined in order to all have an even better experience of the robot, because the robot cannot focus on all visitors close-by and far away. To do so, the robot can at certain moments in the story decide to scan for visitors and invite them to join.

4) *When camera lenses are clearly visible in the design of the robot, use them as eyes*

In our field study, a stereo bumblebee camera and a Kinect were clearly visible on the robot. Our experience in this study taught us that visitors see the stereo camera on top of the robot as the eyes of the robot. Therefore, when the camera cannot be hidden, the camera should be designed as eyes, including the design of gaze cues and gaze direction. Using these cues, especially when people expect them already, will probably smoothen the human-robot interaction. In our case, the FROG robot is not a humanoid robot, while the camera is visible. Therefore, we argue that a visible camera should be used as eyes of a robot, because this will support the mental model users will create of the robot.

7 CONCLUSION AND FUTURE WORK

To conclude, the orientation of the robot is important to shape the visitors' reactions. When it was clear to the visitors what to look at (mostly when the robot was oriented towards the exhibit), they became engaged more easily in the robot guided tour. However more people became interested in the robot when it was oriented towards the exhibit. Also, more people lost interest in the robot and the story when it was oriented towards the exhibit than when it was oriented towards the visitors. Therefore, keeping the attention should be done in a different way than capturing the attention of the visitors.

With this research we focused on visitors' orientation and group formations that visitors formed around the tour guide robot. However, in order to design robot behaviours for giving an effective tour, visitors' reactions when the robot is guiding them from one point of interest to the next should also be analysed, and guidelines about how to shape these should be developed. We will further use the recording from this study to

analyse the visitor reactions to the robot guiding behaviour (e.g. following the robot from a distance or really close to the robot, hesitating to follow the robot) as well as visitor reaction at stops at points of interest while following the robot.

The present study has given us insight into how robot orientation and behaviour can influence people's formations and reactions. A future research question, is to find how the combined effects of robot behaviour and visual design of a robot will influence the number of people who stop to see the robot and eventually join the robot guided tour. In the future we will perform more elaborate evaluations including more robot modalities and behaviours.

ACKNOWLEDGEMENT

The research leading to these results received funding from the European Community's 7th Framework Programme under Grant agreement 288235 (http://www.frogrobot.eu/)."

We would like to thank N. Pérez-Higueras, R.R. Vigo and J. Pérez-Lara for preparing and controlling the robot. We would like to thank M.P. Joosse for his help with data analysis.

REFERENCES

[1] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "Experiences with an interactive museum tour-guide robot," *Artif. Intell.*, vol. 114, no. 1–2, pp. 3–55, 1999.

[2] S. Thrun, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz, "MINERVA: A second-generation museum tour-guide robot," in *Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on*, 1999, vol. 3, no. May, pp. 1999–2005.

[3] R. Siegwart, K. O. Arras, S. Bouabdallah, D. Burnier, G. Froidevaux, X. Greppin, B. Jensen, A. Lorotte, L. Mayor, M. Meisser, R. Philippsen, R. Piguet, G. Ramel, G. Terrien, and N. Tomatis, "Robox at Expo.02: A large-scale installation of personal robots," *Rob. Auton. Syst.*, vol. 42, no. 3–4, pp. 203–222, Mar. 2003.

[4] I. R. Nourbakhsh, C. Kunz, and T. Willeke, "The Mobot Museum Robot Installations: A Five Year Experiment," in *2003 IEEE/RJS International Conference on Intelligent Robots and Systems*, 2003, pp. 3636–3641.

[5] M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Interactive humanoid robots for a science museum," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 2006, pp. 305–312.

[6] B. Graf and O. Barth, "Entertainment Robotics: Examples, Key Technologies and Perspectives," *Safety*, vol. 6, no. 6, pp. 8–12, 2002.

[7] K. Best, "Making museum tours better: understanding what a guided tour really is and what a tour guide really does," *Museum Manag. Curatorsh.*, vol. 27, no. 1, pp. 35–52, 2012.

[8] Adam Kendon, "Spacing and Orientation in Co-present Interaction," in *Development of Multimodal Interfaces: Active Listening and Synchrony; Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers*, 2010, p. pp 1–15.

[9] B. Jensen, N. Tomatis, and L. Mayor, "Robots meet Humans-interaction in public spaces," *IEEE Trans. Ind. Electron.*, vol. 52, no. 6, pp. 1530–1546, 2005.

[10] H. Kuzuoka, Y. Suzuki, J. Yamashita, and K. Yamazaki, "Reconfiguring spatial formation arrangement by robot body orientation," in *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction - HRI '10*, 2010, pp. 285–292.

[11] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka, "Precision Timing in Human-Robot Interaction : Coordination of Head Movement and Utterance," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2008, pp. 131–139.

[12] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh, "Where to Look: A Study of Human-Robot Engagement," in *Proceedings of the 9th international conference on Intelligent User Interfaces*, 2004, pp. 78–84.

[13] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior," in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction*, 2009, vol. 2, no. 1, pp. 69–76.

[14] I. R. Nourbakhsh, J. Bobenage, S. Grange, R. Lutz, R. Meyer, and A. Soto, "An affective mobile robot educator with a full-time job," *Artif. Intell.*, vol. 114, no. 1–2, pp. 95–124, Oct. 1999.

[15] P. Marshall, Y. Rogers, and N. Pantidi, "Using F-formations to analyse spatial patterns of interaction in physical environments," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*, 2011, pp. 445–454.

[16] C. Heath, D. Vom Lehn, and J. Osborne, "Interaction and interactives: collaboration and participation with computer-based exhibits," *Public Underst. Sci.*, vol. 14, no. 1, pp. 91–101, 2005.

[17] C. DiSalvo and F. Gemperle, "From seduction to fulfillment: the use of anthropomorphic form in design," in *DPPI '03 Proceedings of the 2003 international conference on Designing pleasurable products and interfaces*, 2003, pp. 67–72.

[18] C. F. Disalvo, F. Gemperle, J. Forlizzi, S. Kiesler, and H. C. Interaction, "All Robots Are Not Created Equal : The Design and Perception of Humanoid Robot Heads," in *DIS '02 Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, 2002, pp. 321–326.

[19] J. D. Gould, T. J. Watson, and C. Lewis, "Designing for usability: key principles and what designers think," *Mag. Commun. ACM*, vol. 28, no. 3, pp. 300–311.

[20] J. M. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," *Qual. Sociol.*, vol. 13, no. 1, pp. 3–21, 1990.

[21] C. Courage and K. Baxter, *Understanding Your Users: A Practical Guide to User Requirements Methods, Tools, and Techniques*, vol. 2, no. 4. Morgan Kaufmann, 2005, p. 704.

[22] J. Mumm and B. Mutlu, "Human-Robot Proxemics : Physical and Psychological Distancing in Human-Robot Interaction," in *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on. IEEE*, 2011, pp. 331–338.

[23] D. E. Karreman, E. M. A. G. van Dijk, and V. Evers, "Contextual Analysis of Human Non-verbal Guide Behaviors to Inform the Development of FROG, the Fun Robotic Outdoor Guide," *Hum. Behav. Underst.*, vol. LNCS 7559, pp. 113–124, 2012.

[24] M. L. Walters, K. Dautenhahn, K. L. Koay, C. Kaouri, R. Te Boekhorst, C. Nehaniv, I. Werry, and D. Lee, "Close encounters: Spatial distances between people and a robot of mechanistic appearance," in *Proceedings of 2005 5th IEEE-RAS International Conference on Humanoid Robots*, 2005, vol. 2005, pp. 450–455.

[25] M. L. Walters, K. Dautenhahn, S. Woods, K. L. Koay, R. Te Boekhorst, and D. Lee, "Exploratory studies on social spaces between humans and a mechanical-looking robot," *Conn. Sci.*, vol. 18, no. 4, pp. 429–439, 2006.

[26] M. Joosse, A. Sardar, and V. Evers, "BEHAVE: A Set of Measures to Assess Users' Attitudinal and Non-verbal Behavioral Responses to a Robot's Social Behaviors," *Soc. Robot.*, vol. LNAI 7072, pp. 84–94, 2011.

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Performing Facial Expression Synthesis on Robot Faces: A Real-time Software System

Maryam Moosaei[1], **Cory J. Hayes, and Laurel D. Riek**

Abstract. The number of social robots used in the research community is increasing considerably. Despite the large body of literature on synthesizing facial expressions for synthetic faces, there is no general solution that is platform-independent. Subsequently, one cannot readily apply custom software created for a specific robot to other platforms. In this paper, we propose a general, automatic, real-time approach for facial expression synthesis, which will work across a wide range of synthetic faces. We implemented our work in ROS, and evaluated it on both a virtual face and 16-DOF physical robot. Our results suggest that our method can accurately map facial expressions from a performer to both simulated and robotic faces, and, once completed, will be readily implementable on the variety of robotic platforms that HRI researchers use.

1 Introduction

Robotics research is expanding into many different areas, particularly in the realm of human-robot collaboration (HRC). Ideally, we would like robots to be capable partners, able to perform tasks independently and effectively communicate their intentions toward us. A number of researchers have successfully designed robots in this space, including museum-based robots that can provide tours [10], nurse robots that can automatically record a patient's bio-signals and report the results [22], wait staff robots which can take orders and serve food [17], and toy robots which entertain and play games with children [30].

To facilitate HRC, it is vital that robots have the ability to convey their intention during interactions with people. In order for robots to appear more approachable and trustworthy, researchers must create robot behaviors that are easily decipherable by humans. These behaviors will help express a robot's intention, which will facilitate understanding of current robot actions or the prediction of actions a robot will perform in the immediate future. Additionally, allowing a person to understand and predict robot behavior will lead to more efficient interactions [18, 20].

Many HRI researchers have explored the domain of expressing robot intention by synthesizing robot behaviors that are human-like and therefore more readily understandable [29, 13, 21, 5]. For example, Takayama et al. [35] created a virtual PR2 robot and applied classic animation techniques that made character behavior more humanlike and readable. The virtual robot exhibited four types of behaviors: forethought and reaction, engagement, confidence, and timing. These behaviors were achieved solely by modifying the robot's body movement. Results from this study suggest that these changes in body movement can lead to more positive perceptions of the robot, such as it possessing greater intelligence, being more approachable, and being more trustworthy.

While robots like the PR2 are highly dexterous and can express intention through a wide range of body movements, one noticeable limitation is that there are some subtle cues they can not easily express without at least some facial features, such as confusion, frustration, boredom, and attention [16]. Indeed, the human face is a rich spontaneous channel for the communication of social and emotional displays, and serves an important role in human communication. Facial expressions can be used to enhance conversation, show empathy, and acknowledge the actions of others [7, 15]. They can be used to convey not only basic emotions such as happiness and fear, but also complex cognitive states, such as confusion, disorientation, and delirium, all of which are important to detect. Thus, robot behavior that includes at least some rudimentary, human-like facial expressions can enrich the interaction between humans and robots, and add to a robot's ability to convey intention.

HRI researchers have used a range of facially expressive robots in their work, such as the ones shown in Figure 1. These robots offer a great range in their expressivity , facial degrees-of-freedom (DOF), and aesthetic appearance. Because different robots have different hardware, it is challenging to develop transferable software for facial expression synthesis. Currently, one cannot reuse the code used to synthesize expressions on one robot's face on another [6]. Instead, researchers are developing their own software systems which are customized to their specific robot platforms, reinventing the wheel.

Another challenge in the community is many researchers need to hire animators to generate precise, naturalistic facial expressions for their robots. This is very expensive in terms of cost and time, and is rather inflexible for future research. A few researchers use commercial off-the-shelf systems for synthesizing expressions on their robots, but these are typically closed source and expensive as well.

Thus, there is a need in the community for an open-source software system that enables low-cost, naturalistic facial expression synthesis. Regardless of the number of a robot's facial DOFs, from EDDIE [34] to Geminoid F [9], the ability to easily and robustly synthesize facial expressions would be a boon to the research community. Researchers would be able to more easily implement facial expressions on a wide range of robot platforms, and focus more on exploring the nuances of expressive robots and their impact on interactions with humans and less on laborious animation practices or the use of expensive closed-source

[1] The authors are with the Computer Science and Engineering department, University of Notre Dame,{mmoosaei,chayes3,lriek}@nd.edu

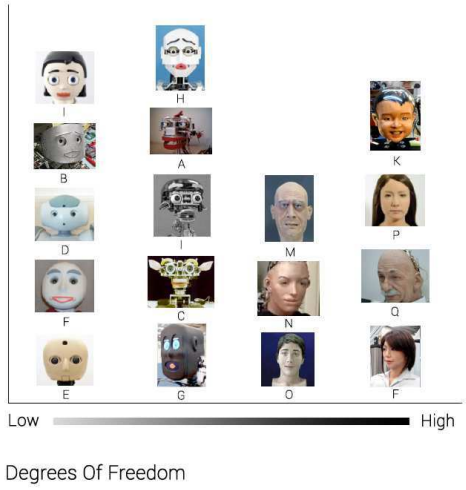


Figure 1. Examples of robots with facial expressivity used in HRI research with varying degrees of freedom. A: EDDIE, B:Sparky, C: Kismet, D: Nao , E: M3-Synchy, F: Bandit, G: BERT2, H: KOBIAN, F: Flobi, K: Diego, M: ROMAN, N: Eva, O: Jules, P: Geminoid F, Q: Albert HUBO , R: Repliee Q2

software.

In this paper, we describe a generalized software framework for facial expression synthesis. To aid the community, we have implemented our framework as a module in the Robot Operating System (ROS), and plan to release it as open source. Our synthesis method is based on performance-driven animation, which directly maps motions from video of a performers face onto a robotic (or virtual) face. However, in addition to enabling live puppeteering or "play-back", our system also provides a basis for more advanced synthesis methods, like shared gaussian process latent variable models [14] or interpolation techniques [23].

We describe our approach and its implementation in Section 2, and its validation in both simulation and on a multi-DOF robot in Section 3. Our results show that our framework is robust to be applied to multiple types of faces, and we discuss these findings for the community in Section 5.

2 Proposed method

Our model is described in detail in the following sections, but briefly our process was as follows: We designed an ROS module with five main nodes to perform performance driven facial expression synthesis for any physical or simulated robotic face.These nodes include:

S, a sensor, capable of sensing the performer's face (e.g., a camera)

P, a point streamer, which extracts some facial points from the sensed face

F, a feature processor, which extracts some features from the facial points coming from the point streamer

T, a translator which translates the extracted features from *F* to either the servo motor commands of physical platforms or the control points of a simulated head

$C : C_1 \dots C_n$, a control interface which can be either an interface to control the animation of a virtual face or motors on a robot.

These five nodes are the main nodes for our synthesis module. However, if desired, a new node can be added to generate any new functionality.

Figure 2 gives an overview of our proposed method. Assume one has some kind of sensor, (*S*), which senses some information from a person's (*pr*) face. This information might consist of video frames, facial depth, or output of a marker/markerless tracker. *pr* can be either a live or recorded performer. In our general method, we are not concerned about identifying the expressions on the *pr*'s face. We are concerned about how to use the expressions to perform animation/synthesis on the given simulated/physical face. *S* senses *pr* and we aim to map the sensed facial expressions onto the robot's face.

Basically, we use a point streamer *P*, to publish information from a provided face. Any other ROS node can subscribe to the point streamer to synthesize expressions for an simulated/physical face. A feature processor *F*, subscribes to the information published by the point streamer and processes this information. *F* extracts useful features out of all of the facial information published by the point streamer. Then, a translator, *T*, translates extracted features to control points of a physical/simulated face. Finally, a control interface $C : C_1 \dots C_n$ moves the physical/simulated face to a position which matches *pr*'s face.

2.1 ROS implementation

Figure 2 depicts the required parts for our proposed method. The software in our module is responsible for three tasks: (1) Obtaining input, (2) Processing input, (3) Actuating motors/control points accordingly

These responsibilities are distributed over a range of hardware components; in this case, a webcam, an Arduino board, a servo shield, and servo motors.

A local computer performs all processing tasks and collects user input. The data is then passed to the control interface, $C : C_1 \dots C_n$ which can either move actuators on an physical robot or control points on a virtual face. While Figure 2 shows the most basic version of our system architecture, other functionality or services can be added as nodes. Below, we describe each of these nodes in detail as well as the ROS flow of our method.

S, the sensor node, is responsible for collecting and publishing the sensor's information. This node organizes the incoming information from the sensor and publishes its message to the topic `/input` over time. The datatype of the message that this node publishes depends on the sensor. For example, if the sensor is a camera, this node publishes all incoming camera images. Examples of possible sensors include a camera, a Kinect, or a motion capture system. This node can also publish information from pre-recorded data, such as all frames of a pre-recorded video.

P, the point streamer node, subscribes to the topic `/input` and extracts some facial points from the messages it receives. This node extracts some facial points and publishes them to the topic `/points`.

F, the feature processor node, subscribes to the topic `/points`. Node *F* processes all the facial points published by *P*. *F* extracts useful features from these points that can be used to map the facial expressions of a person to the physi-

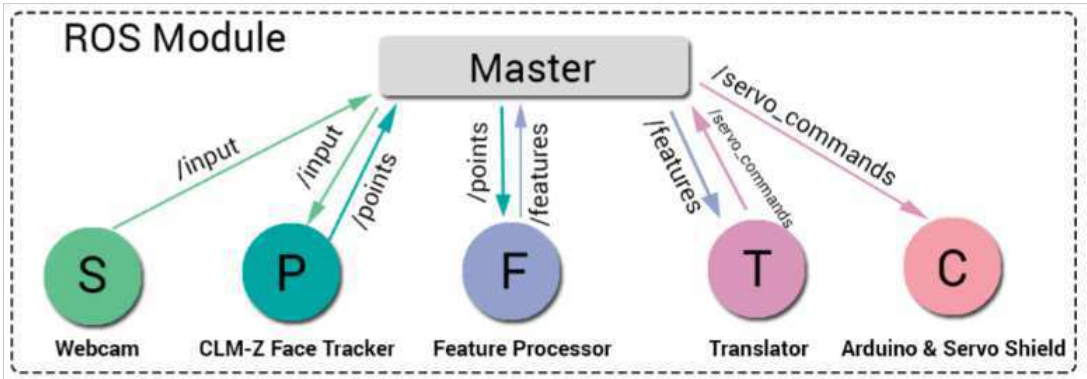


Figure 2. Overview of our proposed method.

cal/simulated face. This node publishes feature vectors to the topic `/features`.

T, the translator node, subscribes to the topic `/features`. and translates the features to DOFs available on the robot's face. Basically, this node processes the message received from the topic `/features` and produces corresponding movements for each of the control points on a robotic or virtual character face. This node publishes its output to the topic `/servo_commands`.

$C : C_1 \dots C_n$: The control interface node subscribes to the topic `/servo_commands` and actuates the motors of a physical robot or control points of a simulated face. We show $C : C_1 \dots C_n$ because a control interface might consist of different parts. For example, in case of a physical robotic head, the control interface might include a microcontroller, a servo shield, etc. We show the combination of all of these pieces as a single node because they cooperate together to actuate the motors. $C : C_1 \dots C_n$ subscribes to the topic `/servo_commands` which contains information about the exact movement for each of the control points of the robotic/simulated face. This node then makes a readable file for the robot containing the movement information and sends it to the robot.

2.2 An example of our method

There are various ways to implement our ROS module. In our implementation in ROS, we used a webcam as *S*. We chose the CLM face tracker as *P*. In *F*, we measured the movement of each of the facial points coming from the point streamer over the time. In *T*, we converted the features to servo commands for the physical robot and slider movements of the simulated head. In *C*, we used an Arduino Uno and a Renbotic Servo Shield Rev2 for sending commands to the physical head. For the simulated faces, *C* generates source files that the Source SDK was capable of processing.

We intended to use this implementation in two different scenarios: a physical robotic face as well as a simulated face. For a physical robot, we used our bespoke robotic head with 16 servo motors. For a simulated face, we used "Alyx", an avatar from video game Half-Life 2 from the Steam Source SDK. We describe each subsystem in detail in the following subsections.

2.2.1 Point streamer P

We employed a Constrained Local Model (CLM)-based face tracker as the point streamer in our example implementation. CLMs are person-independent techniques for facial feature tracking similar to Active Appearance Models (AAMs), with the exception that CLMs do not require manual labeling [12]. In our work, we used an open source implementation of CLM developed by Saragih et al. [1, 33, 11].

We ported the code to run within ROS. In our implementation, `ros_clm` (the point streamer) is an ROS implementation of the CLM algorithm for face detection. The point streamer `ros_clm` publishes one custom message to the topic `/points`. This message to the topic includes 2D coordinates of 68 facial points. This message is used to stream the CLM output data to anyone who subscribes to it.

As shown in the Figure 2, when the *S* node (webcam) receives a new image, it publishes a message containing the image data to the `/input` topic. The master node then takes the message and distributes it to the P node (ros_clm) because it is the only node that subscribes to the `/input` topic.

This initiates a callback in the *P* `ros_clm` node, causing it to begin processing the data which is basically tracking a mesh with 68 facial points over time. The `ros_clm` node sends its own message on the `/points` topic with the 2D coordinates of the 68 facial feature points.

2.2.2 Feature processor F

The *F* node subscribes to the topic `/points`. The *F* node receives these facial points. Using the position of two eye corners, *F* removes the effects of rotation, translation, and scaling. Next, in each frame, *F* measures the distance of each facial point to the tip of the nose as a reference point and saves 68 distances in a vector. The tip of the nose stays static in transition from one facial expression to the other. If the face has any in-plane translation or rotation, the distances of facial points from the tip of the nose will not be affected.

Therefore, any change in the distance of a facial point relative to the tip of the nose point over time would mean a facial expression is occurring. *F* publishes its calculated features to the topic `/features`.

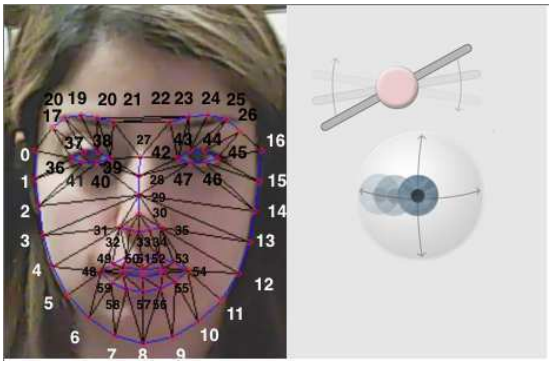


Figure 3. Left: the 68 facial feature points of CLM face tracker, Right: an example robotic eyebrow with one degree of freedom and an example robotic eyeball with two degrees of freedom

2.2.3 *Translator T*

The *T* node subscribes to the topic `/features` and produces appropriate commands for servos of a physical robot or control points of a simulated face. *F* keeps track of any changes in the distances of each facial point to the tip of the nose and publishes them to the `/features` topic. The *T* node has the responsibility of mapping these features to corresponding servo motors and servo ranges of a physical face, or to the control points of a simulated head. *T* performs this task in three steps. The general idea of these steps is similar to the steps Moussa et al. [28] used to map the MPEG-4 Facial Action Parameters of a virtual avatar to a physical robot.

In the first step, for each of the servo motors, we found a group of one or multiple CLM facial points whose movement significantly affected the motor in question. For example, as Figure 3 shows, the CLM tracker tracks five feature points on left eyebrow (22,23,24,25,26). However, the robot face shown in Figure 3 has only one motor for its left eyebrow. Therefore, the corresponding feature group for the robots left eyebrow, would be 22,23,24,25,26.

T converts the movement of each group of the CLM feature points to a command for the corresponding servo motor of a physical robot or control point of a simulated face. We used two examples in this paper, one with a simulated face and one with our bespoke robot. As an example, Table 1 shows the corresponding group of CLM points for each of the 16 servo motors of our bespoke robot

We averaged the movements of all of the points within a given group to compute only one number as the command for each motor/control point. To demonstrate this principle, our bespoke robot has a single motor for the right eyebrow. However, as Figure 3 shows, the CLM face tracker tracks five feature points on right eyebrow. If a performer raises their right eyebrow, the distance of these five points to the tip of the nose increases. We average the movements of these five points and use that value to determine the servo command for the the robot's right eyebrow.

Servo motors have a different range of values than that of feature points. Therefore, in the second step, we created a conversion between these values. The servos in our robot accept values between 1000 and 2000.

To find the minimum and maximum movement of each group of points associated with each servo, we asked a performer to make a wide range of extreme facial movements while seated in front of a webcam connected to a computer running CLM. For example, we asked the performer to raise their eyebrows to their extremities, or open their mouth to its maximum. Then, we manually modified the robot's face to match the extreme expressions on the subject's face and recorded the value of each motor. This way, we found the minimum and maximum movement for each group of facial feature points as well as for each servo motor.

In the last step, we mapped the minimum, maximum, and default values of the CLM facial points and the servo motors. Some servo motors had a reversed orientation with the facial points. For those servos, we flipped the minimum and maximum. In order to find values for a neutral face, we measured the distance of feature points to the tip of the nose while the subject had a neutral face. We also manually adjusted the robot's face to look neutral and recorded servo values.

Using the recorded maximum and minimum values, we applied linear mapping and interpolation (c.f., Moussa et al.) to find the criteria of mapping facial distances to servo values [28]. These criteria are used to translate facial points in each unseen incoming frame to the robot's servo values. The *T* node publishes a set of servo values to the topic `/servo_commands`.

2.2.4 *Control interface* $C : C_1 \dots C_n$

The *C* node subscribes to the topic `/servo_commands` and sends the commands to the robot. The servo motors of our robot are controlled by an interface consisting of an Arduino UNO connected to a Renbotic Servo Shield Rev2. ROS has an interface that communicates with Arduino through the rosserial stack [2]. By using rosserial_arduino, a subpackage of rosserial, one can add libraries to the Arduino source code to integrate Arduino-based hardware in ROS. This allows communication and data exchange between the Arduino and ROS.

Our system architecture uses rosserial to publish messages containing servo motor commands to the Arduino in order to move the robot's motors. The control interface receives the desired positions for the servo motors at 24 frames-per-second (fps). For sending commands to the simulated face, C generates source files that the simulated face is capable of processing.

Table 1. The facial parts on the robot, and corresponding servo motors and CLM tracker points.

Facial Part	Servo Motor #	CLM Points
Right eyebrow	1	17,18,19,20
Left eyebrow	2	23,24,25,26
Middle eyebrow	3	21,22
Right eye	4 (x direction), 5 (y direction)	37,38,40,41
Left eye (x and y direction)	6 (x direction), 7 (y direction)	43,44,46,47
Right inner cheek	8	49,50
Left inner cheek	9	51,52
Right outer cheek	10	49,50,51
Left outer cheek	11	51,52,53
Jaw	12	56,57,58
Right lip corner	13	48
Left lip corner	14	54
Right lower lip	15	57,58
Left lower lip	16	55,56

3 Validation

To ensure our system is robust, we performed two evaluations. First, we validated our method using a simulated face (we used "Alyx", an avatar in the Steam Source SDK [3]). Then, we tested our system on a bespoke robot with 16 DOFs in its face.

3.1 Simulation-based evaluation

We conducted a perceptual experiment in simulation to validate our synthesis module This is a common method for evaluating synthesized facial expressions [9, 25]. Typically, participants observe synthesized expressions and then either answer questions about their quality or generate labels for them. By analyzing collected answers, researchers evaluate different aspects of the expressions of their robot or virtual avatar.

3.1.1 Method

In our perceptual study, we extracted three source videos of pain, anger, and disgust (total of nine videos) from the UNBC-McMaster Pain Archive [24] and MMI database [31], and mapped them to a virtual face. The UNBC-McMaster Pain Archive is a naturalistic database of 200 videos from 25 participants suffering from shoulder pain. The MMI database [31] is a database of images/videos of posed expressions from 19 participants who were instructed by a facial animation expert to express six basic emotions (surprise, fear, happiness, sadness, anger, and disgust). We selected pain, anger, and disgust as these three expressions are commonly conflated, and were replicating the approach taken by Riva et al. [32].

Using our synthesis module, we mapped these nine facial expressions to three virtual characters from the video game Half-Life 2 from Steam Source SDK. We used three different virtual avatars, and overall we created 27 stimuli videos 3 *(Expression: pain, anger, or disgust)* $\times$ 3 *(Gender: androgynous, male, and female)*. Figure 4 shows example frames of the created stimuli videos.

In order to validate people's ability to identify expressions synthesized using our performance-driven synthesis module, we conducted an online study with 50 participants on Amazon MTurk. Participant's ages ranged from 20-57 (mean age = 38.6 years). They were of mixed heritage, and had all lived in the United States for at least 17 years. Participants watched the stimuli videos in randomized order and were asked to label the avatar's expression in each of the 27 videos.

3.1.2 Results and discussion

We found that people were able to identify expressions when expressed by a simulated face using our performance-driven synthesis module (overall accuracy: 67.33%, 64.89%, and 29.56%[2] for pain, anger and disgust respectively) [19, 26]. Riva et al. [32] manually synthesized painful facial expressions on a virtual avatar with the help of facial animation experts, and found 60.4% as the overall pain labeling accuracy rate [32]. Although we did not set out to conduct a specific test to compare our findings to those of manual animation of the same expressions (c.f. Riva et al. [32]), we found our synthesis method achieved arithmetically higher labeling accuracies for pain. These results are encouraging, and suggest that our synthesis module is effective in conveying naturalistic expressions. The next evaluation is to see how well it does on a robot.

[2] Low disgust accuracies are not surprising; it is known to be a poorly distinguishable in the literature [8].

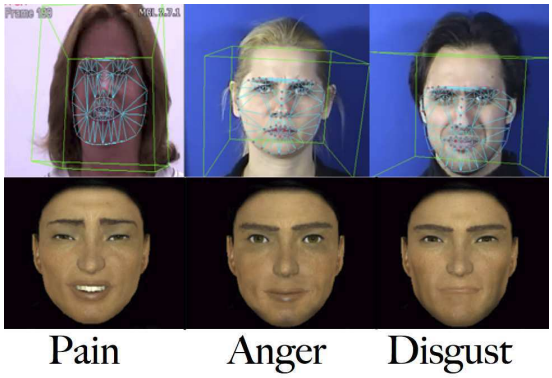


Figure 4. Sample frames from the stimuli videos and their corresponding source videos, with CLM meshes.

3.2 Physical robot evaluation

To test our synthesis method with a physical robot, we used a 16-facial-DOF bespoke robot. Evaluating facial expressions on a physical robot is more challenging than on a simulated face because their physicality changes the physical generation of synthesis. Moving motors in real-time on a robot is far more complex a task due to the number of a robot's motors, their speed, and their range of motion.

We needed to understand if our robot's motors were moving in real time to their intended positions. Since the skin of our robot is still under development, we did not run a complete perceptual study similar to the one we ran in simulation. However, as we were testing how the control points on the robot's head moved in a side-by-side comparison to a person's face, we do not believe this was especially problematic for this evaluation.

3.2.1 Method

We ran a basic perceptual study with 12 participants to test both the real-time nature of the system, and the similarity between expressions of a performer and the robot. We recorded videos of a human performer and a robot mimicking the performer's face. The human performer could not see the robot. However, facial expressions made by the performer were transferred to the robot in real time.

The performer sat in front of a webcam connected to a computer. During the study, the performer was instructed to perform 10 face-section expressions, two times each (yielding a total of 20 videos). The computer instructed the performer to express each of the face-section expressions step by step. Face-section expressions were: *neutral*, *raise eyebrows*, *frown*, *look right*, *look left*, *look up*, *look down*, *raise cheeks*, *open mouth*, *smile*.

We recorded videos of both the performer and the robot mimicking the performer's face. Each video was between 3-5 seconds in length. We ran a basic perceptual study by using side-by-side

Table 2. Full results for each of the 10 face-section expressions.

Face-section expression	Average similarity score	s.d	Average synchrony score	s.d
Neutral	4.12	1.07	4.16	1
Raise eyebrows	4.33	0.86	4.25	1.13
Frown	4	1.02	4.08	1.24
Look right	4.54	0.5	4.66	0.74
Look left	4.5	0.77	4.37	1.08
Look up	2.83	1.29	3.7	1.31
Look down	3.54	1.41	4.45	0.77
Raise cheeks	3.79	1.4	4.25	0.85
Open mouth	4.12	1.16	4.62	0.56
Smile	2.79	1.14	4.41	0.91
Overall	3.85	1.28	4.3	1.01

comparison or "copy synthesis", which we have described in our previous work [27]. In a side-by-side comparison, one shows synthesized expressions on a simulated/physical face side-by-side with the performer's face to participants, and asks them to answer some questions [4, 36].

We showed side-by-side face-section videos of the performer and the robot to participants. Participants viewed the videos in a randomized order. We asked participants to rate the similarity to and synchrony with the performer's expressions and the robot expressions through use of a 5-point Discrete Visual Analogue Scale (DVAS). A five on the scale corresponded to " similar/synchronous" and a one to "not similar/synchronous" .

3.2.2 Results and discussion

Participants were all American and students at our university. Their ages ranged from 20-28 years old (mean age = 22 years). Eight female and four male students participated.

The overall average score for similarity between the robot and the performer expressions was 3.85 (s.d. = 1.28). The overall average score for synchrony between the robot and performer expressions was 4.30 (s.d. = 1.01).

Table 2 reports the full results for each of the 10 face-section expressions. The relatively high overall scores of similarity and synchrony between the performer and the robot expressions suggest that our method can accurately map facial expressions of a performer onto a robot in real-time. However, as this figure shows, we had a low average similarity score for *lookup* and *smile*.

One reason might be that the CLM tracker that we used in our experiment does not accurately track vertical movements of the eyes. Therefore, we could not accurately map the performer's vertical eye movements to the robot. Also, since our robot still does not have skin, its lips do not look very realistic. This might be a reason why participants did not find the robot's lip movements to be similar to the performer's lips movements.

4 General discussion

In this paper, we described a generalized solution for facial expression synthesis on robots, its implementation in ROS using performance-driven synthesis, and its successful evaluation with a perceptual study. Our method can be used both to map facial expressions from live performers to robots and virtual characters, as well as serve as a basis for more advanced animation techniques.

Our work is robust, not limited by or requiring a specific number of degrees of freedom. Using ROS as an abstraction of the code, other researchers may later upgrade the software and increase functionality by adding new nodes to our ROS module.

Our work is also a benefit to the robotics, HRI, and affective agents communities, as it does not require a FACS-trained expert or animator to synthesize facial expressions. This will reduce researchers' costs and save them significant amounts of time. We plan to release our ROS module to these communities within the next few months.

One limitation of our work was that we could not conduct a complete evaluation of our work on a physical robot, since its skin is still under development. Once the robot's skin is completed, we will run a full perceptual test. A second limitation was that the eye-tracking capabilities in CLM are poor, which may have caused the low similarity scores between the robot and performer. In the future as eye tracking technology advances (such as with novel, wearable cameras), we look forward to conducting our evaluation again.

Robots that can convey intentionality through facial expressions are desirable in HRI since these displays can lead to increased trust and more efficient interactions with users. Researchers have explored this domain of research, though in a somewhat fragmented way due to variations in robot platforms that require custom synthesis software. In this paper, we introduced a real-time platform-independent framework for synthesizing facial expressions on both virtual and physical faces. The best of our knowledge, this is the first attempt to develop an open-source generalized performance-driven facial expression synthesis system. We look forward to continuing work in this area.

5 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1253935.

REFERENCES

[1] `github.com/kylemcdonald/FaceTracker`. Accessed: 2014-09-26.

[2] `http://wiki.ros.org/rosserial`. Accessed: 2014-09-26.

[3] Valve Software: Source SDK. `source.valvesoftware.com/sourcesdk.php`.

[4] B. Abboud and F. Davoine. Bilinear factorisation for facial expression analysis and synthesis. In *Proceedings of VISP '05*, pages 327–333, 2005.

[5] H. Admoni, A. Dragan, S. S. Srinivasa, and B. Scassellati. Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In *Proceedings of HRI '14*, pages 49–56. ACM, 2014.

[6] A. Araújo, D. Portugal, M. S. Couceiro, and R. P. Rocha. Integrating arduino-based educational mobile robots in ros. In *Autonomous Robot Systems (Robotica) '13*, pages 1–6. IEEE, 2013.

[7] S. Baron-Cohen. Reading the mind in the face: A cross-cultural and developmental study. *Visual Cognition*, 3(1):39–60, 1996.

[8] D. Bazo, R. Vaidyanathan, A. Lentz, and C. Melhuish. Design and testing of a hybrid expressive face for a humanoid robot. In *IROS '10*, 2010.

[9] C. Becker-Asano and H. Ishiguro. Evaluating facial displays of emotion for the android robot geminoid f. In *WACI '11*, pages 1–8. IEEE, 2011.

[10] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke. Towards a humanoid museum guide robot that interacts with multiple persons. In *Humanoids '05*, pages 418–423. IEEE, 2005.

[11] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *FG '11*, pages 915–920. IEEE, 2011.

[12] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC '06*, volume 2, page 6, 2006.

[13] P. F. Dominey and F. Warneken. The basis of shared intentions in human and robot cognition. *New Ideas in Psychology*, 29(3):260–274, 2011.

[14] C. H. Ek, P. Jaeckel, N. Campbell, N. D. Lawrence, and C. Melhuish. Shared gaussian process latent variable models for handling ambiguous facial expressions. In *American Institute of Physics Conference Series*, volume 1107, pages 147–153, 2009.

[15] P. Ekman. About brows: Emotional and conversational signals. In *Human ethology: Claims and limits of a new discipline: contributions to the Colloquium*, pages 169–248, 1979.

[16] R. El Kaliouby and P. Robinson. Generalization of a vision-based computational model of mind-reading. In *Affective computing and intelligent interaction*, pages 582–589. Springer, 2005.

[17] V. Emeli and H. Christensen. Enhancing the robot service experience through social media. In *RO-MAN '11*, pages 288–295. IEEE, 2011.

[18] F. Eyssel, D. Kuchenbrandt, and S. Bobinger. Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In *Proceedings of HRI '11*, pages 61–68. ACM, 2011.

[19] M. Gonzales, M. Moosaei, and L. Riek. A novel method for synthesizing naturalistic pain on virtual patients. *Simulation in Healthcare*, 2013.

[20] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527, 2011.

[21] H. Knight. Eight lessons learned about non-verbal interactions through robot theater. In *Social Robotics*, pages 42–51. Springer, 2011.

[22] X. Li, B. MacDonald, and C. I. Watson. Expressive facial speech synthesis on a robotic platform. In *IROS '09*, pages 5009–5014. IEEE, 2009.

[23] J. Lieberman and C. Breazeal. Improvements on action parsing and action interpolation for learning through demonstration. In *Humanoids '04*, volume 1, pages 342–365. IEEE, 2004.

[24] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE, 2011.

[25] D. Mazzei, N. Lazzeri, D. Hanson, and D. De Rossi. Hefes: An hybrid engine for facial expressions synthesis to control human-like androids and avatars. In *BioRob '12*. IEEE, 2012.

[26] M. Moosaei, M. J. Gonzales, and L. D. Riek. Naturalistic pain synthesis for virtual patients. In *Fourteenth International Conference on Intelligent Virtual Agents (IVA 2014)*.

[27] M. Moosaei and L. D. Riek. Evaluating facial expression synthesis on robots. In *HRI Workshop on Applications for Emotional Robots at the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013.

[28] M. B. Moussa, Z. Kasap, N. Magnenat-Thalmann, and D. Hanson. Mpeg-4 fap animation applied to humanoid robot head. *Proceeding of Summer School Engage*, 2010.

[29] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita. Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In *Proceedings of HRI '09*, pages 69–76. ACM, 2009.

[30] K. Nagasaka, Y. Kuroki, S. Suzuki, Y. Itoh, and J. Yamaguchi. Integrated motion control for walking, jumping and running on a small bipedal entertainment robot. In *ICRA'04*, volume 4, pages 3189–3194. IEEE, 2004.

[31] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *ICME '05*, 2005.

[32] P. Riva, S. Sacchi, L. Montali, and A. Frigerio. Gender effects in pain detection: Speed and accuracy in decoding female and male pain expressions. *EUR J PAIN*, 2011.

[33] J. M. Saragih, S. Lucey, and J. F. Cohn. Face alignment through subspace constrained mean-shifts. In *ICCV '09*. IEEE, 2009.

[34] S. Sosnowski, A. Bittermann, K. Kuhnlenz, and M. Buss. Design and evaluation of emotion-display eddie. In *IROS '06*, pages 3113–3118. IEEE, 2006.

[35] L. Takayama, D. Dooley, and W. Ju. Expressing thought: improving robot readability with animation principles. In *Proceedings of HRI '11*, pages 69–76. ACM, 2011.

[36] Q. Zhang, Z. Liu, B. Guo, and H. Shum. Geometry-driven photorealistic facial expression synthesis. In *Proceedings of SCA '03*, pages 177–186, 2003.

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Gender, more so than Age, Modulates Positive Perceptions of Language-Based Human-Robot Interactions

Megan Strait and **Priscilla Briggs** and **Matthias Scheutz**[1]

Abstract. Prior work has shown that a robot which uses politeness modifiers in its speech is perceived more favorably by human interactants, as compared to a robot using more direct instructions. However, the findings to-date have been based soley on data aquired from the standard university pool, which may introduce biases into the results. Moreover, the work does not take into account the potential modulatory effects of a person's age and gender, despite the influence these factors exert on perceptions of both natural language interactions and social robots. Via a set of two experimental studies, the present work thus explores how prior findings translate, given a more diverse subject population recruited via Amazon's Mechanical Turk. The results indicate that previous implications regarding a robot's politeness hold even with the broader sampling. Further, they reveal several gender-based effects that warrant further attention.

Figure 1: Scenario: the humanoid MDS robot (Xitome Designs; left) instructs a confederate participant (right) on a brief drawing task.

1 INTRODUCTION

Natural language interactions with virtual and robotic agents are becoming increasingly pervasive, from virtual personal assistants (such as Apple's Siri agent), to socially assistive robots (e.g., elder care robots such as [4]). As the functionality of these artificial agents grows, so does the need to communicate with humans effectively to best serve the human interlocutor [12]. Surprisingly, however, there are very few attempts to date to carefully evaluate the different ways in which artificial agents could talk with humans in the context of a given task based on the agent's physical embodiment. For example, it is unclear whether an artifical agent, depending on its embodiment, should use imperatives when instructing humans (e.g., "turn right at the next intersection") or whether a more polite way of expressing an instruction is required (e.g., "we need to turn right at the next intersection"). Intuitively, a non-embodied agent like a navigation system might get away with syntactically simple, effecient imperatives, while a humanlike embodied robotic agent might have to employ more conventional forms of politeness.

Past work evaluating politeness in natural language interactions with robotic agents supports this intuition. Torrey and colleagues, for example, showed that the use of hedges (e.g., "*I guess*", "*probably*", and "*sort of*") and discourse markers – two "negative" politeness techniques – improves how people perceive a robot instructing a person via natural language. Specifically, they found that polite robots were viewed more positively than robots using more direct speech [22]. Even though negative politeness may be less noticeable than the *please*s of positive politeness, hedging indicates to the listener that the speaker is trying to mitigate the force of the request [7, 14].

[1] Human-Robot Interaction Laboratory, Tufts University, Medford, MA USA

Recent extensions of the above findings show that other negative politeness techniques (e.g., phrasing requests indirectly [9]), as well as positive (e.g., inclusive pronouns), suffice to improve perceptions of human-robot interactions (e.g., [6, 19, 21]). However, this research investigating human perceptions of robot politeness in human-robot interactions ([21, 22]) is predominately based on data drawn from the standard (and relatively homogeneous) university population.

Thus, whether and how these findings transfer to scenarios involving a population that is more diverse (e.g., economically, educationally), remains unknown. In particular, there are several factors (sociolinguistic, cultural, and demographic) in addition to politeness that have been found to modulate perceptions of natural language interactions (e.g., [3, 13, 16, 17, 20]). For instance, contrary to popular stereotypes, Japan is not as robot-positive as the US [2, 8].

Of particular relevance, is the growing amount of evidence that men (relative to women) hold significantly more positive towards robotic entities [5]. While both Torrey et al. ([22]) and Strait et al. ([21]) attempted to control for unintended effects due to gender, their participant samples were nevertheless imbalanced and thereby constrained in their ability to represent the general population. Hence, it is important to revisit these findings with explicit consideration of socio-demographic factors to understand what are their specific influences and how the findings extend beyond the university.

The goal of the present work was thus two-fold: (1) to investigate whether an extension of [21] with more diverse subject demographics would replicate the previously-observed effects of robot politeness (based on interaction observation), and further, (2) how the subject-based factors of age and gender specifically interact with those of the robot (e.g., the robot's use of polite communicatory cues).

To address these questions, we conducted a set of two online experiments via Amazon's Mechanical Turk with the aim of achieving greater diversity in people's age, and educational/geographical backgrounds, as well as more balanced gender demographics. In both, we presented videos depicting a robot instructing a person on a simple drawing task. We solicited people's reactions to these videos to determine the influence of a robot's politeness relative to any modulatory effects of a person's age and gender (**Experiment I**). Owing to a limitation of the first study, we conducted a follow-up to Experiment I to determine whether the findings hold given more naturalistic interaction settings (**Experiment II**).

2 EXPERIMENT I

Based on the previous work outlined in the introduction ([21, 22]), we hypothesized that by using politeness modifiers in its speech, a robot would be perceived more favorably (as evidenced by higher ratings of *likeability* and reduced ratings of *aggression*) than a robot that uses more direct instructions. In addition, we generally explored the modulatory effects of a person's socio-demographic factors – in particular, age and gender – and how they interact with characteristics of a robot to influence perceptions of human-robot interactions.

To test our hypotheses and the age- and/or gender-based modulations thereof, we conducted a fully between-subjects investigation of the effects of a robot's communication strategy on observations of brief human-robot interactions – as influenced by a person's age and gender. In order to obtain a more diverse population than previously, we conducted our investigation online via Amazon's Mechanical Turk. Using a modification of the materials and methods developed in [21], we tasked participants with viewing a short video depicting a robot as it advised a person on creating a simple drawing. Following the video viewing, participants were prompted for their perceptions of the interaction, as rated on several dimensions regarding the likeability and aggression of the robot.

2.1 Materials & Methods

2.1.1 Participants & Procedure

839 participants were recruited via Amazon Mechanical Turk.[2] Prior to participating, subjects were informed the purpose of the study was to investigate factors that influence perceptions of human-robot interactions. Upon informed consent and subsequent completion of a demographic survey, the subject was shown one of 32 videos depicting a robot instructing a human confederate on a simple task. Following the viewing, the subject completed a 12-item questionnaire regarding his/her perceptions of the robot's appearance and behavior. Lastly, to assess attentiveness, participants completed a three-item check regarding salient details of the video clip.

Of these 839 participants, data from 329 were discarded due to several exclusion criteria: a restriction to limit participation to native english speakers (51 participants), and failure to complete the requested tasks (70) or failure on a three-item attention check (with a success threshold of 100%) to ensure participants viewed the presented video (208). Thus, our final sample included data from 510 participants (62% male) from 47 of 50 US states. The average age of this sample was 31.21 (SD=9.71), ranging from 18 to 76 years old. The most common level of education obtained was a bachelor's degree (45%), with an additional 36% of participants having some amount of college-level education. A small percent of participants reported having completed only high school (12%) and a smaller proportion reported obtaining more advanced degrees (7%). Participants also reported relatively high interest in robots (M=5.15, SD=1.32) – though low familiarity with robots (M=3.75, SD=1.49) – based on a 7-point Likert scale with 1=low and 7=$high$.

[2] In anticipation of some loss in data due to exclusion criteria, we chose this sample size to achieve ≥ 15 useable observations in hypothesis testing.

	Comforting	*Considerate*	*Controlling*
Aggressive	−.15	−.11	**.68**
Annoying	**-.62**	−.27	.21
Comforting	**.73**	.30	−.13
Considerate	.21	**.63**	−.15
Controlling	−.11	−.16	**.52**
Eerie	**-.73**		.16
Likable	**.60**	**.59**	
Warm	.22	**.77**	−.24
Eigenvalues	3.63	1.16	.99
Variance Explained	.24	.44	.56

Table 1: Factor loadings for the three-factor EFA solution.

2.1.2 Independent Variables

We employed a $2 \times 3 \times 2$ factorial design in which we systematically manipulated a robot's *politeness* in an advice-giving scenario, using the same conditions as those developed by Strait and colleagues ([21]). We also included participant *age* (three levels) and *gender* to investigate how they affect perceptions of the human-robot interaction. In total, we had the following three independent variables (IVs):

- **Politeness** of the robot's instructions (direct vs. polite). The *polite* condition entailed the robot giving instructions that contained one or more of both positive and negative politeness strategies, such as praise (e.g., "great job") and hedges (e.g., "a *kind of* large circle"). The *direct* speech condition employed the exact same instructions, but with the politeness modifiers removed.
- **Participant age** (three levels). We established three age categories based on a $1/3$ split of all the self-reported ages, resulting in a corresponding to the age of the standard university sample (M_1=22.81 years, SD=1.87), as well as two older adult categories (M_2=28.68, SD=1.99; M_3=42.16, SD=8.86).
- **Participant gender** (female vs. male).

2.1.3 Covariates

In addition to the above, we planned to carefully control for potential effects due to a person's motivations for completing the tasks (i.e., due to his/her purported *interest* in robots), as well as any effects due to characteristics of the stimulus set. To do so, we covaried three factors pertaining to the robot's physical embodiment:

- **Appearance** of the robot (two levels): the humanoid MDS (Xitome Designs) versus the less humanlike PR2 (Willow Garage).
- **Production modality** (synthetic vs. human speech), and
- **Gender** (female vs. male) of the robot's voice.

Thus, a total of four covariates – participants' **interest** in robots, the robot's **appearance** and the **gender** and **production modality** of the robot's voice, – were used in the analyses reported below.

2.1.4 Stimuli

A set of 32 videos (two conditions – polite versus direct speech – with 16 instances per condition) were constructed based on systematic manipulation of the robot-based IVs and covariates. Each video depicted a variant of a robot instructing a male human actor on a pen-and-paper drawing of a koala (cf. [21]). To avoid potential effects of affect, behavior, and/or movement (due to differences between the two robots' abilities), the robots were kept stationary. To avoid unintended effects due to a particular appearance, gender, voice, or the way in which the voice was produced, 16 video instances co-varying the robot's humanoid appearance (MDS versus the PR2), voice production modality (synthetic- versus human-produced speech) and voice gender (four voices – two female, two male) were created per condition. Four adult human actors comprised the set of human voices, with instructions to perform with flat affect. Synthetic voice production was performed using the native Mac OS X text-to-speech (TTS) software with four voices: "Alex","Ava", "Tom", and "Vicki". Following a between-subjects design, participants viewed only one video (selected randomly from the set of 32).

2.1.5 Dependent Variables

Of the set of 12 questionnaire items, three items – *task difficulty*, *interaction difficulty*, and *interest in interacting* – were considered as unique variables. On the remaining 9 items drawn from prior work (cf. [21, 22]), exploratory factor analysis produced a three-factor solution which showed a better fit ($\chi^2(7) = 13.36$, $p = .0638$) than a model where the variables correlate freely.

The criterion for retention of a questionnaire item was a factor loading of $\geq .50$ (see Table 1). We thus interpreted the three latent variables as the following: how **comforting** (four items – comforting, likable, -annoying, and -eerie; Cronbach's α=.83), **considerate** (three items – considerate, likable, and warm; α=.79), and **controlling** (two items – aggressive and controlling; α=.55) the robot was perceived. Items that were negatively correlated are indicated by $-$, and were automatically reversed in the computation of the latent constructs. Further, all dependent measures were normalized (to a scale between 0 and 1) prior to analysis.

2.2 Results

To assess the effects of the three IVs, between-subjects ANCOVAs were conducted on each of the dependent variables (taking into account the four covariates), with homogeneity of variance confirmed using Levene's test. All significant effects are reported below (with significance denoting $\alpha \leq .05$), and all post-hoc tests reflect a Bonferroni-Holm correction for multiple comparisons.

2.2.1 Comforting, Considerate, & Controlling

As expected, the *politeness* manipulation showed marginal ($p<.10$) to significant main effects on all three latent factors – *comforting*, *considerate*, and *controlling* (see Table 2, top). Similarly, participants' *gender* did as well (see Table 2, bottom); however, there were no significant main or interaction effects due to the participants' *age*.

Overall, both politeness and gender tended to increase ratings of the robot as *comforting* and *considerate*, and conversely, decrease those for *controlling*. However, these main effects were eclipsed by a *politeness* $\times$ *gender* interaction on both of the two positive factors: *comforting* ($F(1, 498)$=4.57, p=.03, η^2=.01) and *considerate* ($F(1, 498)$=6.97, $p<.01$, η^2=.01).

	DIRECT ($n = 254$)	**POLITE** ($n = 256$)	$F(1, 498)$	p	η^2
Comforting	.13 (.37)	.19 (.38)	3.26	$= .07$	.01
Considerate	.46 (.16)	.54 (.17)	31.82	$< .01$	**.06**
Controlling	.25 (.17)	.20 (.16)	10.29	$< .01$	.02
	FEMALE ($n = 193$)	**MALE** ($n = 317$)	$F(1, 498)$	p	η^2
Comforting	.21 (.40)	.11 (.36)	9.27	$< .01$	.02
Considerate	.53 (.16)	.48 (.16)	13.42	$< .01$	.03
Controlling	.20 (.16)	.26 (.17)	14.44	$< .01$	.03
Difficulty (t)	.17 (.16)	.21 (.18)	8.20	$< .01$	.02
Difficulty (i)	.24 (.23)	.28 (.21)	5.18	$= .02$	.01
Interest	.48 (.23)	.43 (.21)	4.74	$= .01$	.01

Table 2: Main effects of *politeness* (top) and *gender* (bottom), and relevant descriptive and inferential statistics.

In particular, the interactions showed that – while polite speech tended to improve participants' ratings – it did so primarily for women (see Figure 2, left and center). That is, a robot's use of polite speech significantly improved ratings of *comfort* when viewed by female observers (M=.29, SD=.39, n=94) relative to those by female observers of direct speech (M=.14, SD=.40, n=99; p=.04) and male observers of both direct (M=.11, SD=.34, n=155; p=.01) and polite speech (M=.11, SD=.38, n=162; p=.01). Similarly, though the polite robot significantly improved observers' ratings of *considerateness* for both female (M_{polite}=.59, SD=.16; M_{direct}=.47, SD=.17; $p<.01$) and male observers (M_{polite}=.50, SD=.17; M_{direct}=.45, SD=.15; p=.02), women's ratings were most improved relative to men's ($p<.01$).

With regard to perceptions of the robot as *controlling*, politeness was still broadly effective at decreasing ratings – regardless of the observer's gender, with polite robots receiving lower ratings relative to those more direct in their instructions (see Table 2, top). But, just being female helped as well: with women rating the robot as substantially less controlling than did men (see Table 2, bottom).

2.2.2 Difficulty & Interest

Gender further exerted significant main effects on the dependent variables regarding the perceived *difficulty* of both the task and interaction, as well as the observers' own *interest* in interacting with the depicted robot (see Table 2, bottom). In particular, female participants tended to rate both the task and interaction as less difficult than did males (see Table 2–bottom, *Difficulty*). Furthermore, they tended to show more interest in interacting with the robot agent than their male counterparts (see Table 2–bottom, *Interest*). There were no significant effects (main or interaction) due to *politeness* or *age*.

2.3 Discussion

Do people perceive a robot, which employs politeness modifiers in its speech, more favorably than one that uses more direct speech? Based on previous research by [21, 22], we expected that participants would rate a polite robot more favorably than one that is more direct in its instructions, as evidenced by higher ratings of positive constructs (e.g., *likability*) and lower ratings of negative constructs (e.g., *aggression*). Consistent with that work, the politeness manipulation here showed lower ratings of the robot as *controlling* and higher ratings of the robot as being *considerate* and *comforting*. In particular, our results

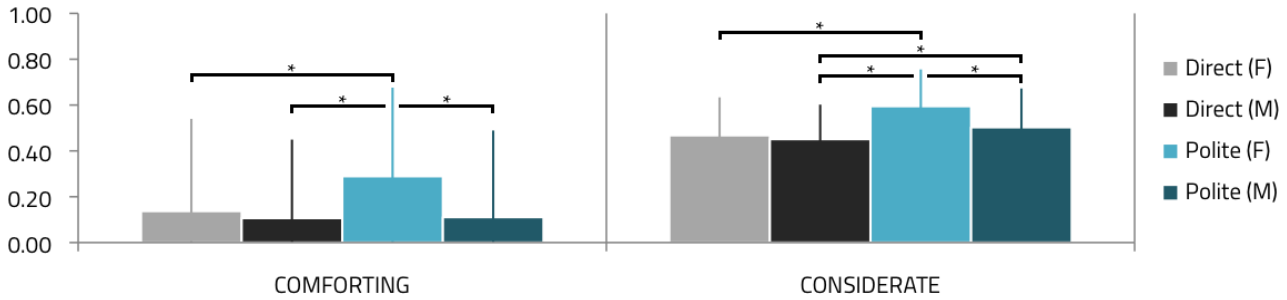


Figure 2: Interaction between robot *politeness* and participant *gender* on the three latent factors – the degree to which the robot was perceived as *comforting*, *considerate*, and *controlling*. Gray bars indicate the use of *direct* speech, versus blue, which indicates *polite* speech. Lighter bars indicate female participants (versus male participants, darker bars). All significant contrasts are shown (indicated by asterisks).

replicate and confirm those of prior work, even with a substantially more diverse subject population.

Does a person's age and gender further modulate perceptions of human-robot interactions? Based on previous suggestions that men and women view and respond to robots in significantly different ways [5], we evaluated the primary and modulatory effects of participants' age and gender. Participants' *gender* exerted a main effect on *all* dependent measures: how *comforting*, *considerate*, and *controlling* the robot was perceived as being, as well as how *difficult* both the task and interaction seemed and participants' *interest* in interacting with the depicted robot. In particular, female participants (relative to their male counterparts) showed more positive responding towards the robots and their interactions with the human confederate, as reflected by increased ratings of interest, comfort, and the robot's considerateness, as well as decreased ratings of the task/interaction difficulty and the robot's aggression. Further, interactions with the politeness manipulation showed that a robot's use of polite speech was effective at increasing women's positive attributions (the robot as being *comforting* and *considerate*), but not men's. Participant *age*, however, showed no main or interaction effects on *any* of the measures.

2.3.1 Implications

Prior work has suggested that a robot's use of politeness modifiers in its speech improves perceptions of human-robot interactions in advice-giving situations [21, 22]. Our results further replicate these findings (with respect to observation of human-robot interactions), and moreover, show the influence of politeness holds given a more general and representative population sample. In particular, our participants came from a wide variety of educational backgrounds (ranging from high school to advanced degrees) and geographical locations within the US (47 states).

In addition, we explicitely considered the effects of a person's age (ranging from the standard university age level to older adult) and their gender, to determine their influence and nature relative to the robot's politeness. This consideration of such socio-demographic factors revealed a methodological consideration for HRI studies – namely, that a person's *gender* should be taken into account when assessing perceptions of language-based human-robot interactions, as it is a modulating influence in addition to a robot's *politeness*.

This was expected, as previous research (e.g., [15, 18, 20]) has found that men exhibit more positivity towards robots than women. But, contrary to prior observations, our results indicate that women respond, in general, more positively towards the depicited robots. This may be due to the difference in the presentation the interactions as, here, video-recordings of human-robot interactions were evaluated by post-hoc observation, whereas, previous work has used scenarios involving the participatory and co-located interaction between the participant and robot of interest [16, 17, 20]. Alternatively (or in addition), it may be due to the difference in interaction: here, the robot interactants were depicted as instructing a human confederate; whereas, the human interactants in prior work were tasked with instructing or working with (rather than subservient to) the robot agent. Despite the conflicting differences in the nature of their effects, our findings add to the growing body of evidence implicating gender as an important methodological consideration in evaluating perceptions of human-robot interactions.

2.3.2 Limitations & Future Directions

Our approach to the investigation of perceptions of polite robots contributes a simple online task to assess the modulatory influences (or lack thereof) of a person's age and gender. In particular, the collection of data with broad socio-demographics augments in-laboratory studies that are limited to small, and relatively homogeneous, participant populations. This contribution here is significant because it replicates the previously reported influences of politeness, and further, sheds light on how such findings might transfer to the general population. That said, our approach also has several limitations (which underscore avenues for further research), three of which we discuss below.

Relevance. First, we note that the effect sizes for the given manipulations are relatively small. The magnitude of the effect of politeness on perceptions of the robot's considerate approaches a medium qualification (η=.10), but nevertheless, the implications of both robot politeness and participant gender are of limited weight. This may also suggest it is worth looking at the specific effects due to other factors such as a person's educational or geographic background (two socio-demographic items for which we did not control).

Mode of Evaluation. Another limitation of relevant consideration is how peoples' evaluations of the interactions were obtained. Here, the interactions were evaluated post-hoc by a third-party observer, who (by definition) was remotely located from the actual robot/interaction. This is particularly important to note, as it has been found that perceptions of human-robot interactions are further modulated by the interaction distance (remote versus co-located) and nature (observatory versus participatory) [21]. Thus, while the video-based interactions and online evaluations allowed us to sample from a broader demographic than that which is available locally, whether and how our gender-based findings apply to actual, co-located human-robot interactions warrants further investigation.

Stimuli. Lastly, there are a number of important limitations to the stimuli used and their presentation. Here the stimuli depicted brief (2 minute) interactions between an inanimate robot and a human confederate, which is an unrealistic interaction scenario in comparison to the intended usage of social robots.

In particular, prior work has shown that movement (however subtle) can impact the efficacy of interactions. For example, Andrist and colleagues have found that averting a robot's gaze (even for robot's without articulated eyes) can improve perceptions of the robot and their interactions [1]. Thus, with regard to the present study – though we limited movement to avoid unintended and/or differential influences (e.g., due to the robots' different capacities for actuation), the absence of movement itself might be affecting the current findings in unknown ways. For instance, the absence of attention-indicating gaze (e.g., looking at the participant when he/she is not performing a drawing instruction) might reduce positive attributions (e.g., considerateness) and/or increase negative attributions. This idea is supported by participants' open responses, which generally showed negative attitudes regarding the robot's lack of movement. Thus, there is the distict possibility that the lack of movement influenced perceptions in some way that may attenuate (or worse, decimate) other influences (e.g., due to politeness). With such considerations in mind, we moved to conduct a follow-up experiment to test the nature and magnitude of effects due to politeness and gender, when the robot was animated in a more naturalistic fashion.

3 EXPERIMENT II

Based on the considerations outlined in the previous section, we composed an exploratory follow-up investigation to Experiment I. We again conducted a between-subjects investigation of the effects of a robot's politeness (as influenced by a person's gender) on perceptions of human-robot interactions – but, with more naturalistic interactions. Specifically, we constructed a second set of video stimuli in which the robot was *animated* with attention-sharing and (human-like) idling movements, based on the naturalistic movements exhibited by a human instructing in such a context.

3.1 Materials & Methods

3.1.1 Participants & Procedure

437 additional participants were recruited via Amazon Mechanical Turk.[3] As in Experiment I, participants were told the purpose of the study was to investigate factors that influence perceptions of human-robot interactions. Upon informed consent and completion of a demographic questionnaire, the subject was shown one of 16 videos (similarly depicting a robot instructing a human confederate on a simple task). Following the viewing, the subject completed the 12-item questionnaire regarding his/her perceptions of the robot's appearance and behavior and the three-item check to assess whether the participant attended to the video.

Of these 437 participants, data from 176 participants were discarded due to: failure to complete the requested tasks (54) or failure on the attention check (122). Thus, our final sample included data from 261 participants (60% male) from 48 of the 50 US states. The average age of this sample was 32.45 (SD=10.45), ranging from 18 to 68 years old. The most common level of education obtained was similarly a bachelor's degree (44%), with an additional 37% of participants having some amount of college-level education. As in Experiment I, a small percent of participants reported having completed only high school (13%) and a smaller proportion reported obtaining more advanced degrees (6%). Participants again reported low familiarity (M=3.79, SD=1.49) with, but relatively high interest (M=5.33, SD=1.39) in robots.

3.1.2 Independent Variables

We again employed a fully factorial design, with the same independent variables as previously:

- The robot's **politeness** (direct vs. polite).
- **Participant age** (three levels): the standard university sample (M_1=22.85 years, SD=2.18), as well as two older adult categories (M_2=29.70, SD=2.01; M_3=43.98, SD=8.65).
- The participant's **gender** (female vs. male).

3.1.3 Covariates

We again planned to control for effects due to a person's *interest* in robots, as well as any due to characteristics of the stimulus set. As there was little variance explained by *production modality*, we excluded it from consideration to help reduce the overall number of videos to remake, thus reducing the number of observations needed to achieve similar sample sizes as Experiment I. As a result, we considered a total of three covariates in our analyses here: two factors pertaining to the robot's physical embodiment (the robot's *appearance* – MDS vs. PR2 – and *gender* of the robot's voice) and one factors pertaining to the participant (their *interest* in robots).

3.1.4 Stimuli

To increase the degree of observable presence/embodiment of the depicited robots, we recreated the videos from Experiment I[4] to animate the robots with select movements during the interaction. The movement modifications were intended to create a sense of "shared attention" and "idle" behaviors, based on the behaviors observed of a human instructor during pretesting of the drawing task with two people. In particular, the attentive behaviors were implemented such that the robot (MDS or PR2) moved its eyes (MDS) or head (PR2) up/down to focus on the human actor when giving instructions or on the actor's drawing (when the actor was drawing). Each robot also performed a set of idle behaviors (initiated based on random timers) throughout the interaction, based on their relative capacities for movement:

- *Blinking* (MDS only) – the MDS robot has two actuated eyelids that were closed and reopened ($500ms$) mimic human blinking.
- *Swaying* (MDS only) – the MDS has three degrees of freedom (DOF) on its center axis, allowing mimicry of slight head tilts (left/right and up/down positioning determined randomly at initiation of each tilt).
- *Breathing* (PR2 only) – the PR2, having fewer DOF with respect to its head movement, was limited to regular up/down undulation of its frontal laser. The rate of the laser movement approximated the average person's resting state heart rate ($70bpm$).

[3] In anticipation of data loss due to our exclusion criteria, we chose this sample size to again achieve ≥ 15 useable observations in hypothesis testing.

[4] As *production modality* was dropped from consideration, we recreated only a subset of the E1 videos – the 16 depicting a robot with a synthetic voice.

	DIRECT ($n = 130$)	POLITE ($n = 131$)	$F(1, 249)$	p	η^2
Comforting	.59 (.20)	.66 (.20)	6.72	$= .01$	.03
Considerate	.59 (.18)	.70 (.17)	26.27	$< .01$	**.11**
Controlling	.24 (.19)	.19 (.15)	6.31	$= .01$	.03

	FEMALE ($n = 104$)	MALE ($n = 157$)	$F(1, 249)$	p	η^2
Comforting	.68 (.20)	.58 (.20)	15.03	$< .01$	**.06**
Considerate	.68 (.18)	.62 (.18)	8.67	$< .01$	.03
Controlling	.20 (.16)	.24 (.17)	4.22	$= .04$	.02
Difficulty (t)	.24 (.23)	.29 (.23)	3.55	$= .06$	.01
Difficulty (i)	.26 (.28)	.35 (.27)	6.40	$= .01$	.03
Interest	.68 (.28)	.61 (.27)	3.45	$= .06$	.01

Table 3: Main effects of *politeness* (top) and *gender* (bottom), and relevant descriptive statistics, in Experiment II.

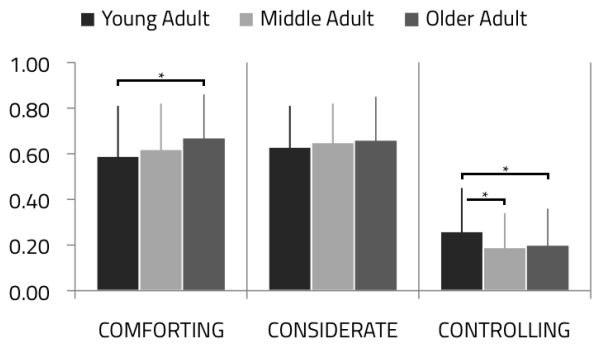


Figure 3: Main effects of participant *age*. Asterisks indicate significant contrasts.

3.1.5 Dependent Variables

We used the same dependent measures as previously: task and interaction *difficulty* and *interest in interacting*, as well as how **comforting**, **considerate**, and **controlling** the robot was perceived as being.

3.2 Results

To assess the effects of robot *politeness* and participant *age*/*gender* – in the context of more naturalistic interactions – between-subjects ANCOVAs were conducted on each of the dependent variables (taking into account the four covariates), with homogeneity of variance confirmed using Levene's test. All significant effects are reported below (with significance denoting $\alpha \leq .05$), and all post-hoc tests reflect a Bonferroni-Holm correction for multiple comparisons.

3.2.1 Robot Politeness

As previously found, *politeness* exerted a significant effect on all three of *comforting*, *considerate*, and *controlling* DVs. Specifically, as expected based on Experiment I and previous literature, the robot's use of polite speech increased participants' comfort and their perceptions of the robot's considerateness. It also reduced perceptions of the robot as controlling (see Table 3, top).

3.2.2 Participant Age & Gender

Similarly, as Experiment I showed, *gender* improved perceptions along all dependent measures (see Table 3, bottom). Specifically, female participants continued here to (1) rate the robot as more *considerate* and less *controlling*, (2) indicate greater *comfort* and *interest in interacting* with the depicted robot, and (3) rate both the interaction and task as less difficult, than did their male counterparts.

Unlike the previous experiment, however, here participant *age* also showed a significant influence on *comfort* with the robot ($F(2, 249) = 3.19$, $p = .04$, $\eta^2 = .03$) and perception of it as *controlling* ($F(2, 249) = 4.07$, $p = .01$, $\eta^2 = .03$). Specifically, participants of the standard university age (young adults) indicated significantly less comfort with the robot ($M = .59$, $SD = .22$, $n = 87$) than the oldest participants ($M = .67$, $SD = .19$, $n = 92$; $p < .01$). Conversely, the younger participants also rated the robot as significantly more controlling ($M = .26$, $SD = .19$, $n = 87$; $p = .01$) than did either of the two older age groups – middle adults ($M = .19$, $SD = .16$, $n = 82$) and older adults ($M = .20$, $SD = .16$, $n = 92$).

3.2.3 Interactions

Furthermore at odds with Experiment I (where the $gender \times politeness$ eclipsed many of the main effects of politeness), there were no significant or even marginally significant interaction effects here. Specifically, in the context of the more naturalistic interactions, the use of polite speech seemed to be effective for both female *and* male participants. This suggests that, while female participants appear to be particularly sensitive to verbal communication (as evidenced by their ratings across both the more naturalistic Experiment II *and* Experiment I), male participants may be more sensitive to *consistency* in verbal and nonverbal communicatory cues.

3.3 Discussion

3.3.1 Summary of Present Findings & Implications

In this follow-up investigation, we explored whether our previous findings in Experiment I – that a robot's use of polite speech improves perceptions (and, that women respond more positively towards such robots) – hold given more naturalistic interaction scenarios (i.e., human-robot interactions in which the robot is animated).

Here we observed that the results, for the most part, reflect those of the previous experiment (despite E1's lack of movement in the shown video interactions). Specifically, the politeness manipulation again resulted in lower ratings of the robot as *controlling* and higher ratings of the robot as being *considerate* and *comforting* (see Figure 4, left). This lends further support of politeness as an effective tool for facilitating more positive responding towards robots (at least for natural language interactions in advice-giving scenarios).

Similarly, participants' *gender* again exerted a main effect on *all* dependent measures: how *comforting*, *considerate*, and *controlling* the robot was perceived as being (see Figure 4, right), as well as how *difficult* both the task and interaction seemed and participants' *interest* in interacting with the depicted robot. In particular, women rated the robots more positively than did male participants as was observed in Experiment I. While this remains in contradition with prior work showing that men respond more positively towards robots than women (e.g., [15, 18, 20]), it nevertheless lends further support towards the methodological implication that gender is a relevant consideration for HRI studies.

Moreover, the results of the present investigation indicate that observatory perspectives of human-robot interactions are not substantially influenced by the robot's animacy. This suggests that simplistic

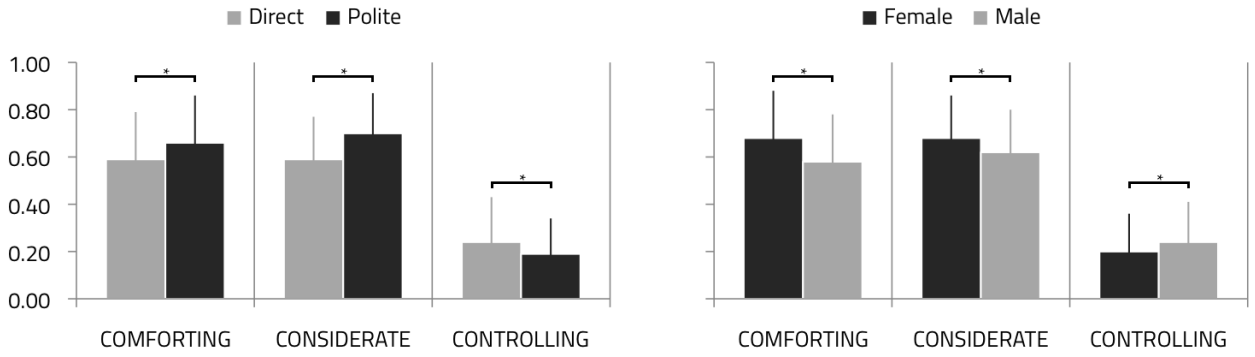


Figure 4: Main effects of *politeness* (left) and participant *gender* (right) on perceptions of the robot as *comforting*, *considerate*, and *controlling*. Dark bars emphasize factors yielding more positive outcomes (polite speech, female participants). All contrasts are significant.

depictions of human-robot interactions, such as in Experiment I, may suffice to investigate perceptions of certain robot behaviors (e.g., a robot's politeness, as perceived by female observers).

However, key differences in findings between the two experiments also underscore the necessity of considering perceptions in more realistic interaction scenarios. Specifically, unlike in Experiment I, Experiment II showed no interactions between any of the three IVs. For example, in the context of the more naturalistic interactions, the use of polite speech was effective at improving ratings regardless of the participant's gender. Whereas, in Experiment I, polite speech was only effective at improving *female participants'* ratings (while male participants of Experiment I were not receptive – the use of politeness modifiers, in the absence of the idling and attention sharing movements, did not improve ratings). This suggests that, while women appear to be particularly sensitive to verbal communication (as evidenced by their ratings across both Experiment I and the more naturalistic Experiment II), men may be more sensitive to *consistency* in verbal and nonverbal communicatory cues. Thus, the findings may imply a need for coherence between a robot's verbal and nonverbal communication (e.g., [10]).

In addition, the present experiment showed a slight influence of age on perceptions of *comfort* with the robot and how *controlling* it seemed (see Figure 3), whereas E1 showed no significant effects owing to participants' age. These effects are somewhat difficult to interpret, however, as it is unclear what aspects of the more realistic interaction would cause the standard university-aged participants (relative to the older adults) to here indicate less comfort with the robot and rate it as more controlling.

3.3.2 Limitations & Future Directions

Here we undertook further investigation of perceptions of robot politeness and potential modulatory factors. Our approach tested a few simple behaviors to assess the influence (or lack thereof) of a robot's movement. In particular, the presentation of human-robot interactions that were more naturalistic (i.e., mimic attention-sharing and idling behaviors exhibited in equivalent human-human interactions) compliments our previous study, which lacked the same degree of social realism. This contribution here is significant because it replicates the influences of politeness of both prior work and our own Experiment I. Further, it sheds light on how subject-based factors (i.e., age and gender) can yield more positive social evaluations. However, as with the previous study, our approach still has its limitations.

In particular, we explored here only a small subset of human-inspired movements. Thus, it is not possible to conclusively say that movement (of any kind) is effective for improving interactions or perceptions thereof. There are substantially more possibilities to try, such as gaze aversion (e.g., [1]) or gesturing (e.g., [11]) to name a few. To determine what extent certain types nonverbal communicatory mechanisms influence perceptions, future work might consider independently manipulating several types of movements, rather than the movement/no-movement meta comparison we made here.

4 GENERAL DISCUSSION

4.1 General Findings & Implications

As expected, Experiment I confirms prior indications that, at least in 3rd-person observation of pre-recorded human-robot interactions ([21, 22]), a robot's use of politeness modifiers in its speech is perceived more favorably relative to a robot that uses more direct speech (e.g., [14, 19, 21, 22]). This is reflected by participants ratings of the polite robot instructors as more *comforting* and *considerate*, and less *controlling* than the robots that were more direct. Moreover, the implications of politeness hold, even for a population that is highly diverse in terms of the socio-demographic factors of education, geographical location, age, and gender. Furthermore, we observed additional validation of the effects owing to a robot's politeness in Experiment II. Thus, consistent with prior indications ([21, 22]), the persistence of effects due to politeness – given the broader population sampling – demonstrate the benefit to using politeness modifiers when a robot communicates with natural language.

The results observed across the two studies further underscore an important methodological consideration – namely, gender – for evaluation of human-robot interactions. Specifically, we found a gender-based divide in the efficacy of the politeness manipulation in both experiments showing that a robot's use of politeness modifiers in its speech is most (and in Experiment I, only) effective for female participants. That is, here women rated polite robots significantly better than those that are more direct, and moreover, their ratings of polite robots are significantly higher than men's ratings of the same robots. Furthermore, the two studies suggest that men are sensitive to consistency in communicatory cues, and more importantly, they are not receptive to polite speech alone. These findings demonstrate the importance of considering gender – either as a systematic manipulation or as a covariate – in the analysis of human-robot interactions.

4.2 General Limitations & Future Work

Our approach to understanding perceptions of polite robots contributes a simple online task to assess the modulatory influences of various situational factors. We emphasize the benefit that the online forum serves for obtaining data with broad socio-demographics versus in-laboratory studies which are limited to smaller and more homogeneous participant populations. This lends the ability towards replicating previously indicated influences of politeness and understanding how such findings might transfer to the general population.

However, we wish to also underscore the limitations of this type of assessment. Despite the benefits to online studies, the results cannot be immediately applied to actual human-robot interactions involving co-located, direct participation, as the present work was conducted from a remote and observatory position (relative to the depicted interactions). Hence, whether (and if so, the extent to which) these findings generalize and apply to in-person, direct interactions with a co-located embodied agent motivates further investigation.

Further, we stress that these findings are preliminary and of limited weight. In particular, we note the small effect sizes observed across both studies. Between the two experiments, the effect sizes reached at most a medium qualification with the influence of politeness on perceptions of the robot as *considerate* ($\eta^2 = .11$ in the more naturalistic interaction scenario of Experiment II, and $\eta^2 = .06$ in Experiment I). Gender also showed an effect of close to a medium size on ratings of *comfort* ($\eta^2 = .06$). However, the size of other effects observed (e.g., due to age) is small ($\eta^2 \leq .03$). Thus, relative to other factors (e.g., the robot's appearance), the robot's politeness and the person's age/gender may be of little importance. While the present work yields implications for both the design of robotic agents and how to evaluate them, future work might consider how relevant gender and politeness are in other contexts or in contrast to other factors.

5 CONCLUSIONS

The primary aim of this research was to investigate whether previous results about human observers' preferences for polite robot speech over more direct speech in an robot instructor would hold for a wider participant demographic, which we were able to confirm. A secondary aim was to explore the modulatory influences of a person's age and gender on perceptions of the robot. Here we obtained several new and important gender effects that hint at a complex interplay of the interaction observers' gender with the observed robot's behavior, which warrants further investigation to elucidate the causal mechanisms responsible for the gender-based differences. Further, owing to a limitation of the design of our first experiment, we explored peoples' perceptions given a more realistic interaction scenario which additionally confirmed the influence of both politeness and gender. These findings are particularly important for the design of future autonomous agents, robotic or virtual, because their success could significantly depend on their ability to adapt, such as to gender-specific expectations of their interactants.

REFERENCES

[1] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu, 'Conversational gaze aversion for humanlike robots', in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 25–32, (2014).

[2] Christoph Bartneck, Tomohiro Suzuki, Takayuki Kanda, and Tatsuya Nomura, 'The influence of peoples culture and prior experiences with aibo on their attitude towards robots', *AI & Society*, **21**(1), 217–230, (2007).

[3] Charles R Crowell, Michael Villano, Matthias Scheutz, and P. Schermerhorn, 'Gendered voice and robot entities: perceptions and reactions of male and female subjects', in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pp. 3735–3741, (2009).

[4] Juan Fasola and Maja J. Mataric, 'A socially assistive robot exercise coach for the elderly', *Journal of Human-Robot Interaction*, **2**(2), 3–32, (2013).

[5] Priska Flandorfer, 'Population ageing and socially assistive robots for elderly persons: the importance of sociodemographic factors for user acceptance', *International Journal of Population Research*, (2012).

[6] ME Foster, A Gaschler, M Guiliani, A Isard, M Pateraki, and RPA Petrick, 'Two people walk into a bar: Dynamic multi-party social interaction with a robot agent', in *International Conference on Multimodal Interaction (ICMI)*, pp. 3–10, (2012).

[7] Bruce Fraser, 'Conversational mitigation', *Journal of Pragmatics*, **4**, 341–350, (1980).

[8] Kerstin Sophie Haring, David Silvera-Tawil, Yoshio Matsumoto, Mari Velonaki, and Katsumi Watanabe, 'Perception of an android robot in japan and australia: A cross-cultural comparison', in *Social Robotics*, 166–175, (2014).

[9] Lewis Hassell and Margaret Christensen, 'Indirect speech acts and their use in three channels of communication', in *Communication Modeling*, p. 9, (1996).

[10] J Hodgins, S Jörg, C O'Sullivan, SI Park, and M Mahler, 'The saliency of anomalies in animated human characters', *ACM Transactions on Applied Perception (TAP)*, **7**(4), 1–14, (2010).

[11] Heeyoung Kim, Sonya S Kwak, and Myungsuk Kim, 'Personality design of sociable robots by control of gesture design factors', in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, pp. 494–499, (2008).

[12] VA Kulyukin, 'On natural language dialogue with assistive robots', in *HRI*, pp. 164–171, (2006).

[13] I Han Kuo, Joel Marcus Rabindran, Elizabeth Broadbent, Yong In Lee, Ngaire Kerse, RMQ Stafford, and Bruce A MacDonald, 'Age and gender factors in user acceptance of healthcare robots', in *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication*, pp. 214–219, (2009).

[14] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski, 'Gracefully mitigating breakdowns in robotic services', in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 203–210, (2010).

[15] Bilge Mutlu, Steven Osman, Jodi Forlizzi, Jessica Hodgins, and Sara Kiesler, 'Task structure and user attributes as elements of human-robot interaction design', in *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication*, pp. 74–79, (2006).

[16] Tatsuya Nomura and Kazuma Saeki, 'Effects of polite behaviors expressed by robots: A psychological experiment in japan', *International Journal of Synthetic Emotions*, **1**(2), 38–52, (2010).

[17] Tatsuya Nomura and Satoru Takagi, 'Exploring effects of educational backgrounds and gender in human-robot interaction', in *Proceedings of the International Conference on User Science and Engineering*, pp. 24–29, (2011).

[18] Natalia Reich and Friederike Eyssel, 'Attitudes towards service robots in domestic environments: The role of personality characteristics, individual interests, and demographic variables', *Journal of Behavioral Robotics*, **4**(2), 123–130, (2013).

[19] Maha Salem, Micheline Ziadee, and Majd Sakr, 'Marhaba, how may I help you?: Effects of politeness and culture on robot acceptance and anthropomorphization', in *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, pp. 74–81, (2014).

[20] Paul Schermerhorn, Matthias Scheutz, and Charles R. Crowell, 'Robot social presence and gender: Do females view robots differently than males?', in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 263–270, (2008).

[21] Megan Strait, Cody Canning, and Matthias Scheutz, 'Let me tell you! Investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality, and distance', in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 479–486, (2014).

[22] Cristen Torry, Susan R. Fussell, and Sara Kiesler, 'How a robot should give advice', in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 275–282, (2013).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Perception of Artificial Agents and Utterance Friendliness in Dialogue

Sascha Griffiths[1] and **Friederike Eyssel**[2] and **Anja Philippsen** and **Christian Pietsch** and **Sven Wachsmuth**[3]

Abstract. The present contribution investigates the construction of dialogue structure for the use in human-machine interaction especially for robotic systems and embodied conversational agents. We are going to present a methodology and findings of a pilot study for the design of task-specific dialogues. Specifically, we investigated effects of dialogue complexity on two levels: First, we examined the perception of the embodied conversational agent, and second, we studied participants' performance following HRI. To do so, we manipulated the agent's friendliness during a brief conversation with the user in a receptionist scenario.

The paper presents an overview of the dialogue system, the process of dialogue construction, and initial evidence from an evaluation study with naïve users ($N = 40$). These users interacted with the system in a task-based dialogue in which they had to ask for the way in a building unknown to them. Afterwards participants filled in a questionnaire. Our findings show that the users prefer the friendly version of the dialogue which scored higher values both in terms of data collected via a questionnaire and in terms of observations in video data collected during the run of the study.

Implications of the present research for follow-up studies are discussed, specifically focusing on the effects that dialogue features have on agent perception and on the user's evaluation and performance.

1 Introduction

Research within the area of "language and emotion" has been identified as one key domain of innovation for the coming years [40, 20]. However, with regard to human-machine communication, we still need better speech interfaces to facilitate human-robot interaction (HRI) [30, 31]. Previous work on human-human communication has already demonstrated that even small nuances in speech have a strong impact on the perception of an interlocutor [1, 38].

In the present work, we have therefore focused on the role of dialogue features (i.e., agent verbosity) and investigated their effects on the evaluation of an embodied conversational agent (ECA) and the user performance. We designed a receptionist scenario involving a newly developed demonstrator platform (see Section 3.2) that offers great potential for natural and smooth human-agent dialogue. To explore how to model dialogues efficiently within actual human-robot interaction we relied on a Wizard-of-Oz paradigm [16, 17].

[1] Queen Mary University of London, UK, email: sascha.griffiths@qmul.ac.uk
[2] New York University, Abu Dhabi, email: fae5@nyu.edu
[3] Bielefeld University, Germany, email: anja.philippsen, christian.pietsch, sven.wachsmuth@uni-bielefeld.de

This HRI scenario involved an embodied conversational agent which served as a receptionist in the lobby of a research center. A similar set-up has been realized in previous studies [2, 24, 25]. Moreover, we draw from existing research on dialogue system design [33] and the acceptance of artificial agents [13, 22].

The question that we seek to answer arises frequently during the implementation of a robot scenario (such as this receptionist scenario) [26], and can also be phrased as how the system should verbalize the information that it is supposed to convey to the user. Obviously, a script has to be provided that covers the necessary dialogue content. The relevant issue is that each utterance can be phrased in a number of ways. This brings up several follow-up questions such as: *Can the perceived friendliness of an agent be successfully manipulated? Is the proposed script a natural way of expressing the intended meaning? Are longer or shorter utterances favourable? How will the user respond to a given wording? Will the script elicit the appropriate responses from the user?*

For the purpose of investigating these questions, we will first discuss related literature and relevant theoretical points. The following section will describe the system. We then turn to the dialogue design and first empirical evidence from a user study.

2 Dialogue Complexity and Perception of Artificial Agents

Obviously, the issue of how to realize efficient dialogue in HRI has been of interest to many researchers in the area of human-machine interaction and principles of natural language generation are generally well understood [39]. However, this is less so the case when taking into account communication patterns between humans and embodied conversational agents and robots.

2.1 Dialogue Complexity and Social Meaning

As Richard Hudson notes, "social meaning is spread right through the language system" [23]. Thus, there is a clear difference between interactions if one commences with the colloquial greeting "*Hi!*" versus one initiated with a more polite "*Good Morning*". However, this does not only concern peripheral elements of language such as greetings, but also syntax. Hudson uses the following example to illustrate this:

1. *Don't you come home late!*
2. *Don't come home late!*

Both sentences differ in terms of syntax and their social meaning. The syntax varies as the first sentence explicitly refers to the subject,

whereas the second sentence does not. The first sentence in the example also appears more threatening in tone than the latter. These subtle differences in the statements' wording lead to a fundamentally different interpretation. Analogously, we assume that in human-agent dialogue subtle manipulations of aspects of that dialogue can result in changes in agent perception. Concretely, we will investigate the role of this kind of linguistic complexity [11] within human-machine interaction.

The impact of changing a dialogue with respect to the social meaning communicated has already been tested in the REA (an acronym for "Real Estate Agent") system [9, 5]. In a study [4] of users' perception of different versions of REA's behaviour, a "normal REA" was tested against an "impolite REA" and a "chatty REA". Results indicated that in the condition in which REA was able to produce a small amount of small talk REA was judged more likeable by participants. In further studies with the system the authors concluded that the interpersonal dimension of interaction with artificial agents is important [8]. It has been shown that implementing a system which achieves task goals and interpersonal goals as well as displaying its domain knowledge can increase the trust a user will have in a system [3]. Cassell [7] also argues that equipping artificial agents with means of expressing social meaning not only improves the users' trust in the domain knowledge that such systems display but also improves interaction with such systems as the users can exploit more of their experience from human-human dialogue.

2.2 Interaction Patterns

The dialogue flow used in the present study was implemented with PaMini, a pattern-based dialogue system which was specifically designed for HRI purposes [32] and has been successfully applied in various human-robot interaction scenarios [35, 36, 37]. The dialogue model underlying the present system (see Section 3.1) is therefore based on generic interaction patterns [33]. Linguistically speaking these are adjacency pairs [29, 10]. In these terms, a dialogue will consist of several invariant elements which are sequentially presented as pairs with one interlocutor uttering one half of the pair in his turn and the other interaction partner responding with an appropriate response.

The full list of generic interaction patterns which are distinguished according to their function given by Peltason et al. [34] includes the following utterance categories: *Greeting, Introducing, Exchanging pleasantries, Task transition, Attracting attention, Object demonstration, Object query, Listing learned objects, Checking, Praising, Restart, Transitional phrases, Closing task, Parting*.

For all these dialogue tasks one can see the interaction as pairs of turns between interlocutors. Each partner has a certain response which fits to the other interlocutor's utterance. Examples of this kind of interaction can be found in Table 1.

Table 1. Examples of adjacency pairs in human-robot interaction (adapted from [34])

Purpose	Example interaction
Greeting	*User: Hello, Vince.* *Robot: Hi, hello.*
Introducing	*User: My name is Dave.* *Robot: Hello, Dave. Nice to meet you.*
Object query	*Robot: What is that?* *User: This is an apple.*
Praising	*User: Well done, Vince.* *Robot: Thank you.*

The problem one faces is that while such dialogues are based on generic speech acts, there is the remaining problem of how the individual items need to be worded. Winograd [46] distinguishes between the ideational function and interpersonal function of language. The ideational function can loosely be understood as the propositional content of an utterance whereas the interpersonal function has more to do with the context of an utterance and its purpose.

3 System Architecture

In the following, we present the system which was constructed both as a demonstrator and as a research platform. We will present the entire set-up which includes an ECA, Vince [42], and a mobile robot platform, Biron [21]. Both of these use the same dialogue manager but only the ECA has been used in this pilot study.

Figure 1 illustrates the architecture of the complete system in autonomous mode. Communication between the components is mainly implemented using the XML-based XCF framework and the Active Memory structure [47]. Three memories are provided for different kinds of information: The short term memory contains speech related information which is inserted and retrieved by the speech recognizer, the semantic processing unit and the dialogue manager. The visual memory is filled by the visual perception components, it contains information about where persons are currently detected in the scene.

The system is designed to provide the visitor verbally with information, but also to guide them to the requested room if necessary[4]. For this purpose, the agent Vince communicates information about the current visitor and his or her needs to the mobile robot Biron via a shared (common ground) memory.

Although Biron is omitted in the present study to reduce complexity, we present the complete system, as Vince and Biron use the same underlying dialogue system. Note that the study could have been conducted also with Biron instead of Vince. Such a study is subject to future work.

3.1 Dialogue Manager

The dialogue manager plays a central role in the overall system as it receives the pre-processed input from the user and decides for adequate responses of the system. A dialogue act may also be triggered by the appearance of persons in the scene as reported by the visual perception component.

Speech input from the user is recognized using the ISR speech recognizer based on ESMERALDA [14]. The semantic meaning is extracted via a parsing component which is possible due to the well defined scenario. Additionally, this component retrieves missing information from an LDAP server that the human might be interested in (e.g. office numbers). The dialogue manager PaMini [35, 36, 37] is based on finite state machines which realize interaction patterns for different dialogue situations as described in Section 2.2. Patterns are triggered by the user or by the robot itself (mixed-initiative). The dialogue component sends the selected response and possibly gesture instructions to the Vince system which synchronizes the speech output and the gesture control internally [28, 27]. Exploiting the information from the visual perception component, Vince attends to the current visitor via gaze following [24].

Biron incorporates a separate dialogue which is coupled with the Vince dialogue. The Biron dialogue at the moment receives input

[4] A short video demonstration of the scenario is provided in this CITEC video: `http://www.youtube.com/watch?v=GOz_MsLel1Y#t=4m32s`. Accessed: March 2, 2015

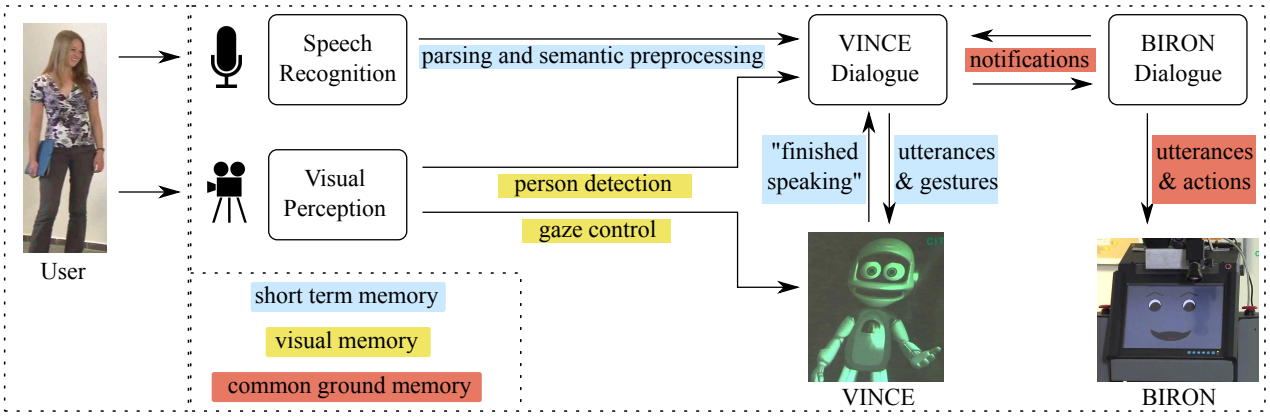


Figure 1. Overview of the architecture of the system in autonomous mode. The colors of the three memories indicate which information is stored in which memory. See Section 3.1 for a thorough description of the information flow.

solely from the Vince dialogue component (not from the user) and communicates the current state to the user. If the visitor wishes, Vince calls Biron and orders him to guide the visitor to the requested room. This feature is currently limited to offices on the ground floor, if visitors are looking for a room on the first or second floor, Biron guides them to the elevator and provides them with information about how to find the room on their own.

3.2 Demonstrator Platform

The embodied conversational agent Vince is installed on a workstation. An Apple Mac Mini is used for this purpose. The system runs a UNIX based operating system (Linux Ubuntu 10.04 32bit). The user interface is controlled by a wireless bluetooth mouse and keyboard or via remote access. The ECA is displayed on a holographic projection screen (i.e. a HoloPro Terminal[5]) in order to achieve a high degree of perceived embodiment. A microphone records speech input and video data are recorded using two cameras. Two loudspeakers are connected to the Mac Mini workstation to provide audio output.

4 Study Design and Realisation

We set up a simplified version of the CITEC Dialogue Demonstrator for the purpose of the study. One difference is that we do not make use of the mobile robot Biron here. Secondly, we rely on Wizard-of Oz teleoperation [12, 45] to trigger interaction patterns by means of a graphical user interface that was designed for our case study.

4.1 Preparation of Dialogues

The dialogues were prepared bottom-up. We tried to leave as little as possible to design by the researchers or a single researcher.

To investigate human-machine dialogue in the context of a receptionist scenario, we initially simulated such dialogues between two human target persons who were given cards which described a particular situation (e.g. that a person would be inquiring about another persons office location).

We recorded two versions of eight dialogues with the two participants, who were asked to take the perspective of a receptionist or a visitor, respectively. The dialogues were then transliterated by a third party who had not been involved in the staged dialogues.

To model the receptionist turns, we extracted all phrases which were classified as greetings, introductions, descriptions of the way to certain places and farewells. We then constructed a paper-and-pencil pre-test in order to identify a set of dialogues that differed in friendliness. 20 participants from a convenience sample were asked to rate the dialogues with regard to perceived friendliness using a 7-point Likert scale.

These ratings were used as a basis to construct eight sample dialogues which differed both in friendliness and verbosity. In a subsequent online pre-test, the sample dialogues were embedded in a cover-story that resembled the set-up of our WoZ scenario.

We used an online questionnaire to test how people perceived these dialogues. On the start screen participants were presented with a picture of the embodied conversational agent Vince and told that he would serve as a receptionist for the CITEC building. On the following screens textual versions of the eight human-agent dialogues were presented. Participants were asked to rate these dialogues with regard to friendliness in order to identify dialogues that would be perceived as either low or high in degree of perceived friendliness of the interaction.

The dialogue with the highest rating for friendliness and the dialogue with the lowest rating for friendliness were then de-composed into their respective parts and used in the main study. The two dialogue versions are presented in Table 2.

4.2 Study

In the main study, the participants directly interacted with the ECA which was displayed on a screen (see Figure 1).

We recruited students and staff at the campus of Bielefeld University to participate in our study on “human-computer interaction”. 20 male and 20 female participants ranging in age from 19 to 29 years (M = 23.8 years, SD = 2.36) took part in the study. Before beginning their run of the study, each participant provided informed consent. Each participant was then randomly assigned to one of two conditions in which we manipulated dialogue friendliness.

The study involved two research assistants (unbeknownst to the participants). Research assistant 1 took over the role of the “wizard” and controlled the ECA’s utterances, while research assistant 2 interacted directly with the participants.

[5] http://www.holopro.com/de/produkte/holoterminal.html Accessed: March 2, 2015

Table 2. Friendly and neutral dialogue version

Dialogue Act	Neutral version	Friendly version
Greeting	Hallo *Hello*	Guten Tag, kann ich Ihnen helfen? *Good afternoon, how can I help you?*
Directions	Der Fragebogen befindet sich in Q2-102. *The questionnaire is located in Q2-102.*	Der Fragebogen befindet sich in Raum Q2 102. Das ist im zweiten Stock. Wenn Sie jetzt zu Ihrer Rechten den Gang hier runter gehen. Am Ende des Gangs befinden sich die Treppen, diese gehen Sie einfach ganz hoch und gehen dann durch die Feuerschutztür und dann ist der Raum einfach geradeaus. *The questionnaire is located in room Q2-102. That is on the second floor. If you turn to your right and walk down the hallway. At the end of the floor you will find the stairs. Just walk up the stairs to the top floor and go through the fire door. The room is then straight ahead.*
Farewell	Wiedersehen. *Goodbye.*	Gerne. *You are welcome.*

Following the Wizard-of-Oz paradigm, research assistant 1 was hidden in the control room and controlled the ECA's verbalisations using a graphical user interface. A video and audio stream was transmitted from the dialogue system to the control room. The "wizard" had been trained prior to conducting the study to press buttons corresponding to the "Dialogue Acts" as shown in Table 2. Importantly, research assistant 1 only knew the overall script (containing a greeting, a description of the route to a room and a farewell), but was blind to the authors' research questions and assumptions.

To initiate the study, research assistant 1 executed "Greeting A" or "Greeting B", depending on whether the "friendly" or "neutral" condition was to be presented, then proceeded to pressing "Directions A" or "Directions B" and finally "Farewell A" and "Farewell B" once the user had reacted to each utterance.

The users then had to follow the instruction given by the agent. Research assistant 2 awaited them at the destination where they had to fill in a questionnaire asking for their impressions of the interaction.

The questionnaire investigated whether differential degrees of dialogue complexity would alter the perception of the artificial agent with respect to a) warmth and competence [15], b) mind attribution [19], and c) usability (system usability scale *SUS*) [6]. We consider these question blocks as standard measures in social psychology and usability studies.

The questionnaire was comprised of three blocks of questions. These do to some extent correspond to the four paradigms of artificial intelligence research listed in Russell & Norvig [41]: "thinking humanly", "acting humanly", "thinking rationally" and "acting rationally". As we were only looking at perception of the artificial agent, we did not look into "thinking rationally". However, warmth and competence are used in research on anthropomorphism, which one can regard as a form of "acting humanly". Mind perception can be related to "thinking humanly". Usability (SUS) is a form of operationalising whether an artificial agent is acting goal driven and useful which holds information on whether it is "acting rationally".

The first block of the questionnaire included four critical items on warmth, and three critical items on competence, as well as nine filler items. The critical questions asked for attributes related to either warmth, such as "good-natured", or competence, such as "skillful".

The second block consisted of 22 questions related to mind perception. These questions asked the participants to rate whether they believed that Vince can be attributed mental states. A typical item is the question whether Vince was capable of remembering events or whether he is able to feel pain.

Finally, the SUS questionnaire consisted of 10 items directly related to usability. Participants were asked question such as whether they found the system easy to use.

Upon completion of the questionnaire, participants were debriefed, reimbursed and dismissed.

5 Results

In the following, two types of results are reported. In Section 5.1, we present results from the questionnaire, in Section 5.2, we present initial results from video data recorded during the study.

5.1 Questionnaire Responses

As aforementioned, 7-point Likert scales (for the warmth, competence and mind question blocks) and a 5-point Likert scale for the SUS questions block) were used to measure participants responses to the dependent measures. For each dependent variable, mean scores were computed with higher values reflecting greater endorsement of the focal construct. Values for the four blocks of questions were averaged for further analysis. The results for the questionnaire are shown in Figure 2.

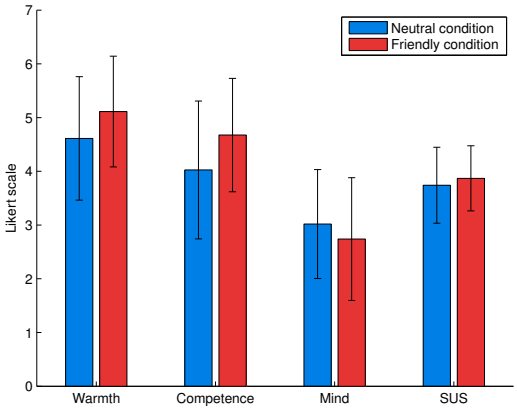


Figure 2. Mean response values for the questionnaire question sets. The mean for the dependent variables warmth, competence, mind and SUS are compared for the two categories neutral (blue) and friendly (red).

5.1.1 Warmth

The mean values for the warmth question set can be seen in Figure 2. It can be notices that the values for the friendly condition are mostly higher than for the neutral condition. The descriptive statistics confirm this. The friendly condition has a maximum value of 7 and a minimum value of 3.25 whereas the neutral condition has a maximum value of 6.75 and a minimum value of 2.25. The mean of the friendly condition is $M = 5.11$ ($SD = 1.14$) and the mean of the neutral condition is $M = 4.61$ ($SD = 1.14$). The mean values suggest that

within the population on which our system was tested the friendly condition is perceived warmer than the neutral condition.

5.1.2 Competence

Similarly, the values for the friendly condition are mostly higher than for the neutral condition. The descriptive statistics confirm this. The friendly condition has a maximum value of 7 and a minimum value of 2.75 whereas the neutral condition has a maximum value of 6.25 and a minimum value of 1.5. The mean of the friendly condition is $M = 4.68$ ($SD = 1.05$) and the mean of the neutral condition is $M = 4.02$ ($SD = 1.28$). The standard deviation shows that there is more variation in the values for the neutral condition. The mean values overall suggest that within the population on which our system was tested the friendly condition is perceived more competent than the neutral condition.

5.1.3 Mind Perception

As Figure 2 shows, the ECA is perceived slightly higher on mind perception in the neutral condition than in the the friendly condition. The neutral condition has a maximum value of 4.9 and a minimum value of 1.32 whereas the friendly condition has a maximum value of 4.93 and a minimum value of 1.09. However, the mean of the neutral condition is $M = 3.02$ ($SD = 1.01$) whereas the mean of the friendly condition is $M = 2.74$ ($SD = 1.14$). The standard deviation suggests that there is more variation in the values for the neutral condition. The mean values overall suggest that within the population on which our system was tested in the friendly condition the participants attributed less mind to the ECA than the neutral condition.

5.1.4 System Usability Scale (SUS)

The values on the system usability scale are slightly higher in the friendly condition than in the neutral condition. The friendly condition has a maximum value of 4.7 and a minimum value of 2.7 whereas the neutral condition has a maximum value of 4.9 and a minimum value of 2.5. The mean of the friendly condition is $M = 3.87$ ($SD = 0.61$) and the mean of the neutral condition is $M = 3.74$ ($SD = 0.71$). The standard deviation suggests that there is more variation in the values for the neutral condition. The mean values overall suggest that within the population on which our system was tested the friendly condition was rated slightly more usable than the neutral condition.

5.2 Further Observations

Further observations that could be made on the dialogue level resulted from the analysis of the video data collected during the runs of the study. The dialogues were transcribed and inspected by one student assistant[6] trained in conversation analysis [18]. The purpose of this was to examine the dialogues to find out whether there were any particular delays in the dialogues and whether participants conformed to the script or not.

[6] Taking this line of research further, we would use two annotators and check for agreement between them. However, this was beyond the scope of the current contribution.

5.2.1 Alignment

We looked at the mean utterance length (MUL) of the participants in interaction with the ECA. We take this as an indicator of how participants align their verbalisations with the agent's verbalisations. The differences between the two conditions can be seen in in Figure 3, the values for the friendly condition are mostly higher than for the neutral condition.

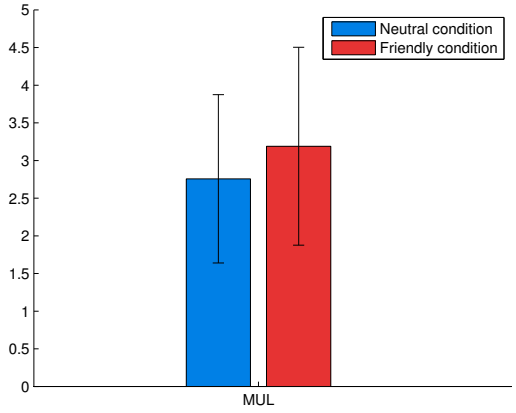


Figure 3. The mean utterance length averaged over the two conditions. The friendly condition has a slightly higher mean value than the neutral condition.

The descriptive statistics confirm this. The friendly condition has a maximum value of 5.5 and a minimum value of 1 whereas the neutral condition has a maximum value of 5.25 and a minimum value of 1. The mean of the friendly condition is $M = 3.12$ ($SD = 1.31$) and the mean of the neutral condition is $M = 2.76$ ($SD = 1.11$). The standard deviation suggests that there is more variation in the values for the friendly condition. The mean values overall suggest that within the population on which our system was tested the friendly condition showed more alignment with the ECA's MUL than the neutral condition.

5.2.2 Irregularities

The video data were reviewed and four types of noticeable effects on the dialogue were determined:

1. Participants returning because they did not understand or forget the ECA's instructions (22.5%, see Section 5.2.3),
2. deviations from the script, i.e. participants trying to do small talk with the ECA (5%, see Section 5.2.4),
3. timing difficulties causing delays in the interaction (25%), and
4. other ways in which the script was altered in small ways (22.5%, e.g. mismatches between the ECA's utterances and the participants utterances).

The overall number of irregularities accumulated across the two categories is summarized in Table 3. In interactions with the neutral condition irregularities can be observed in 75% of the cases, while in the friendly condition only 50% of the interactions show irregularities.

Table 3. Overview of occurred irregularities in the neutral and friendly condition.

	Neutral	Friendly
No irregularities	5	10
Irregularities occur	15	10

5.2.3 Clarity of instructions

Out of the 40 interactions in 9 cases (22.5%) the participants returned because they realized that they could not remember the room number correctly. Out of these the majority, namely 6, were in the neutral condition. Three participants came back for a second short interaction with Vince in the friendly condition.

5.2.4 Small talk

Only two participants (5%) deviated from the script of the dialogue by attempting to do small talk with Vince. Both of these were in the friendly condition. One participant asked the ECA for its name. Another participants tried three deviating questions on Vince during the interaction. The first question was "How are you?", the second "What can you tell me?", and finally the ECA was asked whether they were supposed to actually go to the room after the instructions were given.

6 Discussion

In reporting our results we concentrated on the descriptive statistics and no attempt will be made to generalize beyond this population. Within this first pilot study with the current demonstrator we tried to assess whether manipulating the degree of perceived friendliness has an effect on the interaction.

We now return to the questions asked in the introduction, the main question being how the manipulation affected the interaction between the user and the artificial agent.

6.1 Can the perceived friendliness of an agent be successfully manipulated?

We obtained slightly higher values regarding the perceived warmth in the friendly condition as opposed to the neutral condition. The differences are very small, though. The descriptive statistics point towards a "friendly" version of the dialogue actually being perceived as more friendly by the user. We propose that this will make users more willing to use the services the system can provide. Thus, further research into "friendly agents" seems a productive agenda.

The friendliness level also suggested higher ratings for competence, despite the fact that the friendly dialogue actually led to more misunderstandings. This failure was not reflected in the users judgements directly. Also, participants seem to prefer interacting with the friendly agent.

6.2 Is the proposed script a natural way of expressing the intended meaning?

The results which the video data analysis presented indicate that actually the majority of interactions conducted within this study were smooth and there were no noticeable deviations from the overall "script" in most dialogues. The operator was able to conduct most of the dialogues with the use of just a few buttons. This suggests that one can actually script dialogues of this simple nature quite easily.

However, the wording is crucial and the results suggest that the friendly version of the dialogue is more amicable to clarity. Only three participants did not fully understand or remember the instructions whereas twice as many had to ask for the room a second time in the neutral condition.

6.3 Are longer or shorter utterances favourable?

In a task-based dialogue the artificial agent will ideally demonstrate its knowledge and skill in a domain. However, the pilot-study did not find a very high difference between the two conditions regarding the competence question. The descriptive statistics, however, suggest that the longer utterances in the friendly dialogue received higher competence ratings.

Converse to the prediction, mind perception was slightly higher for the neutral dialogue, though. Thus, the friendly agent is not necessarily perceived as more intelligent by the user.

However, the longer utterances in the friendly version of the dialogue received higher ratings with respect to usability. Also, fewer participants had to come back and ask for the way again in a second interaction in the friendly condition. This suggests that the longer version of the dialogue better conveyed the dialogue content than the neutral version.

6.4 How does the user respond to a given wording?

In the friendly condition, users used longer utterances themselves when speaking to the friendly version of the ECA with more verbose verbalisations. This shows that the participants do align their speech with that of the artificial agent.

One can also tell from the video analysis that only in the friendly condition participants were motivated to further explore the possibilities the system offers. Two participants decided to ask questions which went beyond the script.

6.5 Will the script elicit the appropriate responses from the user?

Participants found it easy to conform to the proposed script. There was only a low percentage of participants who substantially deviated from the script and stimuli presented by the ECA (5% tried to do small talk with the agent). Most dialogues proceeded without the participants reacting in unanticipated ways and only a small percentage of participants failed to extract the relevant information from the verbalisations of the artificial agent.

7 Conclusion

We presented a pilot-study in which participants were confronted with dialogue exhibiting different degrees of friendliness.

While maintaining the same ideational function (see Section 2.2 above) we changed the interpersonal function of the dialogue by using sentences which were obtained through a role-playing pre-study and then rated by participants according to their friendliness.

The obtained dialogues (a friendly and a neutral version) were presented to participants in interaction with an ECA which was implemented via generic interaction patterns. Participants filled in a questionnaire after the interaction which was analysed along with further observational data collected during the study.

The results point towards higher perceived warmth, higher perceived competence and a greater usability judgement for the ECA's

performance in the friendly condition. However, mind perception does not increase in the more friendly dialogue version.

Further research should replicate our findings using a larger sample size. Also, in a similar study the variation of friendliness in interaction had less impact on the participants' perception than the interaction context [43]. Thus, one would have to take a closer look at how politeness and context interact in future studies. In addition, related literature also suggests that anthropomorphic perceptions could be increased by increased politeness [44]. Thus, friendliness can generally be expected to have an effect on the perception of artificial agents.

The dialogue in the present study not only varied in terms of friendliness but also in terms of verbosity. It could be argued that this is not the same and a higher verbosity might have had an unwanted effect, especially on the user's task performance. Future studies could consider whether they can be designed to investigate the effect of friendliness without directly changing agent verbosity.

It would also be interesting to conduct a similar study to explore dialogue usage in the robot Biron. As he is supposed to guide the visitor to the requested room, he spends several minutes with the visitor without exchanging necessary information, thus, is can be expected that the usage of small talk affects the interaction in a positive way.

ACKNOWLEDGEMENTS

The authors would like to thank all colleagues who contributed to the Dialogue Demonstrator: Anja Durigo, Britta Wrede, Christina Unger, Christoph Broschinski, David Schlangen, Florian Lier, Frederic Siepmann, Hendrik Buschmeier, Jan De Ruiter, Johanna Müller, Julia Peltason, Lukas Twardon, Marcin Wlodarczak , Marian Pohling, Patrick Holthaus, Petra Wagner, Philipp Cimiano, Ramin Yaghoubzadeh, Sebastian Ptock, Sebastian Wrede, Thorsten Spexard, Zofia Malisz, Stefan Kopp, and Thilo Paul-Stueve. The research reported here was funded by the Cluster of Excellence "Cognitive Interaction Technology" (EXC 277). Griffiths is also partly supported by ConCreTe: the project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733. The authors would also like to thank three anonymous reviewers of the contribution for their very useful and productive feedback.

REFERENCES

[1] Nalini Ambady, Debi LaPlante, Thai Nguyen, Robert Rosenthal, Nigel Chaumeton, and Wendy Levinson, 'Surgeons' tone of voice: A clue to malpractice history', *Surgery*, **132**(1), 5–9, (July 2002).

[2] Niklas Beuter, Thorsten P. Spexard, Ingo Lütkebohle, Julia Peltason, and Franz Kummert, 'Where is this? - Gesture based multimodal interaction with an anthropomorphic robot', in *International Conference on Humanoid Robots*. IEEE-RAS, (2008).

[3] Timothy Bickmore and Julie Cassell, 'Small talk and conversational storytelling in embodied conversational interface agents', in *Proceedings of the AAAI Fall Symposium on "Narrative Intelligence"*, pp. 87–92, (1999).

[4] Timothy Bickmore and Justine Cassell, '"how about this weather?" social dialog with embodied conversational agents', in *Proceedings of the American Association for Artificial Intelligence (AAAI) Fall Symposium on "Narrative Intelligence"*, pp. 4–8, Cape Cod, MA, (2000).

[5] Timothy Bickmore and Justine Cassell, 'Relational agents: a model and implementation of building user trust', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 396–403. ACM, (2001).

[6] John Brooke, 'SUS – a quick and dirty usability scale', *Usability evaluation in industry*, **189**, 194, (1996).

[7] Justine Cassell, 'Embodied conversational agents: representation and intelligence in user interfaces', *AI Magazine*, **22**(3), 67–83, (2001).

[8] Justine Cassell and Timothy Bickmore, 'Negotiated collusion: Modeling social language and its relationship effects in intelligent agents', *User Modeling and User-Adapted Interaction*, **13**(1-2), 89–132, (2003).

[9] Justine Cassell, Timothy Bickmore, Mark Billinghurst, Lee Campbell, Kenny Chang, Hannes Vilhjálmsson, and Hao Yan, 'Embodiment in conversational interfaces: Rea', in *Proceedings of the CHI'99 Conference*, pp. 520–527. ACM, (1999).

[10] David Crystal, *A Dictionary of Linguistics and Phonetics*, Blackwell Publishers, sixth edn., 2008.

[11] Östen Dahl, *The Growth and Maintenance of Linguistic Complexity*, John Benjamins Publishing Company, Amsterdam/Philadelphia, 2004.

[12] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg, 'Wizard of oz studies: why and how', in *Proceedings of the 1st international conference on Intelligent user interfaces*, pp. 193–200. ACM, (1993).

[13] Friederike Eyssel and Dieta Kuchenbrandt, 'Manipulating anthropomorphic inferences about NAO: The role of situational and dispositional aspects of effectance motivation', in *2011 RO-MAN*, pp. 467–472. IEEE, (July 2011).

[14] Gernot A. Fink, 'Developing HMM-based recognizers with ESMERALDA', in *Text, Speech and Dialogue*, eds., Václav Matousek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, volume 1692 of *Lecture Notes in Computer Science*, pp. 229–234. Springer Berlin Heidelberg, (1999).

[15] Susan T Fiske, Amy J C Cuddy, and Peter Glick, 'Universal dimensions of social cognition: warmth and competence.', *Trends in cognitive sciences*, **11**(2), 77–83, (February 2007).

[16] Norman M Fraser and G Nigel Gilbert, 'Simulating speech systems', *Computer Speech & Language*, **5**(1), 81–99, (1991).

[17] Dafydd Gibbon, Roger Moore, and Richard Winski, *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, 1997.

[18] Charles Goodwin and John Heritage, 'Conversation analysis', *Annual Review of Anthropology*, **19**, 283–307, (1990).

[19] Heather M Gray, Kurt Gray, and Daniel M Wegner, 'Dimensions of mind perception.', *Science (New York, N.Y.)*, **315**(5812), 619, (February 2007).

[20] Sascha Griffiths, Ciro Natale, Ricardo Araújo, Germano Veiga, Pasquale Chiacchio, Florian Röhrbein, Stefano Chiaverini, and Reinhard Lafrenz, 'The ECHORD Project: A General Perspective', in *Gearing up and accelerating cross-fertilization between academic and industrial robotics research in Europe:*, eds., Florian Röhrbein, Germano Veiga, and Ciro Natale, volume 94 of *Springer Tracts in Advanced Robotics*, 1–24, Springer International Publishing, Cham, (2014).

[21] Axel Haasch, Sascha Hohenner, Sonja Hüwel, Marcus Kleinehagenbrock, Sebastian Lang, Ioannis Toptsis, Gernot A. Fink, Jannik Fritsch, Britta Wrede, and Gerhard Sagerer, 'Biron - the Bielefeld robot companion', in *Proc. Int. Workshop on Advances in Service Robotics*, eds., Erwin Prassler, Gisbert Lawitzky, P. Fiorini, and Martin Hägele, pp. 27–32. Fraunhofer IRB Verlag, (2004).

[22] Frank Hegel, Friederike Anne Eyssel, and Britta Wrede, 'The social robot Flobi: Key concepts of industrial design', in *Proceedings of the 19th IEEE International Symposium in Robot and Human Interactive Communication (RO-MAN 2010)*, pp. 120–125, (2010).

[23] Joseph Hilferty, 'Interview with Richard Hudson', *Bells: Barcelona English language and literature studies*, **16**, 4, (2007).

[24] Patrick Holthaus, Ingo Lütkebohle, Marc Hanheide, and Sven Wachsmuth, 'Can I help you? - A spatial attention system for a receptionist robot', in *Social Robotics*, eds., Shuzhi Sam Ge, Haizhou Li, John-John Cabibihan, and Yeow Kee Tan, pp. 325–334. IEEE, (2010).

[25] Patrick Holthaus and Sven Wachsmuth, 'Active peripersonal space for more intuitive HRI', in *International Conference on Humanoid Robots*, pp. 508–513. IEEE RAS, (2012).

[26] Patrick Holthaus and Sven Wachsmuth, 'The receptionist robot', in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp. 329–329. ACM, (2014).

[27] Stefan Kopp, 'Social adaptation in conversational agents', *PerAda Magazine (EU Coordination Action on Pervasive Adaptation)*, (2009).

[28] Stefan Kopp and Ipke Wachsmuth, 'Synthesizing multimodal utterances for conversational agents', *Computer Animation and Virtual Worlds*, **15**(1), 39–52, (2004).

[29] Stephen C Levinson, *Pragmatics*, Cambridge University Press, Cambridge, 1983.

[30] Nikolaos Mavridis, 'A review of verbal and non-verbal human–robot in-

teractive communication', *Robotics and Autonomous Systems*, **63**, 22–35, (2015).
[31] Roger K Moore, 'From talking and listening robots to intelligent communicative machines', in *Robots that Talk and Listen – Technology and Social Impact*, 317 – 336, De Gruyter, Boston, MA, (2014).
[32] Julia Peltason, 'Position paper: Julia Peltason', in *6th Young Researchers' Roundtable on Spoken Dialogue Systems*, pp. 63 – 64, (2010).
[33] Julia Peltason, *Modeling Human-Robot-Interaction based on generic Interaction Patterns*, Ph.D. dissertation, Bielefeld University, 2014.
[34] Julia Peltason, Hannes Rieser, Sven Wachsmuth, and Britta Wrede, 'On Grounding Natural Kind Terms in Human-Robot Communication', *KI - Künstliche Intelligenz*, (March 2013).
[35] Julia Peltason and Britta Wrede, 'Modeling Human-Robot Interaction Based on Generic Interaction Patterns', in *AAAI Fall Symposium: Dialog with Robots*, pp. 80 —- 85, Arlington, VA, (2010). AAAI Press.
[36] Julia Peltason and Britta Wrede, 'Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns', in *SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 229–232, The University of Tokyo, (2010). Association for Computational Linguistics.
[37] Julia Peltason and Britta Wrede, 'The curious robot as a case-study for comparing dialog systems', *AI Magazine*, **32**(4), 85–99, (2011).
[38] Rajesh Ranganath, Dan Jurafsky, and Daniel A McFarland, 'Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates', *Computer Speech & Language*, **27**(1), 89–115, (2013).
[39] Ehud Reiter and Robert Dale, *Building Natural Language Generation Systems*, Studies in Natural Language Processing, Cambridge University Press, Cambridge, 2000.
[40] F Röhrbein, S Griffiths, and L Voss, 'On industry-academia collaborations in robotics', Technical report, Technical Report TUM-I1338, (2013).
[41] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall International, Harlow, third int. edn., 2013.
[42] Amir Sadeghipour and Stefan Kopp, 'Embodied gesture processing: Motor-based integration of perception and action in social artificial agents', *Cognitive computation*, **3**(3), 419–435, (2011).
[43] Maha Salem, Micheline Ziadee, and Majd Sakr, 'Effects of politeness and interaction context on perception and experience of HRI', in *Social Robotics*, eds., Guido Herrmann, Martin J. Pearson, Alexander Lenz, Paul Bremner, Adam Spiers, and Ute Leonards, volume 8239 of *Lecture Notes in Computer Science*, 531–541, Springer International Publishing, (2013).
[44] Maha Salem, Micheline Ziadee, and Majd Sakr, 'Marhaba, how may I help you?: Effects of politeness and culture on robot acceptance and anthropomorphization', in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp. 74–81. ACM, (2014).
[45] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati, 'The oz of wizard: simulating the human for interaction research', in *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*, pp. 101–107. IEEE, (2009).
[46] Terry Winograd, *Language as a cognitive process (Vol. 1)*, Addison-Wesley, Reading, MA, 1983.
[47] Sebastian Wrede, Jannik Fritsch, Christian Bauckhage, and Gerhard Sagerer, 'An XML based framework for cognitive vision architectures', in *Proc. Int. Conf. on Pattern Recognition*, number 1, pp. 757–760, (2004).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Turn-yielding cues in robot-human conversation

Jef A. van Schendel and **Raymond H. Cuijpers**[1]

Abstract. If robots are to communicate with humans in a successful manner, they will need to be able to take and give turns during conversations. Effective and appropriate turn-taking and turn-yielding actions are crucial in doing so. The present study investigates the objective and subjective performance of four different turn-yielding cues performed by a NAO robot. The results show that an artificial cue, flashing eye-LEDs, lead to significantly shorter response times by the conversational partner than not giving any cue and was experienced as an improvement to the conversation. However, stopping arm movement or head turning cues showed, respectively, no significant difference or even longer response times compared to the baseline condition. Conclusions are that turn-yielding cues can lead to improved conversations, though it depends on the type of cue, and that copying human turn-yielding cues is not necessarily the best option for robots.

1 INTRODUCTION

"Beep boop!" Will our future robot partners communicate with us like Star Wars' R2D2? A more desirable future would be one where we can interact with robots in a fluent and pleasant manner, using the same natural language we use to talk to other people.

As robots grow more advanced, they are able to help us out in more areas of our lives. An area of interest is for instance elderly care, since healthcare costs in European countries are on the rise [6], and the 80+ population in Europe is expected to more than double from 2013 to 2050 [23]. Robots could increase cost-efficiency and have shown positive effects in this area [5].

But no matter what type of work, socially assistive robots as they are called [22], should be not just able to successfully perform their tasks, but deal with human beings in an appropriate, respectful and productive manner. This requires a way to naturally communicate with them, which involves taking and giving turns. This is also called managing the conversational floor.

1.1 Turn-taking

To manage the conversational floor, humans make use of turn-taking and turn-yielding cues [8]. One way to give such cues is through speech itself: the intention to yield a turn can be made clear through syntax (for instance, ending with a direct question) but also changes in intonation or speaking rate [10, 13]. Using these cues requires understanding what is being said, which is difficult for robots. Another way is through non-verbal cues, given through body movement or gaze direction [16]. The major advantage of non-verbal cues is that they do not require speech to be intelligible.

[1] Human Technology Interaction group, Eindhoven University of Technology, the Netherlands, email: r.h.cuijpers@tue.nl

Existing research has investigated ways for robots and other agents to shape and guide a conversation. Positive results have been found when robots have been used to implement conversational gaze behavior [2, 18, 21] and gestures [14, 17], likewise with agents who make use of eye gaze [1, 7, 19], especially when it is appropriate in context [9, 12]. Other researchers investigated both gestures and eye gazing by robots and, in certain combinations, found positive effects on message retention [24] and persuasion [11]. Others still moved on from dyadic sessions to conversations where a robot speaks with multiple people, so-called multiparty settings [3, 4, 15, 18, 25].

Since non-verbal cues have shown promising results in studies such as these, and can be implemented relatively easily for robots, they are of interest for the present study.

While turn-taking has been investigated in many studies, most of them evaluate a combination of turn-yielding cues as a whole and do not compare the effectiveness of isolated turn-yielding cues. Some authors, such as [4], have built interaction models for agents that include turn-yielding. In their study, the assessment of turn-yielding behavior is mixed with other types of interaction. Additionally, the subjective assessment is based on a single condition and is not compared to other models, which makes it difficult to understand the relative contribution of different turn-yielding cues. Therefore, we designed a study in which we can compare the effectiveness and user evaluation of a number of non-verbal turn-yielding cues. The response time of the conversation partner is used as an objective measure, because a shorter response time could mean better and more fluent conversational flow. Shiwa and colleagues [20] already showed that this does not necessarily signify a more pleasant interaction, which is why a questionnaire is used to evaluate the participants' opinion on the value of the different cues. This study will give us further insights in how to employ non-verbal turn-yielding and turn-taking cues during human-robot interaction.

1.2 Turn-yielding cues

Four different turn-yielding cues were selected, based on existing literature.

The first two were based on common human cues and labelled *turn head* and *stop arms*. The former means that the speaker directs its gaze away from the conversational partner during speaking, then returns to the partner when yielding the turn [16]. For the latter, the speaker uses co-speech gestures while talking, but stops doing so when finished. It is based on the idea that interlocutors make certain continuous movements during speaking, but stop moving as a sign that their turn is over [16].

For the third cue, an artificial action was chosen, namely *flash eyes*, where the robot briefly increases the brightness of its eye-LEDs. This condition was added to investigate whether cues have to be based on existing human behavior or not. This cue is not natural in the sense

that it is humanlike, but it is a very common way to communicate non-verbally for robots (and many other technical devices).

The last cue was called *stay silent* and served as the baseline condition. Here, the robot simply stopped speaking with no further action.

These four cues were performed by a robot in dyadic sessions with human conversational partners. In order to generate a large number of turn-yielding events we developed a new task where the participant and the robot took turns to verbally cite the letters of the alphabet. As soon as the robot stopped citing, the participant continued citing letters. After a few letters, the robot continued again. The turn-yielding cues employed by the robot were manipulated.

2 METHOD

2.1 Participants

A total of 20 participants took part in the experiment. One was unable to complete the task and therefore the data in question was not used in the analysis. Roughly half of the participants were recruited from the J.F. Schouten participant database, while the others were recruited through word-of-mouth and invitations via social networks. The only requirement set beforehand was that the participants were able of hearing. Of the 19 participants, 13 were female. All participants were offered monetary compensation or course credits for their time.

2.2 Design

The performed experiment had a within-subjects, repeated measures design with four conditions.

The independent variable in this study was the turn-yielding cue used by the robot. The four conditions, as described under 1.2, were labelled *stay silent*, *stop arms*, *turn head* and *flash eyes*. These were randomly selected by the robot during the experiment.

The dependent variable was the response time of the participant. Specifically, this time was defined as the length in milliseconds between the start of the robot's turn-yielding cue and the beginning of the participant's speech.

Additionally, the participants filled out a questionnaire after the experiment. The questionnaire began by asking the participants which of the four cues they remember noticing. Then, a number of questions asked about their opinion on the four conditions, using a five-point Likert scale. The order of the questions was randomized for each participant in order to minimize ordering bias.

2.3 Setup

This study used a 58-centimeter tall humanoid robot called NAO, developed by Aldebaran Robotics. It has 25 degrees of freedom for movement and various sensors. Of particular interest for this study was its microphone, however, due to unsatisfactory performance during pre-tests, an external microphone was used for the experiment. Both the NAO and the microphone were connected to a laptop, used for controlling the experiment and saving the data.

The experiment took place in the GameXPLab, a laboratory modelled after a living room at Eindhoven University of Technology. Participants were seated in front of a small desk, with the NAO on top of the desk and a small wireless microphone placed between them.

2.4 Procedure

During a short introduction, the participants were given their task: together with the NAO, they were to repeatedly cite the letters of the alphabet. The NAO would start and after a randomly chosen amount of letters it would stop speaking and perform one of the turn-yielding cues. Then, the participant would continue until the NAO started speaking again. The robot autonomously decided when to speak by listening for 2, 3 or 4 utterances after which it waited for a silence to start speaking. The number of utterances determines which letter should be used next. Occasionally, the robot made a mistake (e.g. when it mistook another sound for an letter) or interrupted a person, but this was never a problem from the user's point of view. A small timing delay (0.5s) was added to make the flow as natural as possible. This cycle continued for roughly 15 minutes with each participant.

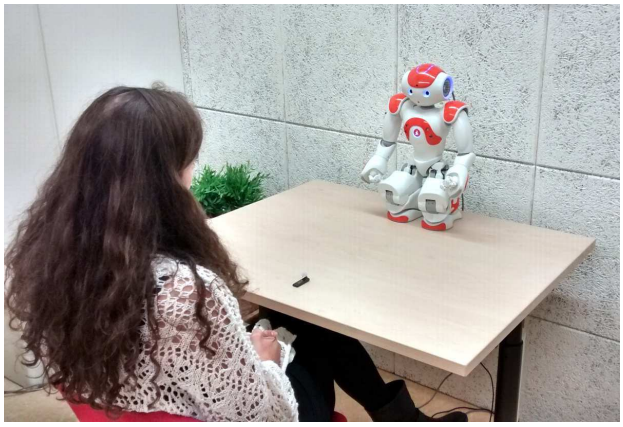


Figure 1. Experiment setup

This particular task was chosen for several reasons. First, the answers by the participants would mostly be single-syllable words, which would make them easier to accurately detect with the microphone and enable the robot to count them, so it would know where to continue the series. The second reason was the assumption that the participants would be able to recall the letters of the alphabet with minimal effort, thereby minimizing the influence of recollection time. Thirdly, the advantage of using a fixed sequence would be to avoid the need for the participant to decide on what to say. In other words, the aim was to control for possibly confounding variables such as recollection time or deliberation time.

Afterwards, the participants filled out a questionnaire (further described under 2.2).

3 RESULTS

3.1 Experiment results

The experiment data was edited and analyzed using SPSS. A number of false positives were recorded as notes during the experiment. After these were removed, a total of 1310 valid data points were left, or about 68.9 recorded measurements per participant.

The distribution of the response time data was found to be skewed right ($\text{skewness} = 1.520 \pm 0.068$) and peaked ($\text{kurtosis} = 5.370 \pm 0.135$). To increase normality it was logarithmically transformed. Histograms of the original (a) and log-transformed (b) data can be found in Figure 3.1. As can be seen, the normality was much improved: the distribution of the transformed data is approximately symmetric ($\text{skewness} = -0.079 \pm 0.068$) and less peaked

(kurtosis $= 0.421 \pm 0.135$).

Table 1 shows the reaction times of the four conditions. Since the distribution of reaction times is skewed we transformed the data using the natural logarithm (ln) before computing the means and standard errors (middle two columns). The last two columns show the reaction times transformed back to the normal time domain.

A one-way ANOVA showed that there was a significant difference between groups ($F(3, 1306) = 15.407$, $p < 0.001$). Levene's test indicated equal variances ($p = 0.644$).

A Tukey HSD post-hoc test revealed that the response time was significantly lower for the *flash eyes* action ($M = 854$ ms, $p = 0.006$) yet significantly higher for the *turn head* action ($M = 1033$ ms, $p = 0.003$) when compared to the *stay silent* condition (M = 944 ms). There was no significant difference between the *stay silent* condition and the *stop arms* action ($M = 916$ ms, $p = .829$).

Additionally, the mean response time for the *turn head* condition was significantly higher than both the *stop arms* ($p < 0.001$) and *flash eyes* ($p < 0.001$) conditions. There was, however, no significant difference between the *flash eyes* and *stop arms* conditions ($p = 0.071$). Post-hoc results are shown in Table 2. A bar chart visualising the means of the four conditions can be found in Figure 4.

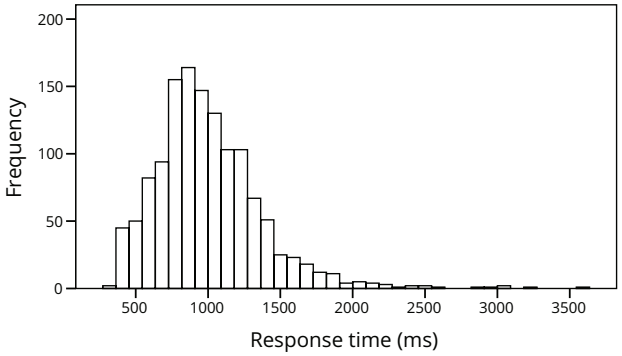


Figure 2. Original data

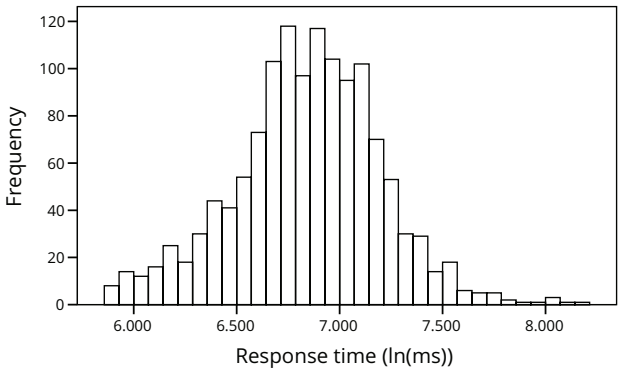


Figure 3. Histograms showing the distribution of response times

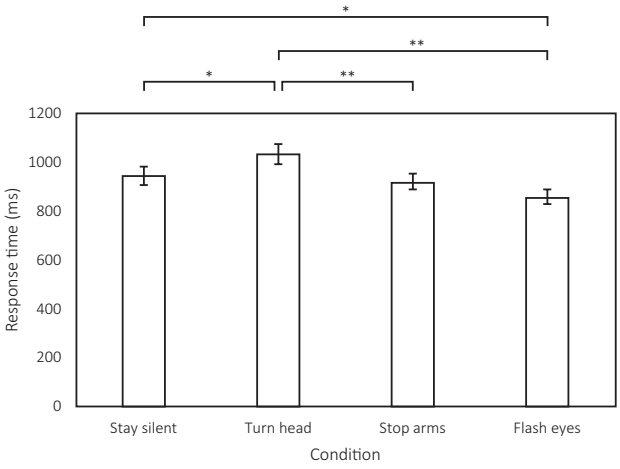


Figure 4. Means of the four conditions. Error bars represent 95% CI. Bars denoted with * differ at significance level < 0.01, bars with ** at significance level < 0.001.

Linear regression on the response times with trial number as the independent variable showed that these times did not decrease after sequential trials (*stay silent* $p = 0.759$; *turn head* $p = 0.224$; *flash eyes* $p = 0.368$), except for the *stop arms* condition ($p = 0.001$). For this last condition, response times decreased by 207 ms after 115 trials, as shown in Figure 5.

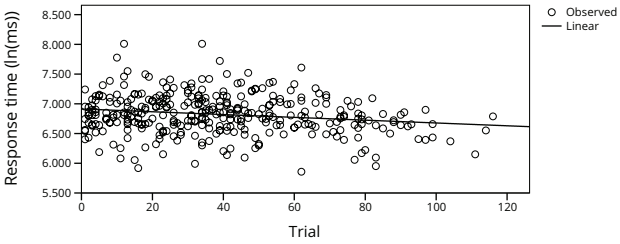


Figure 5. Scatter plot and fitted line of all response times in the *stop arms* condition.

3.2 Questionnaire results

The data gathered with the questionnaire (N = 19) was edited and analyzed using SPSS, in several steps.

The first part of the questionnaire was used as a confirmation of which cues were noticed by the participants. Cues that went unnoticed were excluded from the data.

Furthermore, the questionnaire included pairs of opposite questions, phrased positively and negatively, to avoid acquiescent bias. An example of such a pair is "...improved the flow of the conversation" and "...did not improve the conversation". Before analysis, negatively phrased questions had their answers mirrored.

Principle component analysis was used to identify the underlying factors and group the variables. After applying varimax rotation,

Table 1. Reaction times of the four conditions in the log-transformed and normal domain. SE is the standard error of sample mean. N is the number of turn yields (1310 in total).

Condition	N	Mean (ln(ms))	SE (ln(ms))	Mean (ms)	SE (ms)
Stay silent	331	6.85	.020	944	±19
Turn head	337	6.94	.019	1033	20/-19
Stop arms	334	6.82	.018	916	17/-16
Flash eyes	308	6.75	.020	854	±17

Table 2. Post-hoc test results of the response times

(I) condition	(J) condition	Mean difference (I-J, ln(ms))	SE (ln(ms))	Sig.
Stay silent	Turn head	-0.95	.027	.003
	Stop arms	.023	.027	.829
	Flash eyes	.091	.028	.006
Turn head	Stay silent	.095	.027	.003
	Stop arms	.118	.027	.000
	Flash eyes	.186	.028	.000
Stop arms	Stay silent	-.023	.027	.829
	Turn head	-.118	.027	.000
	Flash eyes	.068	.028	.071
Flash eyes	Stay silent	-.091	.028	.006
	Turn head	-.186	.028	.000
	Stop arms	-.068	.028	.071

three components were found with an Eigenvalue over 1, accounting for 35.1, 28.2 and 13.2 percent, respectively, of the total variance.

The rotated component matrix, shown in Table 3, shows which questions load on which components after rotation. Based on this data, the three components were named *Pleasant*, *Improvement* and *Noticeable*. Table 4 shows which questions make up which components.

After identifying the components, a one-way ANOVA on the combined questions showed that there was a significant difference between groups for the *Improvement* ($F(3, 292) = 8.998$, $p < 0.001$) and *Noticeable* ($F(3, 70) = 3.081$, $p = 0.033$) components, but not for the *Pleasant* component ($F(3, 218) = 0.602$, $p = 0.614$).

A Tukey HSD post-hoc test performed on the *Improvement* and *Noticeable* components showed that there were several significant differences between the means of the questionnaire responses. *Flash eyes* scored significantly higher on *Improvement* than both *stop arms* ($p < 0.001$) and *stay silent* ($p = 0.001$). Also, *stop arms* scored higher than *stay silent* on *Noticeable* ($p = 0.040$).

The post-hoc test results for the *Improvement* and *Noticeable* components can be found in Table 5 and 6, respectively. A graphical summary of all the components can be found in Figure 6.

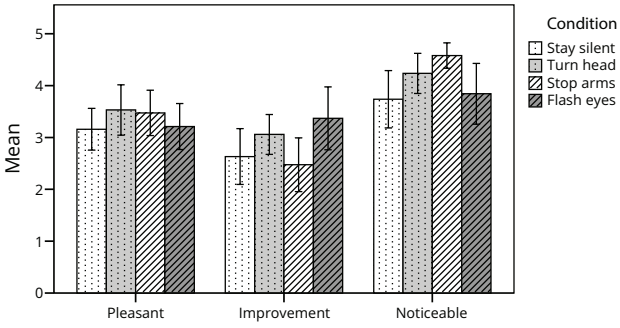


Figure 6. Means of the four conditions for every component. Error bars represent 95% CI.

4 DISCUSSION

The present study investigated different turn-yielding cues to be used by a robot in robot-human conversation. An experiment and questionnaire measured the performance and rating of the different cues. The results show that using a turn-yielding cue can lead to faster response times by the conversational partner compared to the baseline condition. One of the cues, namely *flash eyes*, produced the lowest response times and was rated higher on Improvement than the baseline condition and any other cue. The results, therefore, partially confirm the hypothesis that turn-yielding cues by a robot can improve robot-human conversation.

4.1 Different types of cues

The *flash eyes* cue lead to faster response times and had the highest *Improvement* rating by the participants. However, other cues showed different results. The *turn head* cue showed significantly longer response times compared to *staying silent*. Moreover, while the *stop arms* condition was rated as more noticeable than *staying silent*, there was no significant difference between the mean response times of these two cues.

There was a difference of 179 ms between the means of the response times for the *flash eyes* and *turn head* cues. A conclusion could be that while turn-yielding cues have the potential to lead to decreased response times, the type of cue matters a great deal.

Table 3. Rotated component matrix. Questions marked with * were mirrored.

Question	Component 1	Component 2	Component 3
...made it obvious it was my turn	.913	.094	.035
...had no clear meaning*	.868	.023	-.110
...did not improve the conversation*	.723	.459	.067
...improved the flow of the conversation	.703	.465	-.143
...was uncomfortable*	.074	.871	-.101
...was friendly	.142	.863	.155
...felt natural	.415	.560	-.096
...was hard to notice*	-.060	-.007	.986

Table 4. Components and related questions. Questions marked with * were mirrored.

Component 1, *Pleasant*	Component 2, *Improvement*	Component 3, *Noticeable*
...was uncomfortable*	...made it obvious it was my turn	...was hard to notice*
...was friendly	...had no clear meaning*	
...felt natural	...did not improve the conversation*	
	...improved the flow of the conversation	

4.2 Artificial cue

While a decrease in response time can be a hint that the cue improves the conversation, this does not necessarily have to be the case. Results from the questionnaire, however, were in line with the results from the experiment when it came to the *flash eyes* cue. It was seen as an improvement to the conversation and to have a clearer meaning when compared to the *stop arms* and *stay silent* cues.

Some anecdotal evidence from the experiment pointed the same way. Several participants remarked that they appreciated the *flash eyes* cue, one of them explaining "It signals that he is done, and that he won't interrupt me". Multiple participants also described the cue as "natural", which is interesting for an artificial cue that human conversational partners are unable to perform.

Thus, one of the interesting things here is that the cue with the lowest response time was an artificial cue, as opposed to the *turn head* and *stop arms* cues, which were based on literature from human-human interaction. There appears to be a difference between a human being using such cues and the NAO doing the same. This could have several causes. One possible cause is that the NAO did not perform the cue correctly, and therefore its meaning was unclear to the participants. Results from the questionnaire are inconclusive on this point: these cues were not rated significantly lower on this point, and their means center around "Neither agree nor disagree". Another reason could be that the participants found the cues with movement to be unexpected and therefore hesitated in their responses.

4.3 Movement cues

The cues that were based on movement, namely *turn head* and *stop arms*, showed worse performance compared to *flash eyes*, which did not involve movement. The movements made by the robot could be a source of distraction or hesitation for the participants, which could explain the longer response times.

Some anecdotal evidence from the experiment pointed this way. Some of the participants talked about the *turn head* and *stop arms* cues, explaining that they found many of the robot's movements to be distracting, and were sometimes confused as to the meaning of these movements. The data from the questionnaire shows that the *stop arms* cue was rated as significantly higher on the *Noticeable* component. Could it have been too noticeable, thereby distracting the participant?

Additionally, during the experiment it often seemed that when the NAO started moving, the participant hesitated to continue, preferring to wait to see where the robot was going with this. One of them remarked that he did not recognize the movement of *turn head* as a cue to start speaking, so instead he "just waited until it was done".

The movements could have simply been unexpected. Linear regression showed that for at least the *stop arms* cue, the mean response time decreased after subsequent trials, suggesting the participants were faster to respond and perhaps got used to the cue. Perhaps after longer interaction with the robot, this cue could have lead to response times similar to *flash eyes*.

Whether these findings are specific to the NAO robot is unclear, but fact is that this particular robot makes distinct sounds during movements and that it remains completely static outside of the performed cues. This could make movement cues highly salient by default.

4.4 Improvements to the experiment

A critical component of the experiment was accurately measuring the response time. The external microphone made it possible to relatively accurately and precisely measure the points at which the participant started speaking. However the beginning of the measurement, defined as the point at which the NAO stopped speaking, was harder to measure accurately. In the experiment, the timer started running after the NAO signalled it was done. However further investigation revealed that there is in fact a pause between the actual end of the sound and this signal, of around 225 ms on average. Though this issue could unfortunately not be avoided during this experiment, it could have an impact on the results. In practice it means that the turn-yielding cue could be performed sooner after speaking, possibly leading to a larger decrease in response times and an even stronger effect. Indeed, if we subtract 225ms form the reaction times for all non-verbal cues except the *stay silent* cue in Figure 4, we obtain a graph where all non-verbal cues lead to a reaction time improvement

Table 5. Post-hoc test results for the *Improvement* component

(I) condition	(J) condition	Mean difference (I-J)	SE	Sig.
Flash eyes	Turn head	.449	.193	.096
	Stop arms	.921	.188	.000
	Stay silent	.724	.188	.001
Turn head	Flash eyes	-.449	.193	.096
	Stop arms	.472	.193	.072
	Stay silent	.275	.193	.488
Stop arms	Flash eyes	-.921	.188	.000
	Turn head	-.472	.193	.072
	Stay silent	-.197	.188	.720
Stay silent	Flash eyes	-.724	.188	.001
	Turn head	-.275	.193	.488
	Stop arms	.197	.188	.720

Table 6. Post-hoc test results for the *Noticeable* component

(I) condition	(J) condition	Mean difference (I-J)	SE	Sig.
Flash eyes	Turn head	-.393	.319	.608
	Stop arms	-.737	.310	.091
(p ¡ .001)	Stay silent	.105	.310	.986
Turn head	Flash eyes	.393	.319	.608
	Stop arms	-.344	.319	.704
	Stay silent	.498	.319	.406
Stop arms	Flash eyes	.737	.310	.091
	Turn head	.344	.319	.704
	Stay silent	.842	.310	.040
Stay silent	Flash eyes	-.105	.310	.986
	Turn head	-.498	.319	.406
	Stop arms	-.842	.310	.040

compared to the *stay silent* cue. However, the *flash eyes* cue would still be most salient and the relative effectiveness of these cues remains the same.

5 CONCLUSIONS

The present study explored the use of turn-yielding cues by a robot. We found that such turn-yielding cues can improve both performance and user experience during human-robot conversation. These results on turn-yielding are in line with earlier findings that show that non-verbal cues can influence turn taking in conversations [2, 18]. Our study adds to earlier research by specifically focusing on the relative effect of turn-yielding cues and it shows that the type of cue is of importance for both performance and user experience.

An important question is how these conclusions are to be used in the development of socially assistive robots. Should one, for instance, always make use of an eye-flashing cue? It is clear that turn-yielding cues have the potential to improve a conversation, but in our study at most one cue was presented at a time (in addition to the stay silent cue). While the eye-flashing cue showed the most promise during this experiment, its meaning is, in general, ambiguous. Flashing LEDs are used to signal all sorts of events. In that sense the *turn head* and *stop arms* cues are much better, because they not only inform the observer about the timing of an event but also that the event is a turn-yield. So we expect that these cues are more useful in complex interactions. Finally, it would be interesting to see how these cues interact. A head turn could disambiguate a LED flash, so that in combination the turn-yield cues are effective and robust.

REFERENCES

[1] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher, 'Designing effective gaze mechanisms for virtual agents', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 705–714, (2012).

[2] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu, 'Conversational gaze aversion for humanlike robots', in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp. 25–32, (2014).

[3] Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke, 'Integrating vision and speech for conversations with multiple persons', in *International Conference on Intelligent Robots and Systems 2005*, pp. 2523–2528, (2005).

[4] Dan Bohus and Eric Horvitz, 'Multiparty turn taking in situated dialog: Study, lessons, and directions', in *Proceedings of the SIGDIAL 2011 Conference*, pp. 98–109, (2011).

[5] Joost Broekens, Marcel Heerink, and Henk Rosendal, 'Assistive social robots in elderly care: a review', *Gerontechnology*, **8**(2), 94–103, (2009).

[6] Sarah Chaytor and Uta Staiger, 'The future of healthcare in europe', *UCL European Institute*, (2011).

[7] Alex Colburn, Michael F. Cohen, and Steven Drucker, 'The role of eye

gaze in avatar mediated conversational interfaces', *Sketches and Applications, Siggraph'00*, (2000).
[8] Starkey Duncan, 'Some signals and rules for taking speaking turns in conversations.', *Journal of personality and social psychology*, **23**(2), 283, (1972).
[9] Maia Garau, Mel Slater, Simon Bee, and Martina Angela Sasse, 'The impact of eye gaze on communication using humanoid avatars', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 309–316, (2001).
[10] Agustín Gravano and Julia Hirschberg, 'Turn-yielding cues in task-oriented dialogue', in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 253–261, (2009).
[11] Jaap Ham, René Bokhorst, Raymond Cuijpers, David van der Pol, and John-John Cabibihan, 'Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power', in *Social Robotics*, 71–83, Springer, (2011).
[12] D. K. J. Heylen, Ivo Van Es, Anton Nijholt, and E. M. A. G. van Dijk, 'Experimenting with the gaze of a conversational agent', (2002).
[13] Anna Hjalmarsson, 'The additive effect of turn-taking cues in human and synthetic voice', *Speech Communication*, **53**(1), 23–35, (2011).
[14] Chien-Ming Huang and Bilge Mutlu, 'Modeling and evaluating narrative gestures for humanlike robots.', in *Robotics: Science and Systems*, (2013).
[15] Martin Johansson, Gabriel Skantze, and Joakim Gustafson, 'Head pose patterns in multiparty human-robot team-building interactions', in *Social Robotics*, 351–360, Springer, (2013).
[16] Adam Kendon, 'Some functions of gaze-direction in social interaction', *Acta psychologica*, **26**, 22–63, (1967).
[17] Chaoran Liu, Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita, 'Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction', in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, pp. 285–292, (2012).
[18] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita, 'Footing in human-robot conversations: how robots might shape participant roles using gaze cues', in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pp. 61–68, (2009).
[19] David G. Novick, Brian Hansen, and Karen Ward, 'Coordinating turn-taking with gaze', in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pp. 1888–1891, (1996).
[20] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, 'How quickly should communication robots respond?', in *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, pp. 153–160, (March 2008).
[21] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh, 'Where to look: a study of human-robot engagement', in *Proceedings of the 9th international conference on Intelligent user interfaces*, pp. 78–84, (2004).
[22] Adriana Tapus, Mataric Maja, and Brian Scassellatti, 'The grand challenges in socially assistive robotics', *IEEE Robotics and Automation Magazine*, **14**(1), (2007).
[23] UN, 'World population prospects, the 2012 revision', *New York: Department for Economic and Social Affairs*, (2013).
[24] Elisabeth T. Van Dijk, Elena Torta, and Raymond H. Cuijpers, 'Effects of eye contact and iconic gestures on message retention in human-robot interaction', *International Journal of Social Robotics*, **5**(4), 491–501, (2013).
[25] Roel Vertegaal, Robert Slagter, Gerrit Van der Veer, and Anton Nijholt, 'Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 301–308, (2001).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Robot Learning from Verbal Interaction: A Brief Survey

Heriberto Cuayáhuitl[1]

Abstract. This survey paper highlights some advances and challenges in robots that learn to carry out tasks from verbal interaction with humans, possibly combined with physical manipulation of their environment. We first describe what robots have learnt from verbal interaction, and how do they do it. We then enumerate a list of research limitations to motivate future work in this challenging and exciting multidisciplinary area. This brief survey points out the need of bringing robots out of the lab, into uncontrolled conditions, in order to investigate their usability and acceptance by end users.

1 INTRODUCTION

Intelligent conversational robots are an exciting and important area of research because of their potential to provide a natural language interface between robots and their end users. A learning conversational robot can be defined as an entity which improves its performance over time through verbally interacting with humans and/or other machines in order to carry out abstract or physical tasks in its (real or virtual) world. The vision of such kinds of robots is becoming more realistic with technological advances in artificial intelligence and robotics. The increasing development of robot skills presents boundless opportunities for them to perform useful tasks for and with humans. Such development is well suited to robots with a physical body because they can exploit their input and output modalities to deal with the complexity of public spatial environments such as homes, shops, airports, hospitals, etc. A robot learning from interaction, rather than a robot that does not learn, is particularly relevant because it is not feasible to pre-program robots for all possible environments, users and tasks. Even though many robotic systems can be scripted or programmed to behave just as expected, the rich nature of interaction with the physical world, or with humans, demands flexible, adaptive solutions to deal with dynamic, previously unknown, or highly stochastic domains. Therefore, robots should be able to refine their already learned skills over time and/or acquire new skills by (verbally) interacting with its users and its spatial environment. An emerging multidisciplinary community at the intersection of machine learning, human-robot interaction, natural language processing, robot perception, robot manipulation and robot gesture generation, among others, seeks to address challenges in realising such robots capable of interactive learning.

This paper will provide a brief survey on robots that learn to acquire or refine their verbal skills from example interactions using machine learning. Conversational robots that draw on hand-coded behaviours, or robots learning from non-verbal interaction [3, 14], are therefore considered out of scope here.

[1] Heriot-Watt University, United Kingdom, email: hc213@hw.ac.uk

2 ADVANCES

2.1 What have robots learnt from conversational interaction?

The following list of representative conversational robots shows a growing interest in this multidisciplinary field, see Figure 1.

- The mobile robot *Florence* is a nursing home assistant [20, 17]. The tasks of this robot include providing the time, providing information about the patient's medication schedule and TV channels, and motion commands such as go to the kitchen/bedroom. The learning task consists in inducing a dialogue strategy under uncertainty, where the actions correspond to physical actions (motion commands) and clarification or confirmation actions. The robot's goal is to choose as many correct actions as possible.
- Iwahashi's non-mobile robot with integrated arm+hand+head learns to communicate from scratch by physically manipulating objects on a table [11]. The tasks of this robot include (a) acquisition of words, concepts and grammars for objects and motions; (b) acquisition of the relationships between objects; and (c) the ability to answer questions based on beliefs. The robot's goal is to understand utterances and to generate reasonable responses from a relatively small number of interactions.
- The mobile robot *SmartWheeler* is a semi-autonomous wheelchair for assisting people with severe mobility impairments [19]. The task of the robot is to assist patients in their daily locomotion. The learning task is similar as in the Florence robot, the induction of a dialogue manager under uncertainty, but with a larger state space (situations). The robot's goal is to reduce the physical and cognitive load required for its operation.
- A mobile robotic forklift is a prototype for moving heavy objects from one location to another [25]. Example commands include going to locations, motion commands, and picking up and putting down objects. The learning task consists in understanding natural language commands in the navigation and object manipulation domain. The robot's goal is to ground natural language commands (mapping commands to events, objects and places in the world [18]) in order to output a plan of action.
- The humanoid robot *Simon* manipulates physical objects on a table from human teachers [2]. The task of the robot includes pouring cereal into bowls, adding salt to salads, and pouring drinks into cups. The learning task is to ask questions to human demonstrators from three different types: label queries (Can I do it like this?), demonstration queries (Can you show me how to do it?) , and feature queries (Should I keep this orientation?). The robot's goal is to ask as good questions as possible in order to achieve fast learning from physical demonstrations.
- A KUKA mobile platform with manipulator ensembles simple furniture [24]. The task of the robot is to assemble IKEA furniture such as tables based on STRIPS-like commands. The learning

Figure 1. Example learning conversational robots: (a) Florence nursebot [20], (b) Iwahashi's robot [11], (c) Kuka furniture assembler [24], (d) Nao giving directions [1], (e) Nao playing quizzes [7], (f) Simon robot learning from demonstrations [2], (g) James bartender robot [12], (h) Forklift robot [25], (i) SmartWheeler [19], (j) PR2 learning new words [15], (k) Gambit picking up objects [16], and (l) Cosero receiving verbal commands [21]. See text in Section 2.

tasks consists in learning to ground language and to train a natural language generator in order to ask for help to humans (by generating words from symbolic requests) when the robot encounters a failure situation. The robot's goal is to ensemble furniture as independently as possible and to ask for help when failures occurred.

- The torso robot *James* serves drinks to people in a pub [12]. The task of the robot is to approach customers in natural language, to ask for the drinks they want, and to serve the requested drinks. The learning task consists in inducing a dialogue manager for multi-party interaction. The robot's goal is to serve as correct drinks as possible based on socially acceptable behaviour due to the presence of multiple customers at once in the robot's view.
- The humanoid robot *NAO* has been used to play interactive quiz games [7, 6]. The robot's tasks include engaging into interactions, asking and answering questions from different fields, and showing affective gestures aligned with verbal actions. The learning task consists in inducing a dialogue strategy optimising confirmations and flexible behaviour, where users are allowed to navigate flexibly across subdialogues rather than using a rigid dialogue flow. The robot's goal is to answer correctly as much as possible and to ask as many questions as possible from a database of questions.
- The humanoid robot *NAO* has been used to give indoor route instructions [1]. The task of the robot is to provide directions, verbally and with gestures, to places within a building such as offices, conference rooms, kitchen, cafeteria, bathroom, etc., based on a predefined map. The learning task is to induce a model of engagement to determine when to engage, maintain or disengage an interaction with the person(s) in front of the robot. The robot's goal is to direct people to the locations they are looking for.
- The mobile robot *PR2* has been used to acquire new knowledge of objects and their properties [15]. The tasks of the robot include to spot unknown objects, to ask how unknown objects look like, and to confirm newly acquired knowledge. The learning task is to extend its knowledge base of objects via descriptions of their physical appearance provided by human teachers. The robot's goal is to answer questions of its partially known environment.
- The robot arm *Gambit* has been used to study how users users refer to groups of objects with speech and gestures. The tasks of the robot is to move indicated objects in a workspace, via verbal descriptions of object properties and possibly including gestures. The learning task is to understand user intentions without requiring specialized user training. The robot's goal is to select, as correctly as possible, the referred objects on the table.
- The mobile robot *Cosero* has been used in the RoboCup at home competition, which has won several of them in recent years [21]. The tasks of the robot include to safely follow a person, to detect an emergency from a person calling for help, to get to know and recognise people and serve them drinks, and to bring objects from one location to another. The learning task is to extend its knowledge of locations, objects and people. The robot's goal is to carry out tasks autonomously—provided in spoken language—as expected and in a reasonable amount of time.

ID	Dimension / Reference	[20]	[11]	[19]	[25]	[2]	[24]	[12]	[7]	[1]	[15]	[16]	[21]	ALL
01	Learning To Interpret Commands	1	1	1	1	1	1	1	1	1	1	1	1	12
02	Dialogue Policy Learning	1	0	1	0	1	0	1	1	0	0	0	0	5
03	Learning To Generate Commands	0	1	0	0	0	1	0	0	0	0	0	0	2
04	Learning To Engage	0	0	0	0	0	0	1	1	1	0	0	0	3
05	Grammar Learning	0	1	0	0	0	0	0	0	0	0	0	0	1
06	Flexible Interaction	0	0	0	0	0	0	0	1	0	0	1	1	3
07	Speech-Based Perception	1	1	1	0	1	0	1	1	1	1	1	1	10
08	Language Grounding	0	1	0	0	0	1	0	0	0	0	1	0	3
09	Speech Production	1	1	1	0	1	0	1	1	1	1	0	1	9
10	Multimodal Fussion	0	1	1	0	1	0	1	0	1	1	1	1	8
11	Multimodal Fission	0	1	1	0	0	0	1	1	1	0	0	1	6
12	Multiparty Interaction	0	0	0	0	0	0	1	0	1	0	0	0	2
13	Route Instruction Giving	0	0	0	0	0	0	0	0	1	0	0	0	1
14	Navigation Commands	1	0	1	1	0	1	0	0	0	0	0	1	5
15	Object Recognition and Tracking	0	1	0	1	1	1	1	0	0	1	1	1	8
16	Human Activity Recognition	0	0	0	0	0	0	0	0	1	0	0	1	2
17	Localisation and Mapping	1	0	1	1	0	0	0	0	0	0	0	1	4
18	Gesture Generation	0	0	0	0	1	0	0	1	1	1	0	1	5
19	Object Manipulation	0	1	0	1	1	1	1	0	0	0	1	1	7
20	Supervised Learning	0	1	0	1	0	1	1	1	1	0	1	0	7
21	Unsupervised Learning	0	0	1	0	0	0	0	0	0	0	1	0	2
22	Reinforcement Learning	1	0	1	0	0	0	1	1	0	0	0	0	4
23	Active Learning	0	0	0	0	1	0	0	0	0	0	0	0	1
24	Learning From Demonstration	0	0	0	0	1	0	0	0	0	1	0	1	3
25	Evaluation w/Recruited Participants	1	0	0	0	1	1	1	1	1	0	1	1	8
26	Evaluation in Noisy/Crowded Spaces	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1. Features of robots acquiring/using their verbal skills. While boolean values are rough indicators, real values are better indicators but harder to obtain.

2.2 How do conversational robots learn to interact?

Machine learning frameworks are typically used to equip robots with learning skills, and they differ in the way they treat data and the way they process feedback [13, 8]. Some machine learning frameworks addressed by previous related works are briefly described as follows:

- *Supervised learning* can be used whenever it comes to the task of classifying and predicting data, where the data consists of labelled instances (pairs of features and class labels). The task here is to induce a function that maps the unlabelled instances to labels. This function is known as a classifier when the labels are discrete and as a regressor when the labels are continuous. Conversational robots make use of classifiers to predict spatial description clauses [25], grounded language [11, 24], social states [12], dialogue acts [7], gestures [16], and engagement actions [1], among others.
- *Reinforcement Learning* makes use of indirect feedback typically based on numerical rewards given during the interaction, and the goal is to maximise the rewards in the long run. The environment of a reinforcement learning agent is represented with a Markov Decision Process (MDP) or a generalisation of it. Its solution is a policy that represents a weighted mapping from states (situations that describe the world) to verbal and/or physical actions, and can be found through a trial and error search in which the agent explores different action strategies in order to select the one with the highest payoff. This framework can be seen as a very weak form of supervised learning, where the impact of actions is rated according to the overall goal (e.g. fetching and delivering an object or playing a game). This form of learning has been applied to design the dialogue strategies of interactive robots using MDPs [12], Semi-MDP to scale up to larger domains [7], and Partially Observable MDPs to address interaction under uncertainty [20, 19].
- *Unsupervised learning* addresses the challenge of learning from unlabelled data. Since it does not receive any form of feedback, it has to find patterns in the data solely based on its observable features. The task of an unsupervised learning algorithm is thus to uncover hidden structure in unlabelled data. This form of machine learning has been used by [19] to cluster the observation space of a POMDP-based dialogue manager, by [12] to cluster social states for multiparty interaction, and by [16] to select features for gesture recognition tasks.
- *Active learning* includes a human directly within the learning procedure assuming three data sets: a small set of labelled examples, a large set of unlabelled examples, and chosen examples. The latter are built in an interactive fashion by an active learning algorithm who queries a human annotator for labels it is most uncertain of. This form of learning has been applied to *learning from demonstration* scenarios by [2] and closely related by [15, 21].

Other forms of machine learning that can be applied to conversational robots include transfer and multi-task learning, lifelong learning, and multiagent learning, among others [8, 4]. Furthermore, while a single form of learning can be incorporated into conversational robots, combining multiple forms of machine learning can be used to address perception, action and communication in a unified way. The next section describes some challenges that require further research for the advancement of intelligent conversational robots.

3 Challenges: What is missing?

Table 1 shows a list of binary features for the robots described above. These features are grouped according to language, robotics, learning, and evaluation. The lowest numbers in the last column indicate the dimensions that have received little attention. From this table, it can observed that the main demand to be addressed is conversational robots that interact with real people in uncontrolled environments rather than recruited participants in the lab. The research directions demanding further attention are briefly described as follows:

- **Noise and crowds:** most (if not all) interactive robots have been trained and tested in lab or controlled conditions, where no noise or low levels of noise are exhibited–see Table 1. A future direction concerning the whole multidisciplinary community lies in training and evaluating interactive robots in environments including people with real needs. This entails dealing with dynamic and varying levels of noise (from low to high), crowded environments on the move, distant speech recognition and understanding [26, 23] possibly combined with other modalities [5], and real users from the general population rather than just recruited participants.
- **Unknown words and meanings:** most interactive robots have been equipped with static vocabularies and lack grammar learning (see line 5 in Table 1), where the presence of unseen words lead to misunderstandings. Equipping robots with mechanisms to deal with the unknown could potentially make them more usable in the real world. This not only involves language understanding but also language generation applied to situated domains [9].
- **Fluent and flexible interaction:** when a robot is equipped with verbal skills, it typically uses a rigid turn-taking strategy and a predefined dialogue flow (see line 6 in Table 1). Equipping robots with more flexible turn-taking and dialogue strategies, so that people can say or do anything at any time, would contribute towards more fluent and natural interactions with humans [7].
- **Common sense spatial awareness:** most conversational robots have been equipped with little awareness of the dynamic entities and their relationships in the physical world (see lines 13 and 16 in Table 1). When a robot is deployed in the wild, it should be equipped with basic spatial skills to plan its verbal and non-verbal behaviour. In this way, spatial representations and reasoning skills may not only contribute to safe human-robot interactions but also with opportunities to exhibit more socially-acceptable behaviour. See [22, 10] for detailed surveys on social interactive robots.
- **Effective and efficient learning from interaction:** interactive robots are typically trained in simulated or controlled conditions. If a robot is to interact in the wild, it should be trained with such kinds of data. Unfortunately, that is not enough because moving beyond controlled conditions opens up multiple challenges in the way we train interactive robots such as the following:
  - robot learning from unlabelled or partially labelled multimodal data (see lines 21 and 23 in Table 1) should produce safe and reasonable behaviours;
  - altering the robot's behaviour, even slightly, should be straightforward rather than requiring a substantial amount of human intervention (e.g. programming);
  - inducing robot behaviours should exploit past experiences from other domains rather than inducing them from scratch; and
  - learning to be usable and/or accepted by people from the general population is perhaps the biggest challenge.

4 Conclusion

Previous work has shown the increase in multidisciplinary work to realise intelligent conversational robots. Although several challenges remain to be addressed by specialised communities, addressing them as a whole is the end-to-end challenge that sooner or later it has to be faced. This challenge involves two crucial actions with little attention so far (a) to bring robots out of the lab to public environments, and (b) to demonstrate that they are usable and accepted by people from the general public. We hope that the topics above will encourage further multidisciplinary discussions and collaborations.

REFERENCES

[1] Dan Bohus, Chit W. Saw, and Eric Horvitz, 'Directions robot: In-the-wild experiences and lessons learned', in *AAMAS*, (2014).

[2] Maya Cakmak and Andrea Lockerd Thomaz, 'Designing robot learners that ask good questions', in *HRI*, (2012).

[3] Sonia Chernova and Andrea Lockerd Thomaz, *Robot Learning from Human Teachers*, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2014.

[4] Heriberto Cuayáhuitl and Nina Dethlefs, 'Dialogue systems using online learning: Beyond empirical methods', in *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Comm.*, (2012).

[5] Heriberto Cuayáhuitl, Lutz Frommberger, Nina Dethlefs, Antoine Raux, Matthew Marge, and Hendrik Zender, 'Introduction to the special issue on machine learning for multiple modalities in interactive systems and robots', *TiiS*, **4**(3), (2014).

[6] Heriberto Cuayáhuitl and Ivana Kruijff-Korbayová, 'An interactive humanoid robot exhibiting flexible sub-dialogues', in *NAACL-HLT*, (2012).

[7] Heriberto Cuayáhuitl, Ivana Kruijff-Korbayová, and Nina Dethlefs, 'Nonstrict hierarchical reinforcement learning for interactive systems and robots', *TiiS*, **4**(3), 15, (2014).

[8] Heriberto Cuayáhuitl, Martijn van Otterlo, Nina Dethlefs, and Lutz Frommberger, 'Machine learning for interactive systems and robots: A brief introduction', in *MLIS*. ACM ICPS, (2013).

[9] Nina Dethlefs and Heriberto Cuayáhuitl, 'Hierarchical reinforcement learning for situated natural language generation', *Natural Language Engineering*, **FirstView**, (12 2014).

[10] Terrence Fong, Illah R. Nourbakhsh, and Kerstin Dautenhahn, 'A survey of socially interactive robots', *Robotics and Autonomous Systems*, **42**(3-4), 143–166, (2003).

[11] Naoto Iwahashi, 'Robots that learn language: Developmental approach to human-machine conversations', in *International Workshop on the Emergence and Evolution of Linguistic Communication, EELC*, (2006).

[12] Simon Keizer, Mary Ellen Foster, Zhuoran Wang, and Oliver Lemon, 'Machine learning for social multiparty human-robot interaction', *TiiS*, **4**(3), 14, (2014).

[13] V Klingspor, Y Demiris, and M Kaiser, 'Human robot communication and machine learning', *Applied A.I.*, **11**, 719–746, (1997).

[14] Jens Kober, J. Andrew Bagnell, and Jan Peters, 'Reinforcement learning in robotics: A survey', *I. J. Robotic Res.*, **32**(11), 1238–1274, (2013).

[15] Evan Krause, Michael Zillich, Thomas Williams, and Matthias Scheutz, 'Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues', in *AAAI*, (2014).

[16] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox, 'Learning from unscripted deictic gesture and language for human-robot interactions', in *AAAI*, (2014).

[17] Michael Montemerlo, Joelle Pineau, Nicholas Roy, Sebastian Thrun, and Vandi Verma, 'Experiences with a mobile robotic guide for the elderly', in *AAAI*, (2002).

[18] Raymond J. Mooney, 'Learning to connect language and perception', in *AAAI*, (2008).

[19] Joelle Pineau and Amin Atrash, 'Smartwheeler: A robotic wheelchair test-bed for investigating new models of human-robot interaction', in *AAAI Spring Symposium on Socially Assistive Robotics*, (2007).

[20] Nicholas Roy, Joelle Pineau, and Sebastian Thrun, 'Spoken dialogue management using probabilistic reasoning', in *ACL*, (2000).

[21] Max Schwarz, Jörg Stückler, David Droeschel, Kathrin Gräve, Dirk Holz, Michael Schreiber, and Sven Behnke, 'Nimbro@home 2014 team description', in *RoboCup @ Home League*, (2014).

[22] Luc Steels, 'Social learning and verbal communication with humanoid robots', in *Humanoids*, (2001).

[23] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, 'Convolutional neural networks for distant speech recognition', *IEEE Signal Process. Lett.*, **21**(9), 1120–1124, (2014).

[24] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy, 'Asking for help using inverse semantics', in *RSS*, (July 2014).

[25] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy, 'Understanding natural language commands for robotic navigation and mobile manipulation', in *AAAI*, (2011).

[26] Jean-Marc Valin, François Michaud, Jean Rouat, and Dominic Létourneau, 'Robust sound source localization using a microphone array on a mobile robot', in *IROS*, (2003).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Embodiment, emotion, and chess: A system description

Christian Becker-Asano and **Nicolas Riesterer** and **Julien Hué** and **Bernhard Nebel**[1]

Abstract. We present a hybrid agent that combines robotic parts with 3D computer graphics to make playing chess against the computer more enjoyable. We built this multimodal autonomous robotic chess opponent under the assumption that the more life-like and physically present an agent is the more personal and potentially more effective the interaction will be. To maximize the life-likeness of the agent, a photo-realistic animation of a virtual agent's face is used to let the agent provide verbal and emotional feedback. For the latter an emotion simulation software module has been integrated to drive the agent's emotional facial expressions in parallel to its verbal utterances.

1 Introduction

Chess has been called the "Drosophila of artificial intelligence" [1] meaning that in the same way as the drosophila melanogaster has become the model organism for biological research, chess served at least for many years as a standard problem for artificial intelligence research. When in 1997 Garry Kasparov, who was ranked first at that time, lost against IBM's supercomputer "Deep Blue" [10], this problem was assumed to be solved and chess engines would nowadays outclass the best players. Altogether this triggered researchers to shift their attention to other games, such as Go. Today, for a casual chess player it can be rather frustrating to play against the computer, because he or she will lose most of the times and the computer moves its pieces with seemingly no hesitation.

Recently it was found, however, that different embodiments of the computer opponent change a human chess player's motivation to engage in a game of computer chess. These attitude changes are rooted in the humans' tendency to treat machines as social actors and this effect seems to be stronger the more human-like the machine is designed to appear [16]. With our development of the hybrid chess-playing agent MARCO, the Multimodal Autonomous Robotic Chess Opponent, we aim to investigate this research question.

The remainder of the paper is structured as follows. After discussing related work in the next section, our general motivation is explained and two research questions are introduced. Then the Elo rating will be explained together with how the employed chess engine evaluates board positions. Subsequently, MARCO's hardware components are detailed, before the interconnection of its software components is laid out. Then, the complete system is explained. Finally, we present our ideas concerning experimental protocols for evaluating MARCO. We conclude this paper with a general discussion.

[1] Artificial Intelligence Lab, University of Freiburg, 79110 Freiburg, Germany, email: [basano,riestern,hue,nebel]@cs.uni-freiburg.de

2 Related work

This section describes research projects involving chess playing robots [15, 18, 13]. They aim to answer different research questions and, therefore, they employ systems of different size and complexity.

"Gambit" is a good example for an engineer's solution to an autonomous chess-playing robotic system [15]. With their "robot manipulator system" the authors created a "moderate in cost" (i.e. 18K USD) manipulator that is able to play chess with arbitrary chess sets on a variety of boards without the need to model the pieces. Although their system does not have any anthropomorphic features, it includes a "natural spoken language interface" to communicate with the human opponent. Most importantly, "Gambit" tracks both the board and the human opponent in real time so that the board does not need to be fixed in front of the robot. With its available six degrees of freedom (DoF) and the USB camera mounted on top of its gripper the robot arm reliably grasps a wide array of different chess pieces, even if they are placed poorly. In result, it outperformed all robotic opponents at the 2010 AAAI Small Scale Manipulation Challenge. Unfortunately, no data on human players' enjoyment is available.

In contrast to the remarkable technical achievements behind the development of "Gambit", the "iCat" from Philips was combined with a DGT chess board to investigate the influence of embodiment on player enjoyment in robotic chess [13]. The authors conducted a small-scale empirical trial with the emotional iCat opponent either presented in its virtual or robotic form. Using a modified version of the GameFlow model [20], it was found that overall the virtual version is less enjoyable than the robotic one. A subsequent long term study [14] with the robotic iCat playing chess repeatedly against five children showed, however, that these children lost interest in the robot. Presumably, iCat's complete lack of any manipulation capability together with its cartoon-like appearance let the children ignore the robot completely after the initial curiosity is satisfied.

Similar to our approach, Sajó et al. [18] present a "hybrid system" called "Turk-2" that consists of a "mechanically simple" robot arm to the right of the human player and a rather simple 2D talking head presented on a computer display. "Turk-2" can analyze three emotional facial expressions, namely *sad*, *neutral*, and *happy*, and additional image processing enables the system to monitor the chess board. Interestingly, the authors decided to artificially prolong the system's "thinking time", details of which are unfortunately not reported. The transitions between the talking head's facial expressions *neutral*, *sad*, *happy*, and *bored* are controlled by a state machine that takes the human's emotion (as derived from its facial expression) and the game state into account. Similar to our approach, the talking head will change into a bored expression after some time without input has passed. An empirical study on the effect of the presence of the talking head revealed that without the talking head the players mostly ignored the robotic arm to the right of them, even when it was mov-

ing. With the talking head in front of them, however, the players not only looked at the talking head but also started smiling and laughing.

Regarding the effects of a virtual agent's facial expression of emotions on human performance in a cognitive task, an empirical trial resulted in no significant differences [8]. In addition, the study showed that for such a serious task it made no difference, if the agent's emotions were generated based on a set of hard-coded rules or by making use of a sophisticated and complex emotion simulation architecture. The authors speculate that a less cognitively demanding and more playful task might be better suited to search for such effects.

A prototype of the MARCO system has been demonstrated recently at an international conference [17] and, although conference attendees clearly enjoyed playing and loosing against the agent, several opportunities to improve the system were mentioned. The most noticeable deficiency seemed to be the use of a much too small display for presenting the agent's virtual face. Accordingly, our system now employs a much bigger display.

3 Motivation and research questions

These previous results in combination motivated us to include the following features in MARCO, our Multimodal, Autonomous, Robotic Chess Opponent:

1. A low-cost robotic arm that enables MARCO to autonomously move the chess pieces instead of having to rely on the human opponent's assistance (as in [13])
2. A custom built, robotic display presenting a highly anthropomorphic virtual agent's head to realize a hybrid embodiment combining the best of both worlds, cp. [13, 18]
3. A flexible software architecture that relies on an established emotion simulation architecture as one of its core modules (following up on [8])

The resulting MARCO system will help answering research questions that are motivated by the previous work presented above:

1. Is it more enjoyable to play chess against the robotic arm with or without the virtual agent?
2. Is it more enjoyable to play against the hybrid agent (i.e. the robotic arm with the virtual agent) when the agent expresses emotions as compared to when it remains equally active but emotionally neutral?
3. Is the most human-like and emotional agent evaluated as more social/mindful than the less complex/human-like versions of it? Does this subjective evaluation depend on how experience the human chess player is?

The first question will provide a baseline for the hardware components of our system and will be compared with those reported in [18] with regard to "Turk-2". It is not taken for granted that a more complex system will always be preferable to a simpler system from the perspective of a human player. The second question, however, is targeting the role that artificial emotions might or might not play and it is motivated by previous results [8]. Finally, MARCO allows us to tackle systematically the general question of how and when "mindfullness" is ascribed to machines [16].

4 Background and Preliminaries

4.1 Elo rating

The skill of chess players is usually measured in terms of a single integer value, the so-called Elo Rating [12]. It represents the relative strength of a player, the higher the better, and it increases or decreases with his or her chess match results. Currently, ELO rating in chess goes from 1000 (complete beginner) to 2880 (Magnus Carlsen World Champion).

Differences in the evaluations of our system might correlate with or even depend on the ELO ratings of the human players. In addition, our system might be used as a virtual coach for novice players to improve their chess skills and the ELO rating provides a standard means to compare player strength before and after training.

4.2 Chess Engine

Computer chess engines evaluate the board position using an alpha-beta algorithm with a depth d given as parameter based on a number of criteria like: pieces left on the board, activity of these pieces, security of the king, etc. The greater the depth the more precise is the evaluation. The position evaluation function results in a real number e ranging from $[-\infty, +\infty]$ where 0 means that the position is equal, $-\infty$ that black is winning and $+\infty$ that white is winning. A $+1$ valuation roughly represents the advantage equivalent to a pawn, $+3$ to a knight or a bishop, and so on according to the standard valuation of chess pieces.

We denote by $e_{t,d}$ the evaluation given by the chess engine at move t with depth d. We write e when it is clear from the context. In practice, once $|e| \geq 5$ the game is more or less decided.

Our first prototype [17] was based on the TSCP chess engine [2] for its simplicity and in order to make our results comparable to previous work on the iCat playing chess [13], for which the same engine was used. The communication between the user and the TSCP chess engine is handled by the XBoard Chess Engine Communication Protocol [3]. Originally implemented as a means to facilitate communication between the GNU XBoard Chess GUI and underlying chess engines, this plain text protocol allows for easy information exchange in a human readable form.

Our modular software architecture allows us, however, to plug in other chess engines. The more advanced Stockfish chess engine [4], for example, would allow us to adjust the strength of MARCO's play dynamically.

5 Hardware components

The complete setup is presented in Figure 1. The hardware used comprises a custom designed, 15.6 inch pan-tilt-roll display to present the virtual agent's face, a robotic arm to the right of the agent to move the chess pieces, and a digital chess board (DGT USB Rosewood) with a chess clock. Each of these components will be described next.

5.1 The pan-tilt-roll agent display

The pan-tilt-roll display component features a 15.6 inch upright TFT LCD display with a physical resolution of 1920×1080 pixels and 18bit color depth, cp. Fig. 1. It is positioned opposite of the human player to give the impression of the virtual agent overlooking the complete chess board. Three Dynamixel AX-12A servos (cp. Fig. 2(a)) are connected to USB2Dynamixel interface to allow for control over the display's orientation during the game along all three axes. Thereby, for example, the agent can follow its own arm's movements dynamically as presented in Fig. 2(b).

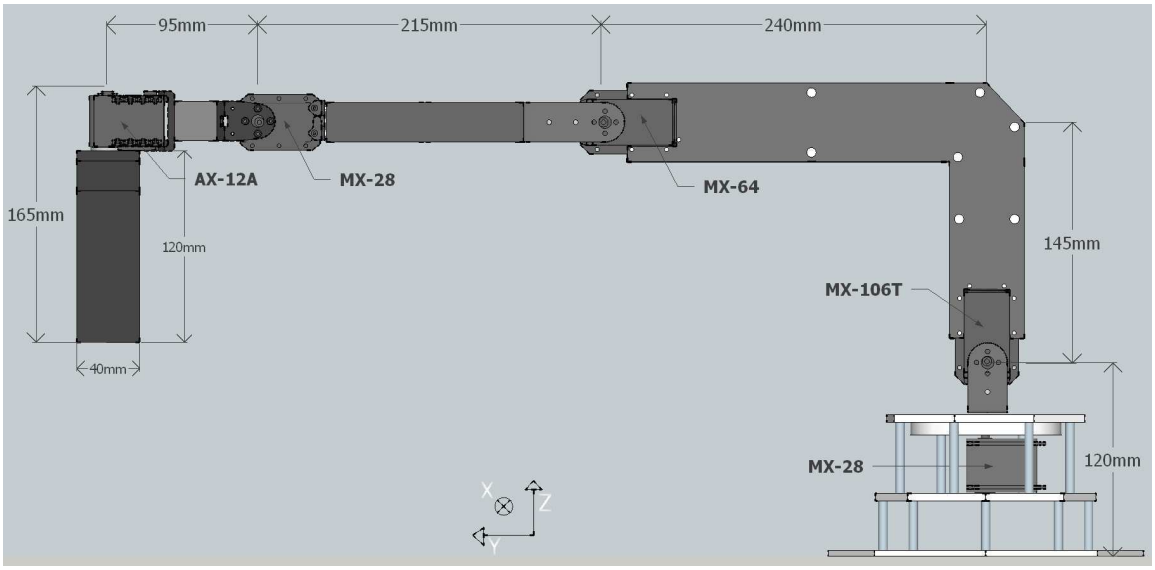
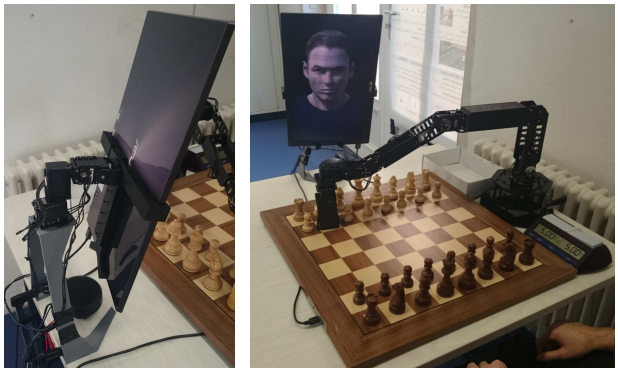


Figure 3. A schematic of the robotic arm with annotations of link lengths and Dynamixel servos used for each joint position

Figure 1. The pan-tilt-roll agent display, the robotic arm, and the digital chess board together with the digital chess clock



(a) Detail view (b) Front view

Figure 2. Pan-tilt-roll mount of the 15.6 inch display presenting the virtual agent's face

5.2 The robotic arm

The hybrid agent's robotic arm is a modification of the "WidowX Robotic Arm Kit Mark II" [5] available from Trossen Robotics. Apart from the rotational base all other parts needed to be extended to allow the agent to pick-and-place all pieces on any of the 64 squares of the board. The upper arm was extended to measure $240mm$, the forearm to measure $215mm$ and the gripper needed to be prolonged to $120mm$ (cp. Fig. 3). These extensions for the arm as well as the extra parts to realize the display mount were printed with a MakerBot 3D printer. Five Dynamixel servos move the robot's arm, cp. Fig. 3. For the base and wrist two MX-28 servos are used. An MX-64 servo moves the robot's elbow and an MX-106 servo its shoulder. The modified gripper is opened and closed by an AX-12A servo, cp. Fig. 4. It can reliably pick-and-place all Staunton chess pieces on the DGT board regardless of their height or size.

5.3 The DGT digital chess board

The DGT chess board is a wooden board with standard Staunton pieces and $55mm \times 55mm$ squares. Each piece is equipped with a unique RFID chip that makes it recognizable. The board is con-

(a) Open state (b) Closed state

Figure 4. The two states of the robot's custom designed gripper picking up a white bishop

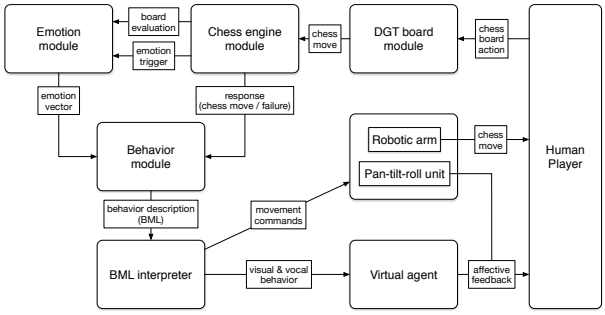


Figure 5. An outline of the software modules and their connections

nected to the computer with a USB cable, and it transmits the position in FEN format to the DGT board module every time a change is performed.

6 Software components

Except for the external MARC framework (see Section 6.3), all components are implemented in C++ using Qt5 [7] in combination with the Robot Operating System (ROS; [6]) to achieve a modular design and cross-platform functionality. The hardware components (i.e. the DGT chess board and the Dynamixel servos) are encapsulated into ROS nodes to establish a flexible communication infrastructure.

6.1 Overview of system components

The following five main software components can be distinguished, which are connected by the ROS message protocol (cp. Fig. 5):

- A DGT board module to detect moving pieces on the physical chess board
- A Chess engine model for position evaluation and chess move calculation
- An Emotion module to simulate MARCO's emotions
- A Behavior module to integrate the chess move with emotional states into a behavior description
- A Behavior Markup Language (BML) Interpreter to prepare the multimodal realization of the behavior
- Robotic components to move the chess pieces on the board and control the virtual agent's pan-tilt-roll unit
- The MARC framework to create the agent's visual appearance on the display

When the human player (cp. Fig. 5, right) performs her move, the DGT board module recognizes the change on the board, derives the move information by comparing the current board configuration with the previous one, and sends this information to the chess engine module. Here, the chess move is verified for correctness and either (1) a failure state, or (2) the chess engine's move is transmitted as MARCO's response to the behavior model. The board evaluation function of the chess engine also provides the emotion module with input. After the emotion module integrated the board evaluation into the agent's emotion dynamics (see Section 6.2), it concurrently updates the behavior module with a vector of emotion intensities. The behavior module integrates the emotional state information with the move calculation into a behavior description in BML [21]. This description is then interpreted by the BML interpreter to drive the virtual agent's visual and vocal behavior as well as the robotic component's actions. While the robotic arm starts to execute the agent's chess move, the pan-tilt-roll unit moves the display to realize affective feedback in combination with the virtual agent's facial expressions.

6.2 Deriving emotional states

The emotion module (cp. Fig. 5) comprises the WASABI Affect Simulation Architecture [9] to simulate the agent's dynamically changing emotional states. As input WASABI needs *valenced impulses* and expectation-based emotions (e.g., *surprised* and *hope*) need to be *triggered* before they can gain positive intensity.

6.2.1 Emotion dynamics

WASABI is based on the idea that emotion and mood are tightly coupled. The term "mood" refers to a relatively stable background state, which is influenced by emotion arousing events, but changes much more slowly as compared to any emotional state. An "emotion", in contrast, is a short-lived psychological phenomenon that more directly impacts behavior than a mood does.

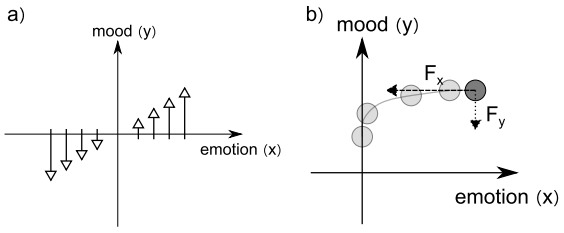


Figure 6. The emotion dynamics of WASABI with (a) the influence of emotional valence on mood, and (b) the effect of the two independent mass-spring systems on the development of the agent's emotional state over time (indicated by the half-transparent circles)

Taking these differences and commonalities as cue, WASABI simulates the positive and negative effects that emotional valence has on

mood, cf. Fig. 6a. In addition, mood and emotion are driven back to zero by two forces independently exerted from two mass-spring systems. Notably, the respective spring constants are set such that the resultant force F_x is always greater that the resultant force F_y, because emotions are longer lasting than mood, cp. Fig. 6b.

MARCO's emotional state as represented in Fig. 6b by the circles is updated with 50Hz letting it move through the space over time. The x and y values are incessantly mapped into PAD space to allow for categorization in terms of emotion labels (cp. Fig. 7; see also [9]).

This dynamic process is started by the arrival of a *valenced impulse* from outside of WASABI that instantaneously changes the emotion value (x) either in the positive or negative direction. How these impulses are derived from the progression of the game is described next.

6.2.2 *Valenced impulses*

The chess engine module continuously calculates board evaluations e_t (at times t during the game). These are converted into *valenced impulses* $val(e_t)$ according to Equation 1.

$$val(e_t) = k \times tanh\left(\frac{e_t}{r}\right) \tag{1}$$

Here, k is a scaling factor and by increasing the denominator $r \in [1, \infty]$ the skewness of the hyperbolic tangent is reduced until a quasi-linear mapping ($val(e_t) = k \times e_t$) is achieved. The hyperbolic tangent is introduced to let us emphasize small values of e_t relative to bigger values of e_t.

For example, choosing $k = 50$ and $r = 2$:

$$\begin{aligned} val(e_t) = 50 \times tanh\left(\frac{e_t}{2}\right) \in (2.5, 25], \\ \forall e_t \in \{x \in \mathbb{R} \mid 0.1 \leq x < 1.1\} \end{aligned} \tag{2}$$

Thus, with these constants any value of e_t between 0.1 and 1.1 results in a weak to medium valenced impulse. Observe that $|val(e_t)| \cong 50, \forall e_t \in \{x \in \mathbb{R} \mid |x| > 5\}$, meaning that a winning (or loosing) board configuration results in the maximum impulse of 50 (or minimum impulse of -50, respectively).

Depending on who plays white, the sign of the scaling factor k is adjusted as to map favorable board positions for MARCO to positively valenced impulses and vice versa. That is, if MARCO plays white k is positive, otherwise it is negative. For the time being, MARCO always plays white letting it perform the first half-move.

Inside the emotion module the *valenced impulses* drive the concurrent simulation of the agent's emotion dynamics. In summary, a positive (negative) impulse has the short term effect of increasing (decreasing) the agent's *emotional valence*, which in turn influences the agent's *mood* in the same direction as a long term effect. A simple mathematical transformation into *pleasure* ($P = \frac{x+y}{2}$) and *arousal* ($A = |x|$) values is performed and the emotion module then uses the PAD space (cf. Fig.7) to categorize the agent's emotional state in terms discrete emotions and their intensities. The *dominance* value is changed in accordance with whether it is MARCO's turn ($D = 1$) or not ($D = 0$). Finally, the resulting set of emotions with positive intensities are transmitted to the behavior module.

6.2.3 *Mapping onto discrete emotions*

In its default configuration, WASABI simulates the primary emotions *annoyed*, *angry*, *bored*, *concentrated*, *depressed*, *fearful*, *happy*, *sad*, and *surprised* as well as the secondary emotions *relief*, *fears-confirmed*, and *hope*; cp. Fig. 7. Five of these 12 emotions (*fearful*, *surprised*, *relief*, *fears-confirmed*, and *hope*) rely on an agent's ability to build expectations about future events, i.e., they are so-called *prospect-based emotions*. For example, one is only surprised about an event, if it is contrary to one's previous expectations, or one fears future events, only if one has reason to expect that bad event is about to happen [9]. Accordingly, in WASABI each of these emotions is configured with zero *base intensity* and needs to be *triggered* (cp. "emotion trigger" in Fig. 5) to give them a chance to gain positive intensity.

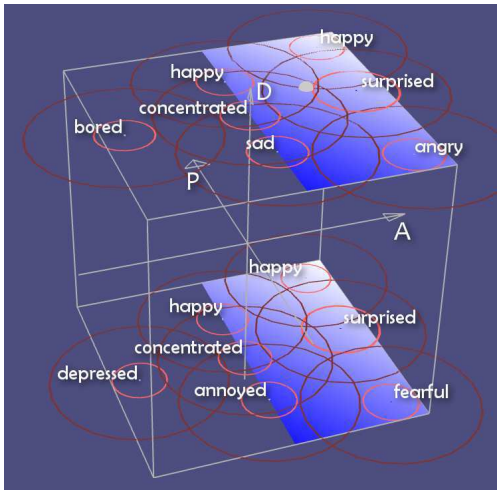


Figure 7. The PAD-space of primary and secondary emotions in WASABI. The primary emotions are distributed as to cover all areas of PAD space. For each of them an activation threshold (outer ring) and a saturation threshold (inner ring) is defined. The two shaded areas represent the distribution of the secondary emotion *hope* in the dominant and submissive subspace, after it was triggered. The grey half-sphere represents MARCO's dynamically changing emotional state. Thus, in this example MARCO would be mildly *happy*, a bit *concentrated*, and quite *hopeful*. If surprise were triggered as well in this moment, MARCO would also be *surprised* to a certain extend.

With respect to chess, our system evaluates the available moves for its opponent. MARCO is able to realize, whenever its last move was less good than previously evaluated, because at time t the evaluation reaches one level deeper into the search tree than at time $t-1$. Accordingly, MARCO might start to fear that the human opponent realizes her opportunity as well. If the evaluation of the situation after the opponent's move is stable, then MARCO's fears are confirmed: the opponent made the right move. On the other hand, if the evaluation comes back to what it was before, i.e., before MARCO made its last move, then the opponent missed the opportunity and MARCO is relieved. The evaluation can be in between these two values and in that case, the agent is neither relieved nor sees its fears confirmed. Nevertheless, the emotion module still receives the negative *valenced impulse* derived from the drop. Formally, Table 1 provides details on how the changing evaluations trigger prospect-based emotions in WASABI.

Notably, the value e_t represents the future directed evaluation of the situation from the robot's perspective. For example, the formula $e_{t-1} - e_t > \epsilon$ lets the behavior trigger *fear* whenever a significant drop in the evaluation function appeared from the previous move to

trigger	*if..*
fear	$e_{t-1} - e_t > \epsilon$
surprise	$\lvert e_{t-1} - e_t \rvert > \epsilon$
fears-confirmed	$fear_{t-1} \wedge (e_{t-1} - e_t < \epsilon)$
hope	$e_{t,d} - e_{t,d-2} > \epsilon$
relief	$fear_{t-1} \wedge (e_t - e_{t-2} < \epsilon)$

Table 1. The conditions under which the prospect-based emotions are triggered in WASABI based on the changes of evaluations over time with ϵ and depth d as custom parameters

Figure 8. The virtual agent expressing *anger*, *neutral*, and *joy* (left to right)

the current one. That is, MARCO realizes at time t that the future seems much worse than evaluated before (in time $t-1$). If subsequently, after the next half-move in $t+1$, the value e_{t-1} turns out to have been correct in the light of the new value e_t (or the situation got even worse than expected), then *fears-confirmed* will be triggered. On the contrary, if it turned out to be much better than expected, *relief* will be triggered. *Surprise* is always triggered when the evaluation changes significantly from one half-move to the next. Finally, *hope* is triggered whenever not taking the full depth of the search tree into account would mean that the key move in the position is hard to reach (requires a computation at depth at least d).[2]

6.2.4 *The emotion vector as input for the behavior module*

It is important to note that, in addition to an emotion being triggered, the *pleasure*, *arousal*, and *dominance* (PAD) values driven by the emotion dynamics must be close enough to that emotion for it to become a member of the *emotion vector* with positive intensity, cp. Fig. 7. Thus, although *surprise* will always be triggered together with *fear*, they will not always both be present in the *emotion vector*, because they occupy different regions in PAD space.

From the *emotion vector* the emotion with the highest intensity is compiled into the BML description driving the MARC framework. The agent comments on particular events like, for example, complimenting the player after it lost a game or stating that the position is now simplified after exchanging the queen.

6.3 The virtual agent provided by the MARC framework

The MARC framework [11] is used to animate the virtual agent, which is presented on the 15.6 inch pan-tilt-roll display facing the human player. The emotional facial expressions (see Fig. 8 for examples) that are provided as part of the BML description are combined inside the MARC framework to create lip-sync animations of emotional verbal utterances. Thanks to the integration of the open-source text-to-speech synthesis OpenMARY [19] the agent's emotion also influences the agent's auditory speech.

[2] An evaluation function is usually set up to an even number, thus the last level of the search tree equals the last two half-moves.

7 Conclusions and future work

This paper detailed the software and hardware components behind MARCO, a chess playing hybrid agent equipped with a robotic arm and a screen displaying a virtual agent capable of emotional facial expressions. A first prototype of the system was demonstrated at an international conference [17] and the experiences gained let to improvements both on concerning the hard- and software components.

Although a limited set of concrete agent behaviors has proven to be fun for the conference participants, we still need to design many more of them. For example, we need to decide which kind of comments are to be given with which timing during the game and how virtual gaze and robotic head movements are to be combined to give the impression of a believable, hybrid agent.

In order to answer the initially stated two research questions, we plan to conduct a series of empirical studies. At first, one group of participants will play against MARCO with the pan-tilt display turned off. Nevertheless, the invisible agent's comments will remain audible in this condition. In the second condition, another group of participants will play against MARCO with an unemotional agent presented on the robotic display. For the third condition, a group of participants will play against the WASABI-driven agent. In all three conditions, player enjoyment will be assessed using the GameFlow [20] questionnaire and video recordings of the human players will be analyzed inspired by [18]. We expect to find significant differences between conditions with the most complete setup (condition three) being most fun for the players.

Nass and Moon claim that imperfect technologies mimicking human characteristics might even increase "the saliency of the computer's 'nonhumanness'." [16, p. 97] In line with their ideas and in addition to the approach outlined above, we plan to compare human-human interaction with human-agent interaction when competing in chess to measure and incessantly improve MARCO's level of human-likeness. This will help to understand how human behavior might be split into computationally tractable components and then realized in robotic agents to improve human-computer interaction.

REFERENCES

[1] http://aitopics.org/topic/chess.
[2] http://www.tckerrigan.com/Chess/TSCP.
[3] http://www.gnu.org/software/xboard/engine-intf.html.
[4] https://stockfishchess.org/.
[5] http://www.trossenrobotics.com/widowxrobotarm.
[6] http://www.ros.org/.
[7] Qt: Cross-platform application and UI framework. http://qt-project.org/, May 2014.
[8] C. Becker-Asano, P. Stahl, M. Ragni, J.-C. Martin, M. Courgeon, and B. Nebel. An affective virtual agent providing embodied feedback in the paired associate task: system design and evaluation. In *Proc. of the 13th. Intl. Conf. on Intelligent Virtual Agents (IVA 2013)*, pages 406–415, Edinburgh, UK, August 2013.
[9] C. Becker-Asano and I. Wachsmuth. Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems*, 20(1):32–49, January 2010.
[10] M. Campbell, A. H. Jr., and F. hsiung Hsu. Deep blue. *Artificial Intelligence*, 134(1 – 2):57 – 83, 2002.

[11] M. Courgeon, J.-C. Martin, and C. Jacquemin. MARC: a Multimodal Affective and Reactive Character. In *Proc. 1st Workshop on AFFective Interaction in Natural Environements*, 2008.

[12] A. E. Elo. *The rating of chessplayers, past and present*. Arco Pub., New York, 1978.

[13] I. Leite, C. Martinho, A. Pereira, and A. Paiva. icat: an affective game buddy based on anticipatory mechanisms. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 3*, AAMAS '08, pages 1229–1232, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems.

[14] I. Leite, C. Martinho, A. Pereira, and A. Paiva. As time goes by: Long-term evaluation of social presence in robotic companions. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 669–674, Sept 2009.

[15] C. Matuszek, B. Mayton, R. Aimi, M. Deisenroth, L. Bo, R. Chu, M. Kung, L. LeGrand, J. Smith, and D. Fox. Gambit: An autonomous chess-playing robotic system. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 4291–4297, May 2011.

[16] C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103, 2000.

[17] N. Riesterer, C. Becker Asano, J. Hué, C. Dornhege, and B. Nebel. The hybrid agent MARCO. In *Proc. of the 16th Intl. Conf. on Multimodal Interaction*, pages 80–81. ACM, 2014.

[18] L. Sajó, Z. Ruttkay, and A. Fazekas. Turk-2, a multi-modal chess player. *International Journal of Human-Computer Studies*, 69(7–8):483–495, 2011.

[19] M. Schröder. OpenMARY sources. https://github.com/marytts/marytts, April 2013.

[20] P. Sweetser and P. Wyeth. Gameflow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3):3–3, 2005.

[21] H. Vilhjálmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. Thóisson, H. Welbergen, and R. J. Werf. The behavior markup language: Recent developments and challenges. In *Intelligent Virtual Agents*, volume 4722 of *LNCS*, pages 99–111. Springer Berlin Heidelberg, 2007.

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Towards a Child-Robot Symbiotic Co-Development: a Theoretical Approach

Vicky Charisi[1], Daniel Davison[1], Frances Wijnen[2], Jan van der Meij[2], Dennis Reidsma[1], Tony Prescott[3], Wouter van Joolingen[4], Vanessa Evers[1]

Abstract. One of the main characteristics for an effective learning is the possibility for learners to choose their own ways and pace of learning, according to their personal previous experiences and needs. Social interaction during the learning process has a crucial role to the skills that learners may develop. In this paper, we present a theoretical approach, which considers relevant theories of child's development in order to proceed from a child-child collaborative learning approach to a child-robot symbiotic co-development. In this symbiotic interaction, the robot is able to interact with the learner and adapt its behaviours according to the child's behaviour and development. This sets some theoretical foundations for an on-going research project that develops technologies for a social robot that facilitates learning through symbiotic interaction.

1 INTRODUCTION

This paper discusses the conceptualization and some initial investigation of children's collaborative learning through symbiotic child-robot interaction in a specific educational setting. According to Douglas [1], biologist Heinrich Anton de Bary used the term "symbiosis" in 1879 to describe any association between *different* species. In this context, symbiotic learning describes the process, during which members of a team mutually influence each other resulting in an alteration of their behaviour. However, relationships among members may sustain imbalances. In order to support symbiotic interactions in learning, special considerations should be given to the orchestration of the relationships and the process between members of the team, from which they all benefit. The core motivating principle of symbiosis and the collaboration within it is reciprocity. Thus, learning emerges through a harmonized openness, responsiveness and adaptation. Elements of this kind of interaction may appear also in collaborative learning settings, which may not be especially designed for symbiotic interactions. Identifying elements of symbiotic interaction in children's collaborative learning may provide us with features for a more effective interaction design and for the design of robot behaviours as the child's co-learner.

In the following sections, we describe some constructivist aspects of child learning focusing on the need for learners to take responsibility for the regulation of the form and pace of learning. We then describe how symbiotic interaction can provide a theoretical and practical framework for understanding child-robot inter-dependence.

2 ASPECTS OF CHILDREN'S LEARNING PROCESSES

According to Foston and Perry [2], learning is a constructive activity that occurs through the interaction of individuals with their surroundings. Stages of development are understood as constructions of the active re-organization of learner's knowledge. This view builds on the constructivist framework of Piagetian developmental theory [3] according to which learning is a dynamic process comprising successive stages of adaption to reality, and during which learners actively construct knowledge by creating and testing their own theories and beliefs.

Two aspects of Piaget's theory underpin the pedagogical approach adopted here: First, an account of the four main stages of cognitive development through which children pass [4]. Since their birth, children go through (i) the sensori-motor stage (0-2 years), (ii) the pre-operational (2-7 years), (iii) the concrete operational (7-12 years) and (iv) the formal operational stage (12 years and onwards). For this project, we consider children in the age group between 7 and 12 years. During this stage, children are able to imagine "what if" scenarios, which involve the transformation of mental representation of things they have experienced in the world. These operations are "concrete" because they are based on situations that children have observed in the environment.

Second, an account of the mechanisms by which cognitive development takes place [5], which we consider in relation to environmental, social and emotional elements of child's development. These mechanisms describe how children actively construct knowledge by applying their current understanding.

2.1 Learning as a dynamic process

According to Piaget's classic constructivist view, learning occurs in a sequence of stages from one uniform way of thinking

[1] Dept. of Electrical Engineering, Mathematics and Computer Science, HMI, University of Twente, Enschede, the Netherlands. Email:{v.charisi, d.p.davison, d.reidsma, v.evers}@utwente.nl

[2] Faculty of Behavioural Management and Social Sciences, ELAN, University of Twente, Enschede, the Netherlands, Email: {f.m.wijnen, j.vandermeij}@utwente.nl

[3] Sheffield Robotics, University of Sheffield, Mappin Street, Sheffield, UK. Email: t.j.prescott@sheffield.ac.uk

[4] Freudenthal Institute for Science and Mathematics Education, Faculty of Science, Utrecht University, Utrecht, the Netherlands, Email: {w.r.vanjoolingen@uu.nl}

to another. Cognitive conflict, arising from discrepancies between internal representations and perceived events, functions as the motivating force for changing from concrete modes of thinking to more abstract forms. Although these stages relate to the child's genetic predispositions and biological development, environmental factors affect the transition from one stage to the next in complex ways. However, since Piaget first defined his framework it has been recognized that developmental transitions are not necessarily age specific events, but it occurs within an age range that can differ from child to child [6]. Additionally, the relationship between child development and the context in which this occurs, is bi-directional which results in a dynamical, iterative process; children affect and, simultaneously, they are affected by factors of their environment [7]. This can happen either in informal settings [8] which support tinkering and learning by doing or by following more formal and standardized processes, such as the inquiry cycle process [9], which will be described in 2.1.2 of this paper.

2.1.1 Child and the natural need for learning through exploration

In order for a child to be strongly engaged with a task it has to be meaningful for them. Since children have an inherent motivation to explore and understand their surroundings, the relevance of the task will stimulate their curiosity and willingness for exploration. Science education provides a formal learning setting that should share some of the characteristics of informal settings in order to help children acquire new concepts and develop transferable skills. Building on constructivist principles, children's natural enthusiasm for play can be a key factor in learning. During play, children can explore the real world, logically organize their thoughts, and perform logical operations [10]. However, this occurs mainly in relation to concrete objects rather than abstract ideas [8]. Children are also able to reflect on their intentional actions which may result in a self-regulated process of change [11].

2.1.2 Inquiry cycle: a systematic process of learning

'Inquiry is an approach to learning that involves a process of exploring the natural or material world, and leads to asking questions, making discoveries, and rigorously testing those discoveries in the search for new understanding. Inquiry should mirror as closely as possible the enterprise of doing real science' [12] (p.2). The main claim of inquiry learning, in relation to science learning, is that it should engage learners in scientific processes to help them build a personal scientific knowledge base. They can then use this knowledge to predict and explain what they observe in the world around them [13]. Thus, having as a starting point child's tendency for informal exploration, with developmental appropriate scaffolding, children develop their scientific thinking. This transferable skill can then facilitate child learning in different contexts.

There are many models that represent the processes of inquiry, but all include the processes of (1) *hypothesis generation* in which learners formulate their ideas about the phenomena they are investigating, (2) *experimentation* in which children perform experiments to find evidence for rejection or confirmation of their hypotheses and (3) *evidence evaluation* in which learners try to find logical patterns in their collected data and to interpret this data to form a conclusion [14, 15].

Banchi and Bell [9] describe a four-level continuum to classify the levels of inquiry in an activity, focusing on the amount of information and guidance that is presented to the learner [9, 16]:

Confirmation inquiry: In this form of inquiry learners are provided with the research question, method of experimentation and the results that they should find. This is useful if, for example, the goal is to introduce learners to the experience of conducting investigations or to have learners practice a specific inquiry skill such as collecting data.

Structured inquiry: Here, the question and procedure are still provided but the results are not. Learners have to generate an explanation supported by the evidence they have collected. In this case learners do know which relationship they are investigating.

Guided inquiry: In this form learners are provided only with the research question. Learners need to design the procedure to test their question and to find resulting explanations.

Open inquiry: This is the highest level of inquiry. Here, learners have the opportunities to act like scientists, deriving questions, designing and performing experiments, and communicating their results. This level requires the most scientific reasoning and is the most cognitive demanding. This low- to higher-level continuum of inquiry is important to help learners gradually develop their inquiry abilities [9]. The obtained inquiry skills are transferable to other contexts.

2.2 The zone of proximal development (ZPD)

The level of potential development is the level at which learning takes place. It comprises cognitive structures that are still in the process of maturing, but which can only mature under the guidance of or in collaboration with others. Vygotsky [17] distinguished between two developmental levels: the level of actual development and that of potential development. The actual development is the level, which the learner has already reached and she can solve problems independently. The level of potential development, which is also known as the *Zone of Proximal Development (ZPD),* describes the place where child's spontaneous concepts meet the systematic reasoning under the guidance or in collaboration with others [18]. In that way, Vygotsky argues that the interpersonal comes before the intrapersonal. This is considered to be as one of the fundamental differences between Vygotsky's conceptualization of child development and that of Piaget.

Learning takes place within the ZPD and here a transition occurs in cognitive structures that are still in the process of maturing towards the understanding of scientific concepts. The level of potential development varies from child to child and is considered a fragile period for child's social and environmental support through the educational praxis. In this context, Vygotsky introduced the notion of 'scaffolding', to describe the expansion of the child's zone of proximal development that leads to the construction of higher mental processes [19]. However, only if we define what causes the expansion of ZPD, we will be able to provide appropriate scaffolding for learners. Siegler [20], for example, has highlighted the question of what

causes change in learning mechanism and he concluded that seemingly unrelated acquisition are products of the same mechanisms or mental process. Scaffolding is considered a core element for the support of child's mental changes in the context of collaborative learning.

2.3 Collaborative learning

Rogoff's [21] definition of collaboration includes mutual involvements and engagement and participation in shared endeavours, which may or may not serve to promote cognitive development. This broad definition allows for flexibility regarding its interpretation and it is adjustable into different contexts. For the present research, we use this definition as a basis for our theoretical approach for collaboration in the context of learning.

Vygotsky [17] emphasized the importance of social interaction with more knowledgeable others in the zone of proximal development and the role of culturally developed sign systems that shape the psychological tools for thinking.

In addition to the development of their cognitive skills, children's social interactions with others during the learning process may trigger their meta-cognitive skills, as well. Providing explanations during collaboration in which children reflect on the process of their learning (meta-cognitive skills) leads to deeper understanding when learning new things [22, 23]. There are two forms of explanation: (1) self- explanation, which refers to explanation of the subject of interest to oneself, and (2) interactive explanation, which refers to explanation to another person [24]. In both cases, the presence of a social partner facilitates children's verbalization of their thinking. However depending on the type of the social partner, children may exhibit different behaviours, which relate to different kind and quality of learning.

The following sections describe two different types of social partners as mediators for children's learning to occur.

2.3.1 Child – tutor

With regard to adult-child interactions, Wood et *al.* [25] defined tutoring as 'the means whereby an adult or 'expert' helps somebody who is less adult or less expert' (p.89). Receiving instructions from a tutor is a key experience in childhood learning (ibid.). This definition of tutoring implies a certain mismatch in the knowledge level between the parties involved, in such a way that the tutor has superior knowledge or skill about a subject which is then passed on to a child via tutoring mechanisms.

2.3.2 Child – child

In combination with tutoring, peer learning has been defined by Topping [26] as 'the acquisition of knowledge and skill through active helping and supporting among status equals or matched companions' (p.1). Topping continues to describe that peer learning 'involves people from similar social groupings who are not professional teachers helping each other to learn and learning themselves by so doing' [26]. This learning method has proven to be very effective amongst children and adults and has been widely researched over the past decades. Peer learning assumes a matched level of initial knowledge of both parties. In ideal peer learning situations, both parties will increase their knowledge levels at a similar pace through collaborative learning mechanisms.

2.4 Emotional engagement and social interaction (in learning)

The importance of positive feelings during the learning process has been reported as crucial [27]. They promote the individual's openness to new experiences and resilience against possible negative situations [28]. It has been reported that dynamic behaviours involve reciprocal influences between emotion and cognition [29]. For instance, emotions affect the ways in which individuals perceive the reality, pay attention and remember previous experiences as well as the skills that are required for an individual to make decisions.

3 SYMBIOTIC INTERACTION

The educational and developmental theories outlined in the previous sections describe various forms of collaborative learning. Social interaction between learners is emphasised as an important factor in successful collaborative learning, where both students co-develop at a complementary pace through shared experiences.

Within the context of this co-development we define *symbiotic interaction* as the dynamic process of working towards a common goal by responding and adapting to a partner's actions, while affording your partner to do the same.

The fundamental requirements for team collaboration have been discussed in detail by Klein and Feltovich [30]. They argue that in order to perform well on *joint activities*, or collaborative tasks, there must be some level of *common ground* between teammates. These concepts have been introduced by Clark [31] to describe the intricate coordination and synchronization processes involved in everyday conversations between humans.

Common ground between team participants is the shared mutual knowledge, beliefs and assumptions, which are established during the first meeting and continuously evolve during subsequent interactions. A strong common ground can result in more efficient communication and collaboration during joint activity, since a participant can assume with relative safety that other participants understand what she is talking about without much additional explanation [30].

Klein and Feltovich [30] argue that in order for a task to qualify for effective joint activity, there must firstly be an *intention* to cooperate towards a common goal and secondly the work must be interdependent on multiple participants. As long as these preconditions are satisfied, a joint activity requires *observable, interpretable* and *predictable* actions by all participants. Finally, participants must be open to *adapt* their behavior and actions to one another. The different processes of the joint activity are choreographed and guided by clear *signaling of intentions* between participants and by using several

coordination devices such as agreement, convention, precedent and salience.

3.1 Intention to act towards a common goal

An important precondition for symbiotic interaction is the *awareness* of a certain common goal, and a clear *intention* to work towards this goal. During the process of establishing and maintaining common ground, both parties will (implicitly or explicitly) become aware of the goals of the other. Maintaining common ground relies on being able to effectively signal your intent to a partner, while at the same time interpreting and reacting to the intent of his or her actions [30].

3.2 Observability of actions and intentions

Equally important to being able to effectively *signal* intent is the ability of the partner to *observe* and *interpret* this intent. A sense of *interpredictability* can be achieved when such signals can be naturally and reliably generated, observed and interpreted by both partners. A healthy level of interpredictability between partners can contribute to an increased common ground and mutual trust between partners [30].

3.3 Interpredictability, adaptability and trust

Within the context of an interaction, predictability means that one's actions should be predictable enough for others to reasonably rely on them when considering their own actions. Over the course of an interaction, certain situations arise which allow a person to estimate the predictability of a partner's actions, or in other words, the amount of *trust* you place in the predictability of your partner. Simpson [32] argues that in human-human interaction, trust levels are often established and calibrated during trust-diagnostic situations "in which partners make decisions that go against their own personal self-interest and support the best interests of the individual or the relationship" [32]. This *willingness* to act predictably and *adapt* one's behavior to support a partner's best interests is a key component of building mutual trust and supporting a symbiotic relationship [33].

In summary, an effective joint activity relies on signaling, observing and interpreting the intent of actions towards a common goal. By establishing a strong common ground, both partners achieve a level of interpredictability. An important factor in building trust is to expose a willingness to act predictively and adapt one's behavior to match the common goals shared with a partner.

4 CHILD-ROBOT INTERACTION

The work reported in this paper is part of a project on social robots in learning scenarios. Social interaction with a robot affects the child's independence during the learning process. Robots can take either end of the spectrum depending on its role, in other words, it can be either tutor-like or peer-like for child learning [34]. Depending on the amount of support needed for the child's learning, the robot might adapt its role to fit this need, shifting either more towards the tutor or the peer role. This adaptive behavior fits the theories on symbiotic interactions outlined above. Together with clear signaling of intents, which contribute to an increased level of predictability, it is this adaptability that proves to be an important factor in building a long-term symbiotic relationship.

Belpaeme et al. [35], for example, have reported the importance of adaptive behavior of the robot when it interacts with children with diabetes. In this study, researchers adapted robot behaviour according to children personality (extroverted / introverted) and to the difficulty level of the task. They concluded that adaptation to user characteristics is an effective aid to engagement.

In the context of the learning process, a robot may adapt its behavior to the child's cognitive, social and emotional characteristics with a purpose to facilitate the expansion of children's zone of proximal development. Thus, the robot can scaffold the process of change by adapting its behaviour according to the user. It shows its awareness and willingness to be influenced by others. The robot then will adapt to the child's next level in order to contribute to the iterative process of development. In this way, we create a learning context based on symbiosis of the child and the robot.

5 FUTURE AGENDA

Inspired by the insights derived from the previously introduced theoretical framework for co-development in learning, we outline our future goals, which focus on the elaboration of aspects of this framework and explore its utility for designing robot-child interactions for inquiry learning. To conclude this paper we briefly describe a contextual analysis we are performing to validate the framework in the specific pedagogic setting of inquiry learning. Thereafter we briefly present some of our ideas for future experiments.

5.1 Some first insights from a contextual analysis

An initial contextual analysis is being performed based on observations of twenty-four children who are working in pairs on a balance beam task. The balance beam task is a specific implementation of a type of structured inquiry learning. Using the balance beam children investigate the weight of several provided objects, exploring both the influence of weight rations and the distance of the object to the pivot.

The setting for this contextual analysis was as follows: a total of 11 pairs of two children (aged 6-9 years) received a structured assignment, which they could complete by using the balance beam that was presented. This assignment was designed according to the processes of structured inquiry (e.g. hypothesis generation, experimentation, evidence evaluation). The children could place pots that differed in weight on different places on the balance, make predictions about what would happen to the balance (tip left, tip right, or stay in equilibrium), perform experiments by removing wooden blocks that held the balance in equilibrium, observe what happened with the balance and

draw conclusions about the variables that influence the balance (weight, distance). These procedures were videotaped and than annotated. These annotations are not yet fully analysed, but a few first indications will be described here.

1. It appeared that children who followed the steps of the assignment correctly were engaging in the different processes that are typical for inquiry learning, and were interacting with each other about the process and the outcome of the task.
2. Most children were able to identify the influence of the two variables (weight and distance) on the balance eventually.
3. Several children asked for additional guidance from the experimenter during the task.

These first insights from the contextual analysis have been taken into account for our next steps for the design of child-robot interaction in the same context. We observed that children in this age may follow the inquiry process during the activity. However, in order for them to reflect on this process, verbalize their thoughts and explain the scientific phenomenon under investigation, they needed the support from a social partner. The teacher facilitated child's process by different types of interaction, such as supporting children's inquiry process by probing questions or asking for explanations and summarizations. In addition to the verbal interaction, we considered non-verbal cues of social interactions that appeared during this contextual analysis. The emerging types of social interactions have informed our design for future experiment on child-robot interaction.

5.2 Planned experiments

Our next steps include two experiments on child-robot interaction. In the first experiment we will focus on the influence of a social robot on explanatory behavior. Explanatory behavior includes the verbalization of scientific reasoning of the child.

The experiment is comprised of two conditions. In the experimental condition the child will be working on an inquiry assignment with the robot. The background story of the robot is that he comes from another planet. He has an assignment from his teacher to study the effects of balance on earth. The robot wants to explore this phenomenon with like-minded people: children. The robot is presented as a peer learner but he does have well-developed inquiry skills. Therefore, the robot will provide instructions and ask questions to help learners explore the phenomenon of balance with the balance beam. The children will provide their answers by talking to the robot. The input of the state of the learning material for the robot will be controlled by a 'Wizard of Oz' technique.

In the control condition learners will be working on the same assignment but without the robot. In this case the tablet provides instruction and will pose exactly the same questions to help learners explore the phenomenon of balance. In the control condition there is no background story, but children are asked to do the assignment as part of their educational program. The children will provide their answers verbally, and it will seem as if the tablet records the answers. In both conditions video recordings will be made of the children working on the task. It is hypothesized that when working on the task in an appropriate social context, in this case being accompanied by the robot, giving answers to the questions will result in more verbal explanatory behavior. Verbally explaining to another person can facilitate greater understanding of one's own ideas and knowledge [23] and might therefore lead to better learning and transfer [36].

The second experiment will focus on the expected cognitive competence children believe the robot has. There will be three conditions. In all conditions the robot will make some incorrect suggestions. The difference between the conditions is that the children are primed to believe that the robot is (1) an expert, (2) a novice or (3) no priming. The goal is to find out how competent and trustworthy children believe the robot is before and after the experiment.

In this paper, we have described some aspects of an initial theoretical framework that we use to design our experiments and user studies to investigate child-robot symbiotic interaction. We are going to give an emphasis to the process of learning in different contexts, focusing on collaborative learning and exploiting the robot as an adaptive co-learner. Thus the robot can scaffold the child to go through an effective learning process. For the future work we aim to investigate how a social robot can scaffold child's inquiry process by facilitating the expansion of ZPD in an effective and enjoyable way focusing on the development of children's meta-cognitive skills.

ACKNOWLEDGMENT

This project has received funding from the European Union Seventh Framework Programme (FP7-ICT-2013-10) as part of EASEL under grant agreement n° 611971.

REFERENCES

[1] Douglas, A.E., *The Symbiotic Habit*. Princeton: Princeton University Press. (2010).

[2] Foston, C.T. and R.S. Perry, *Constructivism: A psychological Theory of Learning*, in *Constructivism*, C.T. Foston, Editor., Teachers College, Columbia University: New York. (2005).

[3] Piaget, J. and B. Inhelder, *The psychology of the child*. New York: Basic Books. (1969).

[4] Piaget, J., *The theory of stages in cognitive development*, in *Measurement and Piaget*, D.R. Green, Editor., McGraw-Hill: New York. p. 1-11. (1971).

[5] Papert, S., *Papert on Piaget*, in *Time: The Century's Greatest Minds*. p. 105. (1999).

[6] Donaldson, M., *Children's minds*. Glasgow: Fontana. (1978).

[7] Smith, L.B. and E. Thelen, *Development as a dynamic system.* Trends in Cognitive Sciences, **7**(8): p. 343-348, (2003).

[8] Resnick, M. and E. Rosenbaum, *Designing for tinkerability*, in *Design, make, play: growing the next generation of STEM Innovators*, M. Honey and D. Kanter, Editors., Routledge: New York. p. 163-181. (2013).

[9] Banchi, H. and R.L. Bell, *The many levels of inquiry.* Science and Children, **46**: p. 26-29, (2008).
[10] Pellegrini, A.D., *The Role of Play in Human Development*. New York: Oxford University Press. (2009).
[11] Whitebread, D. and D. Pino Paternak, *Metacognition, self-regulation and meta-knowing*, in *International Handbook of Psychology in Education*, K. Littleton, C. Wood, and J. Kleine Staarman, Editors., Emerald: Bingley, UK. (2010).
[12] NSF, *An introduction to inquiry*, in *Inquiry: Thoughts, Views and Strategies for K-5 Classroom*. National Science Foundation: Washington. p. 1-5. (2000).
[13] van Joolingen, W.R., T. de Jong, and A. Dimitrakopoulou, *Issues in computer supported inquiry learning in science.* Journal of Computer Assisted Learning, **23**: p. 111-119, (2007).
[14] Klahr, D., *Exploring science: the cognition and development of discovery processes*. Cambridge: The MIT Press. (2000).
[15] Klahr, D. and K. Dunbar, *Dual search space during scientific reasoning.* Cognitive Science, **12**: p. 1-48, (1988).
[16] Bell, R.L., L. Smetana, and I. Binns, *Simplifying inquiry instruction.* The Science Teacher, **72**: p. 30-33, (2005).
[17] Vygotsky, L.S., *Mind in Society: The Development of Higher Psychological Process*. Cambridge, MA: Harvard University Press. (1930-34/1978).
[18] Corsaro, W., *The sociology of childhood*. Thousand Oaks, California: Pine Forge Press. (1997).
[19] Vygotsky, L.S., *Play and its role in the development of the child.* Soviet Psychology, **12**(6): p. 62-76, (1933/1966).
[20] Siegler, R.S., *Microgenetic Analyses of Learning*, in *Handbook of Child Psychology*, D. Kuhn and R.S. Siegler, Editors., Wiley: Hoboken, NJ. p. 464-510. (2006).
[21] Rogoff, B., *Cognition as a Collaborative Process*, in *Handbook of Child Psychology*, D. Kuhn and R.S. Siegler, Editors., Wiley: New York. p. 679-744. (1998).
[22] Chi, M.T.H., et al., *Self-explanations: how students study and use examples in learning to solve problems.* Cognitive Science, **13**: p. 145-182, (1989).
[23] Holmes, J., *Designing agents to support learning by explaining.* Computers and Education, **48**: p. 523-547, (2007).
[24] Ploetzner, R., et al., *Learning by explaining to oneself and to others*, in *Collaborative learning: Cognitive and Computational Approaches*, P. Dillenbourg, Editor., Elsevier: Oxford. p. 103-121. (1999).
[25] Wood, D., J.S. Bruner, and G. Ross, *The role of tutoring in problem solving.* Journal of Child Psychology and Psychiatry, **17**(2): p. 89-100, (1976).
[26] Topping, K.J., *Trends in Peer Learning.* Educational Psychology, **25**(6): p. 631-645, (2005).
[27] Kort, B., R. Reilly, and R.W. Picard, *An affective model of interplay between emotions and learning: reenfineering educational pedagogy - Building a learning companion*, in *International Conference of Advanced Learning Technologies*. IEEE Computer Society: Washington, DC. (2001).
[28] Fredrickson, B.L., *The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions.* American Psychologist, **56**(3): p. 218-226, (2001).
[29] Dolcos, F., A.D. Iordan, and S. Dolcos, *Neural correlates of emotion-cognition interactions: Areview of evidence from brain imaging investigations.* Journal of Cognitive Psychology, **23**(6): p. 669-694, (2011).
[30] Klein, G. and P. Feltovich, *Common Ground and Coordination in Joint Activity.* Organizational simulation: p. 1-42, (2005).
[31] Clark, H., *Using Language.* (1996).
[32] Simpson, J.A., *Psychological Foundations of Trust.* Current Directions in Psychological Science, **16**: p. 264-268, (2007).
[33] Lee, J.D. and K.A. See, *Trust in Automation: Designing for Appropriate Reliance.* Human Factors: The Journal of the Human Factors and Ergonomics Society, **46**: p. 50-80, (2004).
[34] Kanda, T., et al., *Interactive robots as social partners and peer tutors for children: a field trial.* Human-Computer Interaction, **19**(1): p. 61-84, (2004).
[35] Belpaeme, T., et al., *Multimodal Child-Robot Interaction: Buiding Social Bonds.* Journal of Human-Robot Interaction, **1**(2): p. 33-53, (2012).
[36] Rittle-Johnson, B., M. Saylor, and K. Swygert, *Learning from explaining: does it matter if mom is listening?* Journal of Child Psychology, **100**(3): p. 215-224, (2008).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Does anyone want to talk to me? – Reflections on the use of assistance and companion robots in care homes

Kerstin Dautenhahn[1] and **Anne Campbell**[2] and **Dag Sverre Syrdal**[1]

1 Introduction: Robotic Companions for Elderly People

A growing number of research efforts worldwide aim at developing assistive robots to help elderly people in their own homes or in care homes. The rationale for home assistance robotic technology is based on demographic changes in many countries worldwide, with an ageing population. For example, it is predicted that in the European Union the number of people over 65 years will almost double (by 2060) and the number of people between 15-64 years will decrease by over 10%. Health care costs are also rising [33]. Developments into home companions and solutions for Ambient Assisted Living (AAL) in elderly peoples homes or care homes have grown significantly in the EU, see projects such as SRS[12] , Hermes[5] , Florence [4] , KSERA[7] , MOBISERV [9] , Rubicon [11], ACCOMPANY [1] or ROBOT-ERA[10] , to name a few. Recent videos of results on smart home companion robots and the type of assistance they can provide have been illustrated for MOBISERV[29] and ACCOMPANY[15] . Products for robots used in peoples homes are beginning to be marketed, cf. Toyota's Human Support Robot (HSR)[13], Mitsubishi's communication robot Wakamaru[14] , Aldebaran's Pepper robot [2], or Cynthia Breazeal's Jibo robot[6]. These robots come in different shapes and sizes, and appearance and behaviour will influence which roles these robots are being assigned to by their users and the human-robot relationships that may emerge.

One of the authors has been involved in European projects on home assistance robots since 2004, as part of the COGNIRON [3] , LIREC [8] and ACCOMPANY [1] projects. COGNIRON was one of the first projects in Europe on home companion robots. One lesson learnt during the project was the need to move out of the laboratory and into a realistic home setting, which led to the acquisition and development of the University of Hertfordshire Robot House, a smart home equipped with a sensor network and robots being able to detect daily living activities and provide physical, social and cognitive assistance. A second lesson was the need to move away from Wizard-of-Oz (remote controlled) studies. In LIREC the emphasis was on developing fully autonomous home assistant robots, with an emphasis on social assistance. During ACCOMPANY, this direction has been elaborated and extended through allowing the robot to be taught and shown new behaviours and routines by the user, including evaluations with elderly users and their formal and informal carers in long-term studies in three European countries. The ACCOMPANY project has particularly advanced a direction where such autonomously operating companion robots, as part of a smart home infrastructure, socially engage and assist the user, using personalization and human-robot teaching and co-learning for reablement of the user[36] . While these projects have focused primarily on the use of robot within home settings, a separate strand of research within the University of Hertfordshire's work in ACCOMPANY actively elicited the views of residents and staff at a local care home, through the use of theatre prototyping[34] followed by interviews as previously reported in Walters et al.[45]. The current position paper draws on these experiences and findings, as well as those from the other projects, to consider the role that social robots may play in a care home environment.

[1] School of Computer Science, University of Hertfordshire, email: {k.dautenhahn; d.s.syrdal} @herts.ac.uk
[2] School of Health and Social Work, University of Hertfordshire, email a.2.campbell@herts.ac.uk

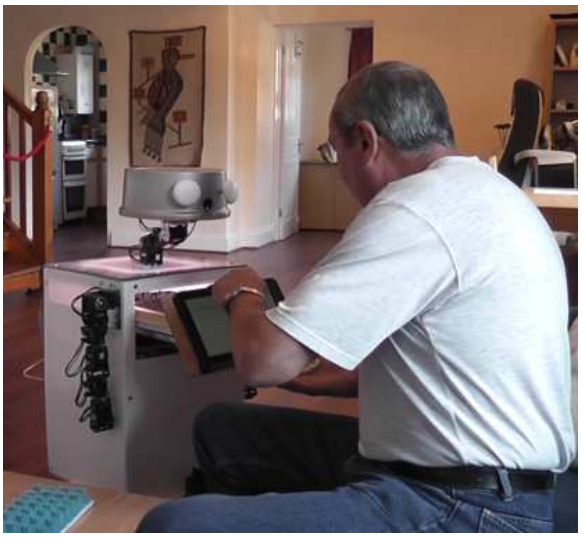
2 Roles of Robots

Different roles of robots in human society have been proposed[21] , including a machine operating without human contact; a tool in the hands of a human operator; a peer as a member of a humaninhabited environment; a robot as a persuasive machine influencing people's views and/or behaviour (e.g. in a therapeutic context); a robot as a social mediator mediating interactions between people; a robot as a model social actor. Opinions on viewing robots either as friends, assistants or butlers have been investigated [23] . It has been suggested the robot can act as a mentor for humans, or information consumer whereby a human uses information provided by a robot[25] . Further roles that have been introduced view robots as a team member in collaborative tasks [19] or roles for robots as learners [39, 28]. Companion robots have been defined as robots that not only can carry out a range of useful tasks, but do so in a socially acceptable manner [22] . This role typically involves both long-term and repeated interaction, as is the case for robots used in an elderly person's home or in a care home. Will people develop human-like relationships with such companion robots? Some studies have tried to address these from a user-centric point of view. Beer et al.[18] found that participants primarily focused on the ability of the robot to streamline and reduce the amount of effort required to maintain their household. However, a recent study based on both recent literature research and focus groups with 41 elderly people, 40 formal caregivers and 32 informal caregivers in the Netherlands, UK and France, the most problematic challenges to independent living were identified mobility, self-care, and interpersonal interaction and relationships [17].

Thus, there seem to be two domains where robots are envisaged to assist in: the physical and/or cognitive domain, providing e.g specific assistance in remembering events and appointments, or to move around, and the domain of social relationships.

This duality of roles do exist in how robots are being proposed to be used in such settings, while surveys of envisaged use scenarios

Figure 1. A companion robot at the University of Hertfordshire



indicate that medical and healthcare personnel see robots as tools that can provide physical assistance with their tasks [40] , however, there are also studies investigating the value of robots as companions in these settings[38].

This approach is grounded in that, apart from physical needs, a key problem in care homes is the resident's loneliness. It impacts upon ' quality of life and wellbeing, adversely affects health and increases the use of health and social care services . A number of interventions have been used, e.g. one-to-one approaches such as Befriending, Mentoring, group services such as lunch clubs, or community engagement through public facilities (sports etc) [46]. Interestingly, in a recent approach chickens have been introduced to a care home, and proved popular with both staff and residents[35]. The impact of robots and animals can be directly compared[16]. Could robots become part of such services?

3 Ethical Issues

While this short position paper cannot comprehensively address the ethical issues involved in the adoption of robots in elder care and the associated literature, we note that elsewhere the danger to anthropomorphise and romanticise robots has been highlighted[20]. The roles that are ascribed to robots and the human—robot relationships discussed in the research community are predominantly based on terms that originally describe human-human interactions. So there is a tendency to use terms robotic 'assistant' or robotic 'carer' and apply the human equivalent literally which automatically implies a whole range of different human-like qualities and abilities, that robots at present cannot address, in terms of their physical and cognitive abilities, as well as in terms of their emotional intelligence, as well as ethical and moral judgements. A number of ethical considerations need to be considered when fostering social relationships between robots and elderly people. Sherry Turkle[41] has previously discussed the danger of 'relational artifacts', i.e. robot designed specifically to encourage people to form a relationship with them. She argued that such 'non-authentic' interaction may lead to people preferring the (relatively easy and predictable and non-judgemental) interaction with a robot compared to interactions with real people. Specifically with regard to eldercare, Amanda and Noel Sharkey[37] pointed out risks involved in using robots in elder care, including the potential for the reduction in the amount of human contact as well as concerns about deception and infantilisation. The theme of deception, infantilisation and the possible reduction in human contact is also emphasized in other reflections on ethical norms of using robots in caring role for elderly people[42, 24].

Interestingly, designing robots as interactive systems that people can engage with, e.g. play games with, is technically feasible. Even pet-like, non-humanoid robots such as Paro have been shown to be successful companions[30]. On the other hand, providing physical assistance involves many technical challenges e.g. in terms of object manipulation, navigation, safety, etc. Thus, if it is 'easier' to build robots as socially interactive companions, and to focus on its role to engage people, shall one concentrate research efforts on this aspect? Is it ethically justifiable, desirable and acceptable by elderly people and their carers, given the above mentioned concerns of deception, infantilisation, and providing non-authentic experiences? In order to shed some initial light on these issues, one of the authors conducted interviews in a care home for elderly people.

4 INTERVIEWS STUDY WITH RESIDENTS AND CARER IN A CARE HOME

An interview study was conducted with carers and residents of a care home in UK. In this study, residents and staff at the residential care home were shown a play which focused on how the adoption of personal home companion impacted the relationships in a domestic household. The play and other aspects of the study is briefly summarised here, details are provided elsewhere[45] . While the play focused on the use of a robot in a different environment, it served to raise awareness of how robots may assist in, and influence the daily life of their users. We would also note that there was no verbal interaction from the robot in the play. Three months after the play, a follow-up study was conducted in which three residents, all with learning disabilities and/or physical disabilities were interviewed, followed by interviews of three experienced registered nurses. The 15-20 min interviews took place in the communal dining room of the home that is familiar and comfortable to both residents and carers. A semi-structured interview technique was used since it is considered a reliable and flexible method and can cater for some of the residents' disabilities[32]. The interviewer wrote down the interview data during the interview, an approach considered less intrusive than audiotaping the interviews. Based on these notes, the interviewer conducted a content analysis of the interview data a number of themes emerged that are described in detail in Walters et al. [45]. Relevant for the present article are the following themes and comments from residents and carers: Concerning acceptable boundaries for care by humans and robots, one resident said that the most important care for her from the robot was psychological care:

> *'Make me feel lovely in myself and give me a boost...make things different...I want to dance with it'.*
> *'I would like the robot to be chatty and to nod his head to show he has heard me'.*

Two other residents wanted the robot to *'Tidy my room and maybe feed me in the future'* and *'comb my hair'*. Regarding conversation and companionship, one of the interviewed residents wanted the robot to be able to start a conversation and then acknowledge that he had heard about her sore knee. Another wanted the robot to dance with her. One theme arising from the interviews of the registered nurses concerned how the robot could provide assistance to staff and

residents, while they still preferred a human to a robot colleague. All 3 nurses thought the robots would help with both physical and psychological care:

'They could provide company, socialise and boost morale'.
'They could be friendly, shake hands and make friendly sounds; talk to them and reduce loneliness'.
'Help with feeding and walking beside them would be helpful'.

Concerning conversation and companionship all three nurses would really value robot that can engage in conversations with residents and provide stimulation:

'Stimulation helps residents feel important'.
'Helpful when staff are busy'.

5 Reflections

The interview study above highlighted a number of issues in favour of robot providing social interaction and communication with residents in a care home in order to help with their loneliness. There are also a number of practical issues, based on experience gained by the second author in care homes, that would support robots in that role:

- The group of residents in care homes is often diverse, ranging from people with dementia, people with learning disabilities, people terminally ill e.g. with cancer, and others. This diversity can impact on the willingness and enjoyment of residents to talk to each other
- Residents in a care home do not know each other prior to joining the care home, they are not a naturally formed unit of friends or family. We cannot expect randomly created groups of people to make friends easily, or even to be interested in talking to each other, while having to live under the same roof under a daily basis.
- Care staff is often very focused on task and efficiency, often under a lot of time-pressure to 'get things done'. There is a large spectrum in the quality of care, but in some care homes social interaction with residents might not be high on the priority list of care staff and their managers.
- From the point of view of care staff, interaction with residents may not always be as enjoyable as one might envisage, e.g. due to memory problems people with dementia may engage in very repetitive conversations.
- In a social environment such as a care home, residents might feel not 'getting along with the others', due to real or perceived conflicts with other residents.
- Some residents may have psychiatric conditions which make them feel paranoid and sometimes aggressive.
- Care home staff and/or residents may not all have English as their first language which affects their ability to communicate with each other smoothly. There may also be differences in intercultural understanding of what is socially acceptable conversation.

Thus, while in an ideal world, care homes should be places where carers and residents live together as 'one happy family', the reality often differs. And it may be useful for robots to provide opportunities for communication and interaction, even if interaction with robots is mechanical, and lacks authenticity and depths of human contact as we have argued elsewhere[41, 22] . For example, present robots cannot replace the gentleness and meaningfulness of a person stroking someone's hair, or touching someone's hands, or a comforting word. This does not always mean that the robot will have to replace carer-resident or resident-resident interactions. Rather, it may function as a social facilitator, or mediator, and may be able to assist residents and carers in overcoming some of the practical issues that often restrict human-human interactions in care homes. Previous research has suggested that the presence of a robot in a care may work to facilitate a greater degree of interaction between the residents of the care home [27, 43] , and this effect may be leveraged further by using features like a memory visualisation system (which uses photos and text to create narratives of previous interaction)[26] to aid further when trying creating common ground between human interactants . In addition, there is also the possibility to adapt and apply research in using robots to increase dyadic interactions in other user-groups [44, 31] in order to further the ability of a robot companion as a social facilitator or mediator. While it can be argued that some of the issues, in particular the staff's focus on task and efficiency can be mediated by the adoption of robots to provide physical support with some of the tasks, this does not necessarily address the other points raised here. We do not argue for robots to replace carers or human contact in general, however, we argue that in situations where residents can expect, and may suffer from, only very little human contact that in such circumstances robots could be beneficial to them and their carers, by helping them to feel less lonely, not only through the direct interaction between the resident and the robot, but also through the robot's ability to mediate interactions between residents and residents and carers — and thus improving the health and well-being of the residents as well as the working conditions and atmosphere at work as experienced by the staff.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper.

REFERENCES

[1] ACCOMPANY. http://rehabilitationrobotics.net/cms2/.
[2] Aldebaran Pepper. https://www.aldebaran.com/en/a-robots/who-is-pepper.
[3] Cogniron. http://www.cogniron.org/final/Home.php.
[4] Florence. http://florence-project.eu.
[5] Hermes. http://fp7-hermes.eu.
[6] Jibo. http://www.myjibo.com/.
[7] KSERA. http://ksera.ieis.tue.nl.
[8] LIREC. http://lirec.eu/project.
[9] MOBISERV. http://www.mobiserv.info/.
[10] Robot-ERA. http://robot-era.eu.
[11] Rubicon. http://fp7rubicon.eu/.
[12] SRS. http://srs-project.eu.
[13] Toyota Human-Support Robot. http://www.toyota-global.com/innovation/partner_robot/.
[14] Wakamaru. https://www.mhi-global.com/products/detail/wakamaru_about.html.
[15] ACCOMPANY. ACCOMPANY Project Video. http://www.youtube.com/watch?v=Z1MJPdhniXc.
[16] Marian R Banks, Lisa M Willoughby, and William A Banks, 'Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs', *Journal of the American Medical Directors Association*, **9**(3), 173–177, (2008).
[17] Sandra Bedaf, Gert Jan Gelderblom, Dag Sverre Syrdal, Hagen Lehmann, Hervé Michel, David Hewson, Farshid Amirabdollahian, Kerstin Dautenhahn, and Luc de Witte, 'Which activities threaten independent living of elderly when becoming problematic: inspiration for meaningful service robot functionality', *Disability and Rehabilitation: Assistive Technology*, **9**(6), 445–452, (2013).
[18] Jenay M Beer, Cory-Ann Smarr, Tiffany L Chen, Akanksha Prakash, Tracy L Mitzner, Charles C Kemp, and Wendy A Rogers, 'The domesticated robot: design guidelines for assisting older adults to age in

place', in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pp. 335–342. ACM, (2012).
[19] Cynthia Breazeal, Andrew Brooks, Jesse Gray, Guy Hoffman, Cory Kidd, Hans Lee, Jeff Lieberman, Andrea Lockerd, and David Chilongo, 'Tutelage and collaboration for humanoid robots', *International Journal of Humanoid Robotics*, **1**(02), 315–348, (2004).
[20] Kerstin Dautenhahn, 'Human-robot interaction', in *The Encyclopedia of Human-Computer Interaction, 2nd Ed*, eds., Mads Soegaard and Rikke Friis Dam.
[21] Kerstin Dautenhahn, 'Roles and functions of robots in human society: implications from research in autism therapy', *Robotica*, **21**(04), 443–452, (2003).
[22] Kerstin Dautenhahn, 'Socially intelligent robots: dimensions of human–robot interaction', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **362**(1480), 679–704, (2007).
[23] Kerstin Dautenhahn, Sarah Woods, Christina Kaouri, Michael L Walters, Kheng Lee Koay, and Iain Werry, 'What is a robot companion-friend, assistant or butler?', in *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pp. 1192–1197. IEEE, (2005).
[24] Heather Draper, Tom Sorell, Sandra Bedaf3 Christina Gutierrez Ruiz, Hagen Lehmann, Michael Hervé, Gert Jan Gelderblom, Kerstin Dautenhahn, and Farshid Amirabdollahian, 'What asking potential users about ethical values adds to our understanding of an ethical framework for social robots for older people.', *MEMCA-14. This Proceedings*, (2014).
[25] Michael A Goodrich and Alan C Schultz, 'Human-robot interaction: a survey', *Foundations and trends in human-computer interaction*, **1**(3), 203–275, (2007).
[26] Wan Ching Ho, Kerstin Dautenhahn, Nathan Burke, Joe Saunders, and Joan Saez-Pons, 'Episodic memory visualization in robot companions providing a memory prosthesis for elderly users', *Assistive Technology*, (2013).
[27] Cory D Kidd, Will Taggart, and Sherry Turkle, 'A sociable robot to encourage social interaction among the elderly', in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pp. 3972–3976. IEEE, (2006).
[28] Katrin S Lohan, Karola Pitsch, Katharina J Rohlfing, Kerstin Fischer, Joe Saunders, Hagen Lehmann, Chrystopher Nehaniv, and Britta Wrede, 'Contingency allows the robot to spot the tutor and to learn from interaction', in *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pp. 1–8. IEEE, (2011).
[29] MOBISERV. MOBISERV Project Video. `http://www.youtube.com/watch?feature=player_embedded&v=6DFJwnwzhPs`.
[30] Wendy Moyle, Marie Cooke, Elizabeth Beattie, Cindy Jones, Barbara Klein, Glenda Cook, and Chrystal Gray, 'Exploring the effect of companion robots on emotional expression in older adults with dementia: a pilot randomized controlled trial.', *Journal of gerontological nursing*, **39**(5), 46–53, (2013).
[31] Fotios Papadopoulos, Kerstin Dautenhahn, and Wan Ching Ho, 'Exploring the use of robots as social mediators in a remote human-human collaborative communication experiment', *Paladyn*, **3**(1), 1–10, (2012).
[32] Denise F Polit and Cheryl Tatano Beck, *Nursing research: Principles and methods*, Lippincott Williams & Wilkins, 2004.
[33] Bartosz Przywara, 'Projecting future health care expenditure at european level: drivers, methodology and main results', Technical report, Directorate General Economic and Monetary Affairs (DG ECFIN), European Commission, (2010).
[34] Mark Rice, Alan Newell, and MAGGIE Morgan, 'Forum theatre as a requirements gathering methodology in the design of a home telecommunication system for older adults', *Behaviour & Information Technology*, **26**(4), 323–331, (2007).
[35] Jessica Salter. Chickens helping the elderly tackle loneliness. `http://www.telegraph.co.uk/news/health/11198410/Chickens-helping-the-elderly-tackle-loneliness.html`.
[36] Joe Saunders, Nathan Burke, Kheng Lee Koay, and Kerstin Dautenhahn, 'A user friendly robot architecture for re-ablement and co-learning in a sensorised homes', in *Assistive Technology: From Research to Practice: AAATE 2013*, eds., Pedro Encarnacao, Luis Azevedo, and Gert Jan Gelderblom, volume 33, IOS Press, (2013).
[37] Amanda Sharkey and Noel Sharkey, 'Granny and the robots: ethical issues in robot care for the elderly', *Ethics and Information Technology*, **14**(1), 27–40, (2012).
[38] Will Taggart, Sherry Turkle, and Cory D Kidd, 'An interactive robot in a nursing home: Preliminary remarks', in *Towards Social Mechanisms of Android Science: A COGSCI Workshop*, (2005).
[39] Andrea L Thomaz and Cynthia Breazeal, 'Teachable robots: Understanding human teaching behavior to build more effective robot learners', *Artificial Intelligence*, **172**(6), 716–737, (2008).
[40] Katherine M Tsui and Holly A Yanco, 'Assistive, rehabilitation, and surgical robots from the perspective of medical and healthcare professionals', in *AAAI 2007 Workshop on Human Implications of Human-Robot Interaction, Technical Report WS-07-07 Papers from the AAAI 2007 Workshop on Human Implications of HRI*, (2007).
[41] Sherry Turkle, 'Authenticity in the age of digital companions', *Interaction Studies*, **8**(3), 501–517, (2007).
[42] Shannon Vallor, 'Carebots and caregivers: Sustaining the ethical ideal of care in the twenty-first century', *Philosophy & Technology*, **24**(3), 251–268, (2011).
[43] Kazuyoshi Wada and Takanori Shibata, 'Robot therapy in a care house-results of case studies', in *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, pp. 581–586. IEEE, (2006).
[44] Joshua Wainer, Ben Robins, Farshid Amirabdollahian, and Kerstin Dautenhahn, 'Using the humanoid robot kaspar to autonomously play triadic games and facilitate collaborative play among children with autism', *Autonomous Mental Development, IEEE Transactions on*, **6**(3), 183–199, (2014).
[45] Michael L Walters, Kheng Lee Koay, Dag Sverre Syrdal, Anne Campbell, and Kerstin Dautenhahn, 'Companion robots for elderly people: Using theatre to investigate potential users' views', in *RO-MAN, 2013 IEEE*, pp. 691–696. IEEE, (2013).
[46] Karen Windle, Karen Francis, and Caroline Coomber. SCIE Research briefing 39: Preventing loneliness and social isolation: interventions and outcomes. `http://www.scie.org.uk/publications/briefings/briefing39/index.asp`.

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Robots Have Needs Too: People Adapt Their Proxemic Preferences to Improve Autonomous Robot Recognition of Human Social Signals

Ross Mead[1] and **Maja J Matarić**[2]

Abstract. An objective of autonomous socially assistive robots is to meet the needs and preferences of human users. However, this can sometimes be at the expense of the robot's own ability to understand *social signals* produced by the user. In particular, human preferences of distance (*proxemics*) to the robot can have significant impact on the performance rates of its automated speech and gesture recognition systems. In this work, we investigated how user proxemic preferences changed to improve the robot's understanding human social signals. We performed an experiment in which a robot's ability to understand social signals was artificially varied, either *uniformly* or *attenuated* across distance. Participants ($N = 100$) instructed a robot using speech and pointing gestures, and provided their proxemic preferences before and after the interaction. We report two major findings: 1) people predictably underestimate (based on a Power Law) the distance to the location of robot peak performance; and 2) people adjust their proxemic preferences to be near the *perceived* location of robot peak performance. This work offers insights into the dynamic nature of human-robot proxemics, and has significant implications for the design of social robots and robust autonomous robot proxemic control systems.

1 Introduction

A social robot utilizes natural communication mechanisms, such as speech and gesture, to autonomously interact with humans to accomplish some individual or joint task [2]. The growing field of socially assistive robotics (SAR) is at the intersection of social robotics and assistive robotics that focuses on non-contact human-robot interaction (HRI) aimed at monitoring, coaching, teaching, training, and rehabilitation domains [4]. Notable areas of SAR include robotics for older adults, children with autism spectrum disorders, and people in post-stroke rehabilitation, among others [25, 17].

Consequently, SAR constitutes an important subfield of robotics with significant potential to improve health and quality of life. Because the majority of SAR contexts investigated to date involve one-on-one face-to-face interaction between the robot and the user, how the robot understands and responds to the user is crucial to successful autonomous social robots [1], in SAR contexts and beyond.

One of the most fundamental social behaviors is *proxemics*, the social use of space in face-to-face social encounters [5]. A mobile social robot must position itself appropriately when interacting with the user. However, robot position has a significant impact on the robot's *performance*—in this work, performance is measured by automated speech and gesture recognition rates. Just like electrical signals, human *social signals* (e.g., speech and gesture) are *attenuated* (lose signal strength) based on distance, which dramatically changes the way in which automated recognition systems detect and identify the signal; thus, a proxemic control system that often varies its location and, thus, creates signal attenuation, can be a defining factor in the success or failure of a social robot [16].

In our previous work [16] (described in detail in Section 2.2), we modeled social robot performance attenuated by distance, which was then used to implement an autonomous robot proxemic controller that maximizes its performance during face-to-face HRI; however, this work begged the question as to whether or not people would accept a social robot that positions itself in a way that differs from traditional user proxemic preferences. Would users naturally change their proxemic preferences if they observed differences in robot performance in different proxemic configurations, or would their proxemic preferences persist, mandating that robot developers must improve autonomous speech and gesture recognition systems before social and socially assistive robot technology can be deployed in the real world? This question is the focus of the investigation reported here.

2 Background

The anthropologist Edward T. Hall [5] coined the term "proxemics", and, in [6], proposed that proxemics lends itself well to being analyzed with performance (as measured through sensory experience) in mind. Proxemics has been studied in a variety of ways in HRI; here, we constrain our review of related work to that of *autonomous* HRI[3].

2.1 Comfort-based Proxemics in HRI

The majority of proxemics work in HRI focuses on maximizing user *comfort* during a face-to-face interaction. The results of many human-robot proxemics studies have been consolidated and normalized in [28], reporting mean distances of 0.49–0.71 meters using a variety of robots and conditions. Comfort-based proxemic preferences between humans and the PR2 robot[4] were investigated in [24], reporting mean distances of 0.25–0.52 meters; in [16], we investigated the same preferences using the PR2 in a conversational context, reporting a mean distance of 0.94 meters. Farther proxemic preferences have been measured in [18] and [26], reporting mean distances of 1.0–1.1 meters and 1.7–1.8 meters, respectively.

[1] University of Southern California, USA, email: rossmead@usc.edu
[2] University of Southern California, USA, email: mataric@usc.edu

[3] There is a myriad of related work reporting how humans adapt to various technologies, but this is beyond the scope of this work. For a review, see [8].

[4] https://www.willowgarage.com/pages/pr2/overview

However, results in our previous work [16] suggest that autonomous speech and gesture recognition systems do not perform well using comfort-based proxemic configurations. Speech recognition performed adequately at distances less than 2.5 meters, and face and hand gesture recognition performed adequately at distances of 1.5–2.5 meters; thus, given current technologies, distances for mutual recognition of these social signals is between 1.5 and 2.5 meters, at and beyond the far end of comfort-based proxemic preferences.

2.2 Performance-based Proxemics in HRI

Our previous work utilized advancements in markerless motion capture (specifically, the Microsoft Kinect) to automatically extract proxemic features based on metrics from the social sciences [11, 14]. These features were then used to recognize spatiotemporal interaction behaviors, such as the initiation, acceptance, aversion, and termination of an interaction [12, 14]. These investigations offered insights into the development of proxemic controllers for autonomous social robots, and suggested an alternative approach to the representation of proxemic behavior that goes beyond simple distance and orientation [13]. A probabilistic framework for autonomous proxemic control was proposed in [15, 10] that considers *performance* by maximizing the sensory experience of each agent (human or robot) in a co-present social encounter. The methodology established an elegant connection between previous approaches and illuminated the functional aspects of proxemic behavior in HRI [13], specifically, the impact of spacing on speech and gesture behavior recognition and production. In [16], we formally modeled (using a dynamic Bayesian network [9]) autonomous speech and gesture recognition systems as a function of distance and orientation between a social robot and a human user, and implemented the model as an autonomous proxemic controller, which was shown to maximize robot performance in HRI.

However, while our approach to proxemic control *objectively* maximized the performance of the robot, it also resulted in proxemic configurations that are atypical for human-robot interactions (e.g., positioning itself farther or nearer to the user than preferred). Thus, the question arose as to whether or not people would *subjectively* adopt a technology that places performance over preference, as it might place a burden on people to change their own behaviors to make the technology function adequately.

2.3 Challenges in Human Spatial Adaptation

For humans to adapt their proxemic preferences to a robot, they must be able to accurately identify regions in which the robot is performing well; however, errors in human distance estimation increase nonlinearly with increases in distance, time, and uncertainty [19]. Fortunately, the relationship between human distance estimation and each of these factors is very well represented by Steven's Power Law, ax^b, where x is distance [19, 23]. Unfortunately, these relationships are reported for distances of 3–23 meters, which are farther away than in those with which we are concerned for face-to-face HRI—thus, we cannot use the reported model parameters and must derive our own.

In this work, we investigate how user proxemic preferences change in the presence of a social robot that is recognizing and responding to instructions provided by a human user. Robot performance (ability to understand speech and gesture) is artificially attenuated to expose participants to success and failure scenarios while interacting with the robot. In Section 3, we describe the overall setup in which our investigation took place. In Section 4, we outline the specific procedures, conditions, hypotheses, and participants of our experiment.

3 Experimental Setup

3.1 Materials

The experimental robotic system used in this work was the Bandit upper-body humanoid robot[5] [Figure 1]. Bandit has 19 degrees of freedom: 7 in each arm (shoulder forward-and-backward, shoulder in-and-out, elbow tilt, elbow twist, wrist twist, wrist tilt, grabber open-and-close; left and right arms), 2 in the head (pan and tilt), 2 in the lips (upper and lower), and 1 in the eyebrows. These degrees of freedom allow Bandit to be expressive using individual and combined motions of the head, face, and arms. Mounted atop a Pioneer 3-AT mobile base[6], the entire robot system is 1.3 meters tall.

A Bluetooth PlayStation 3 (PS3) controller served as a remote control interface with the robot. The controller was used by the experimenter (seated behind a one-way mirror [Figure 2]) to step the robot through each part of the experimental procedure (described in Section 4.1)—the decisions and actions taken by the robot during the experiment were completely autonomous, but the timing of its actions were controlled by the press of a "next" button. The controller was also used to record distance measurements during the experiment, and to provide ground-truth information to the robot as to what the participant was communicating (however, the robot autonomously determined how to respond based on the experimental conditions described in Section 4.2).

Four small boxes were placed in the room, located at 0.75 meters and 1.5 meters from the centerline on each side (left and right) of the participant [Figure 2]. During the experiment (described in Section 4.1), the participant instructed the robot to look at these boxes. Each box was labeled with a unique shape and color; in this experiment, the shapes and colors matched the buttons on the PS3 controller: a green triangle, a red circle, a blue cross, and a purple square. This allowed the experimenter to easily indicate to the robot to which box the user was attending (i.e., "ground-truth").

A laser rangefinder on-board the robot was used to measure the distance from the robot to the participant's legs at all times.

Figure 1. The Bandit upper-body humanoid robot.

[5] http://robotics.usc.edu/interaction/?l=Laboratory:Robots#BanditII
[6] http://www.mobilerobots.com/ResearchRobots/P3AT.aspx

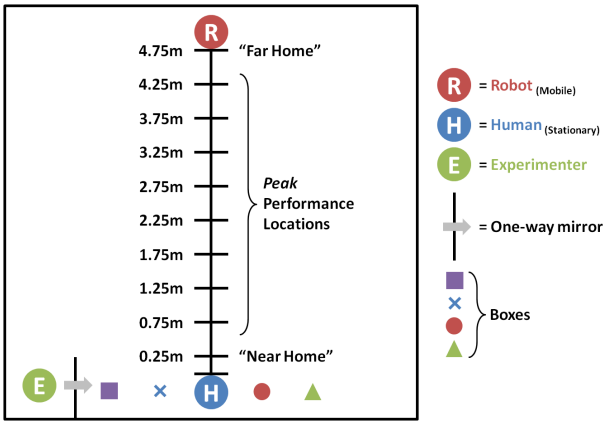


Figure 2. The experimental setup.

3.2 Robot Behaviors

The robot autonomously executed three primary behaviors throughout the experiment: 1) forward and backward base movement, 2) maintaining eye contact with the participant, and 3) responding to participant instructions with head movements and audio cues.

Robot base movement was along a straight-line path directly in front of the participant, and was limited to distances of 0.25 meters (referred to as the "near home" location) and 4.75 meters (referred to as the "far home" location); it returned repeatedly to these "home" locations throughout the experiment. Robot velocity was proportional to the distance to the goal location; the maximum robot speed was 0.3 m/s, which people find acceptable [22].

As the robot moved, it maintained eye contact with the participant. The robot has eyes, but they are not actuated, so the robot's head pitched up or down depending on the location of the participant's head, which was determined by the distance to the participant (from the on-board laser) and the participant's self-reported height. We note that prolonged eye contact from the robot often results in user preferences of increased distance in HRI [24, 18].

The robot provided head movement and audio cues to indicate whether or not it understood instructions provided by the participant (described in Section 4.1.2). If the robot understood the instructions, it provided an *affirmative response* (looking at a box); if the robot did not understand the instructions, it provided a *negative response* (shaking its head). With each head movement, one of two affective sounds were also played to supplement the robot's response; affective sounds were used because robot speech influences proxemic preferences and would have introduced a confound in the experiment [29].

4 Experimental Design

With the described experimental setup, we performed an experiment to investigate user perceptions of robot performance attenuated by distance and its effect on proxemic preferences.

4.1 Experimental Procedure

Participants (described in Section 4.4) were greeted at the door entering the private experimental space, and were informed of and agreed to the nature of the experiment and their rights as a participant, which included a statement that the experiment could be halted at any time.

Participants were then instructed to stand with their toes touching a line on the floor, and were asked to remain there for the duration of the experiment [Figure 2]. The experimenter then provided instructions about the task the participant would be performing.

Participants were introduced to the robot, and were informed that all of its actions were completely autonomous. Participants were told that the robot would be moving along a straight line throughout the duration of the experiment; a brief demonstration of robot motion was provided, in which the robot autonomously moved back and forth between distances of 3.0 meters and 4.5 meters from the participant, allowing them to familiarize themselves with the robot motion. Participants were told that they would be asked about some of their preferences regarding the robot's location throughout the experiment.

Participants were then informed that they would be instructing the robot to look at any one of four boxes (of their choosing) located in the room [Figure 2], and that they could use speech (in English) and pointing gestures. A vocabulary for robot instructions was provided: for speech, participants were told they could say the words "look at" followed by the name of the shape or color of each box (e.g., "triangle", "circle", "blue", "purple", etc.); for pointing gestures, participants were asked to use their left arm to point to boxes located on their left, and their right arm to point to boxes on their right. This vocabulary was provided to minimize any perceptions the person might have that the robot simply did not understand the words or gestures that they used; thus, the use of the vocabulary attempted to maximize the perception that any failures of the robot were due to other factors.

Participants were told that they would repeat this instruction procedure to the robot many times, and that the robot would indicate whether or not it understood their instructions each time using the head movements and audio cues described in Section 3.2.

Participants had an opportunity to ask the experimenter any clarifying questions. Once participant understanding was verified, we proceeded with the experiment.

4.1.1 *Pre-interaction Proxemic Measures (pre)*[7]

The robot autonomously moved to the "far home" location [Figure 2]. Participants were told that the robot would be approaching them, and to say out loud the word "stop" when the robot reached the ideal location at which the participant would have a *face-to-face conversation*[8] with the robot. This pre-interaction proxemic preference from the "far home" location is denoted as pre_{far}.

When the participant was ready, the experimenter pressed a PS3 button to start the robot moving. When the participant said "stop", the experimenter pressed another button to halt robot movement. The experimenter pressed another button to record the distance between the robot and the participant, as measured by the on-board laser.

Once the pre_{far} distance was recorded, the experimenter pressed another button, and the robot autonomously moved to the "near home" location [Figure 2]; the participant was informed that the robot would be approaching to this location and would stop on its own. The process was repeated with the robot backing away from the participant, and the participant saying "stop" when it reached the ideal location for conversation. This pre-interaction proxemic preference from the "near home" location is denoted as pre_{near}.

[7] Measures are provided inline with the experimental procedure to provide an order of events as they occurred in the experiment.

[8] Related work in human-robot proxemics asks the participant about locations at which they feel *comfortable* [24], yielding proxemic preferences very near to the participant. Our general interest is in face-to-face human-robot conversational interaction, with proxemic preference farther from the participant [16, 26, 27], hence the choice of wording.

From pre_{far} and pre_{near}, we calculated and recorded the average pre-interaction proxemic preference, denoted as pre[9].

4.1.2 *Interaction Scenario*

After determining pre-interaction proxemic preferences, the robot returned to the "far home" location. The experimenter then repeated to participants the instructions about the task they would be performing with the robot. When participants verified that they understood the task and indicated that they were ready, the experimenter pressed a button to proceed with the task.

The robot autonomously visited ten pre-determined locations [Figure 2]. At each location, the robot responded to instructions from the participant to look at one of four boxes located in the room [Figure 2]. Five instruction-response interactions were performed at each location, after which the robot moved to the next location along its path; thus, each participant experienced a total of 50 instruction-responses interactions. Robot goal locations were in 0.5-meter intervals inclusively between the "near home" location (0.25 meters) and "far home" location (4.75 meters) along a straight-line path in front of the participant [Figure 2]. Locations were visited in a sequential order; for half of the participants, the robot approached from the "far home" location (i.e., farthest-to-nearest order), and, for the other half of participants, the robot backed away from "near home" location (i.e., nearest-to-farthest order); this was done to reduce any ordering effects [19].

To controllably simulate social signal attenuation at each location, robot performance was artificially manipulated as a function of the distance to the participant (described in Section 4.2). After each instruction provided by the participant, the experimenter provided to the robot (via a remote control interface) the ground-truth of the instruction; the robot then determined whether or not it would have understood the instruction based on a prediction from a performance vs. distance curve (specified by the assigned experimental condition described in Section 4.2), and provided either an *affirmative response* or a *negative response* to the participant indicating its successful or failed understanding of the instruction, respectively.

The entire interaction scenario lasted 10-15 minutes.

4.1.3 *Post-interaction Proxemic Measures (post)*

After the robot visited each of the ten locations, it autonomously returned to the "far home" location. The experimenter then repeated the procedure for determining proxemic preferences described in Section 4.1.1. This process generated post-performance proxemic preferences from the "far home" and "near home" locations, as well as their average, denoted $post_{far}$, $post_{near}$, and $post$[10], respectively.

4.1.4 *Perceived Peak Location Measures (perc)*

Finally, after collecting post-interaction proxemic preferences, the experimenter repeated the procedure described in Section 4.1.1 to determine participant perceptions of the location of peak performance. This process generated perceived peak performance locations from the "far home" and "near home" locations, as well as their average, denoted $perc_{far}$, $perc_{near}$, and $perc$[11], respectively.

[9] Post-hoc analysis revealed no statistically significant difference between pre_{far} and pre_{near} measurements, hence why we rely on pre.

[10] Post-hoc analysis revealed no statistically significant difference between $post_{far}$ and $post_{near}$ measurements, hence why we rely on $post$.

[11] Post-hoc analysis revealed no statistically significant difference between $perc_{far}$ and $perc_{near}$ measurements, hence why we rely on $perc$.

4.2 Experimental Conditions

We considered two performance vs. distance conditions; 1) a **"uniform performance" condition**, and 2) an **"attenuated performance" condition**. Overall robot performance for each condition was held at a constant 40%[12]—that is, for each participant, the robot provided 20 affirmative responses and 30 negative responses distributed across 50 instructions. The way in which these responses were distributed across locations varied between conditions.

In the **uniform performance condition**, robot performance was the same (40%) across across all locations [Figures 3 and 4]. Thus, at each of the ten locations visited, the robot provided two affirmative and three negative responses, respectively. This condition served as a baseline of participant proxemic preferences within the task.

In the **attenuated performance condition**, robot performance varied with distance proportional to a Gaussian distribution centered a location of "peak performance" ($M = peak$, $SD = 1.0$) [Figures 3 and 4]. Due to differences in pre-interaction proxemic preferences, we could not select a single value for $peak$ that provided a similar experience between participants without introducing other confounding factors (e.g., the $peak$ not being at a location that the robot visited or distances beyond the "home" locations). To alleviate this, we opted to select multiple peak performance locations, exploring the space of human responses to robot performance differences at a variety of distances. We selected the eight locations non-inclusively between the "near home" and "far home" locations as the peak performance locations [Figure 2]; the "near home" and "far home" locations were not included in the set of peaks to ensure that participants were always exposed to an actual *peak* in performance, rather than just a *trend*. Peak performance locations were varied between participants.

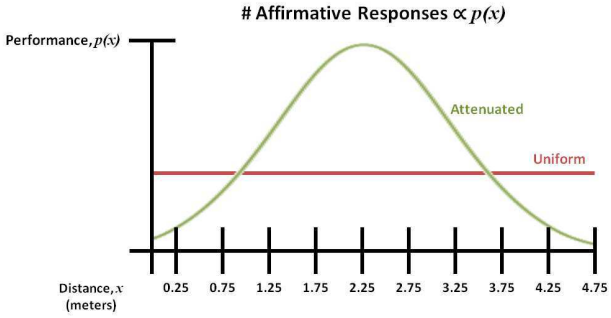


Figure 3. The performance curves of the **uniform** and **attenuated** conditions. In this example, $peak = 2.25$ (in meters), so the attenuated performance curve parameters is $M = peak = 2.25$, $SD = 1.0$. The number of affirmative responses at a distance, x, from the user is proportional to $p(x)$, the evaluation of the performance curve at x.

The distribution of affirmative responses for all conditions is presented in Figure 4. The number of affirmative responses was normalized to 20 (40%) to ensure a consistent user experience of overall robot performance across all conditions. In the **attenuated performance condition**, the number of affirmative responses at $peak$ was always the 5 (i.e., perfect performance), and the number of affirmative responses at other locations were always less than that of the peak to ensure that participants were exposed to an actual peak. At each location, the order in which the five responses were provided was random.

[12] This value was selected because it is an average performance rate predicted by our results in [16] for typical human-robot proxemic preferences.

Performance Condition		# Affirmative Responses vs. Distance									
Distance, x (meters) →		0.25	0.75	1.25	1.75	2.25	2.75	3.25	3.75	4.25	4.75
Uniform →		2	2	2	2	2	2	2	2	2	2
Attenuated: peak → (meters)	0.75	4	***5***	4	3	1	1	1	1	0	0
	1.25	3	4	***5***	4	3	1	0	0	0	0
	1.75	1	3	4	***5***	4	3	0	0	0	0
	2.25	0	1	3	4	***5***	4	3	0	0	0
	2.75	0	0	0	3	4	***5***	4	3	1	0
	3.25	0	0	0	0	3	4	***5***	4	3	1
	3.75	0	0	0	0	1	3	4	***5***	4	3
	4.25	0	0	1	1	1	1	3	4	***5***	4

Figure 4. The distribution of affirmative responses provided by the robot across conditions. Manipulated values are highlighted in **bold italics**.

4.3 Experimental Hypotheses

Within these conditions, we had three central hypotheses:

H1: In the **uniform performance condition**, there will be no significant change in participant proxemic preferences.

H2: In the **attenuated performance conditions**, participants will be able to identify a relationship between robot performance and human-robot proxemics.

H3: In the **attenuated performance conditions**, participants will adapt their proxemic preferences to improve robot performance.

4.4 Participants

We recruited 100 participants (50 male, 50 female) from our university campus community. Participant race was diverse (67 white/Caucasian, 26 Asian, 3 black/African-American, 3 Latino/Latina, and 1 mixed-race). All participants reported proficiency in English and had lived in the United States for at least two years (i.e., acclimated to U.S. culture). Average age (in years) of participants was 22.26 ($SD = 4.31$), ranging from 18 to 39. Based on a seven-point scale, participants reported moderate familiarity with technology ($M = 3.98, SD = 0.85$). Average participant height (in meters) was 1.74 ($SD = 0.10$), ranging from 1.52 to 1.93. Related work reports how human-robot proxemics is influenced by gender and technology familiarity [24], culture [3], and height [7, 21].

The 100 participants were randomly assigned to a performance condition, with $N = 20$ in the **uniform performance condition** and $N = 80$ in the **attenuated performance condition**. In the **attenuated performance condition**, the 80 participants were randomly assigned one of the eight peak performance locations (described in Section 4.2) with $N = 10$ for each $peak$. Neither the participant nor the experimenter was aware of the condition assigned.

5 Results and Discussion

We analyzed data collected in our experiment to test our three hypotheses (described in Section 4.3), and evaluated their implications for autonomous social robots and human-robot proxemics.

To provide a baseline of our robot for comparison in general human-robot proxemics, we consolidated and analyzed pre-interaction proxemic preferences (pre) across all conditions ($N = 100$), as the data had not been influenced by robot performance. The participant pre-interaction proxemic preference (in meters) was determined to be 1.14 ($SD = 0.49$) for our robot system, which is consistent with [18] and our previous work [16], but twice as far away as related work has reported for robots of a similar form factor [28, 24].

5.1 H1: Pre- vs. Post-interaction Locations

To test **H1**, we compared average pre-interaction proxemic preferences (pre) to average post-interaction proxemic preferences ($post$) of participants in the **uniform performance condition**.

A paired *t*-test revealed a statistically significant change in participant proxemic preferences between pre ($M = 1.12, SD = 0.51$) and $post$ ($M = 1.39, SD = 0.63$); $t(38) = 1.49, p = 0.02$. Thus, our hypothesis **H1** is rejected.

The rejection of this hypothesis does not imply a failure of the experimental procedure, but, rather, provides important insights that must be considered for subsequent analyses (and for related work in proxemics). This result suggests that there might be something about the context of the interaction scenario itself that influenced participant proxemic preferences. To address any influence the interaction scenario might have on subsequent analyses, we define a *contextual offset*, θ, as the average difference in participant post-interaction and pre-interaction proxemic preferences ($M = 0.27, SD = 0.48$); this θ value will be subtracted from ($post - pre$) values in Section 5.3 to normalize for the interaction context.

5.2 H2: Perceived vs. Actual Peak Locations

To test **H2**, we compared participant perceived locations of peak performance ($perc$) to actual locations of peak performance ($peak$) in the **attenuated performance conditions** [Figure 5].

Steven's Power Law, ax^b, has previously been used to model human distance estimation as a function of actual distance [19], and is generally well representative of human-perceived vs. actual stimuli [23]. However, existing Power Laws relevant to our work only seem to pertain to distances of 3–23 meters, which are beyond the range of the natural face-to-face communication with which we are concerned. Thus, our goal here is to model our own experimental data to establish a Power Law for $perc$ vs. $peak$ at locations more relevant to HRI (0.75–4.25 meters), which we can then evaluate to test **H2**.

Immediate observations of our data suggested that the data appear to be heteroscedastic [Figure 5]—in this case, the variance seems to increase with distance from the participant, which means we should not use traditional statistical tests. The Breusch-Pagan test for non-constant variance (NCV) confirmed this intuition; $\chi^2(1, N = 100) = 15.79, p < 0.001$. A commonly used and accepted approach to alleviate our heteroscedasticity is to transform the $perc$ and $peak$ data to a log-log scale. While not applicable to all datasets, this approach served as an adequate approximation for our purposes [Figure 6]; it also enabled us to perform a regression analysis to determine parameter values for the Power Law coefficient and exponent, $a = 1.3224$ and $b = 0.5132$, respectively. With these parameters, we identified that $peak$ was a strongly correlated and very significant predictor of $perc$; $R^2 = 0.4951, F(1, 78) = 76.48, p < 0.001$. Thus, our hypothesis **H2** is supported.

This result suggests that people are able to identify a relationship between robot performance and human-robot proxemics, but they will predictably underestimate the distance, x, to the location of peak performance based on the Power Law equation $1.3224x^{0.5132}$. While human estimation of the location of peak performance is suboptimal, it is possible that repeated exposure to the robot over multiple sessions might yield more accurate results. This follow-up hypothesis will be formally tested in a planned longitudinal study in future work (described in Section 6).

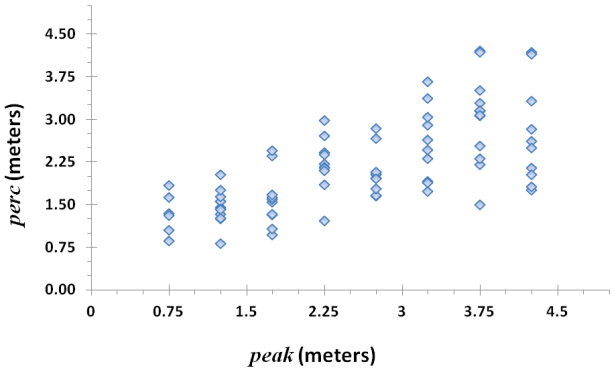


Figure 5. Participant perceived location of robot peak performance ($perc$) vs. actual location of robot peak performance ($peak$). Note the heteroscedasticity of the data, which prevents us from performing traditional statistical analyses without first transforming the data (shown in Figure 6).

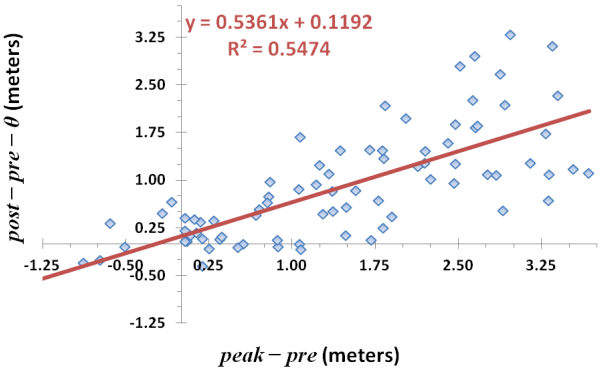


Figure 7. Changes in participant pre-/post-interaction proxemic preferences (pre and $post$, respectively; θ is the *contextual offset* defined in Section 5.1) vs. distance from participant pre-interaction proxemic preference (pre) to the actual location of robot peak performance ($peak$).

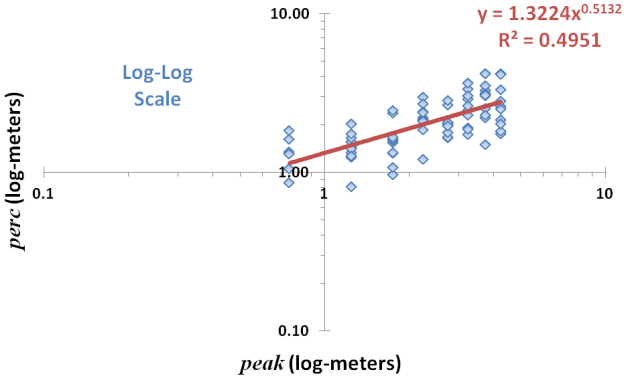


Figure 6. Participant perceived location of robot peak performance ($perc$) vs. actual location of robot peak performance ($peak$) on a log-log scale, reducing the effects of heteroscedasticity and allowing us to perform regression to determine parameters of the Power Law, ax^b.

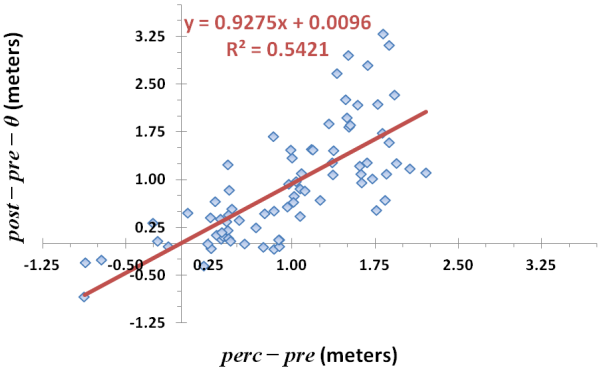


Figure 8. Changes in participant pre-/post-interaction proxemic preferences (pre and $post$, respectively; θ is the *contextual offset* defined in Section 5.1) vs. distance from participant pre-interaction proxemic preference (pre) to the perceived location of robot peak performance ($perc$).

5.3 H3: Preferences vs. Peak Locations

To test **H3**, we compared changes in participant pre-/post-interaction proxemic preferences ($post - pre - \theta$) to the distance from the participant pre-interaction proxemic preference to either a) the actual location of robot peak performance ($peak - pre$) [Figure 7], or b) the perceived location of robot peak performance ($perc - pre$) [Figure 8], both in the **attenuated performance conditions**.

Data for ($post - pre - \theta$) vs. both ($peak - pre$) and ($perc - pre$) were heteroscedastic, as indicated by Breusch-Pagan NCV tests: $\chi^2(1, N = 100) = 18.81, p < 0.001$; and $\chi^2(1, N = 100) = 13.55, p < 0.001$; respectively. This is intuitive, as the data for perceived ($perc$) vs. actual ($peak$) locations of peak performance were also heteroscedastic [Figure 5]. The log-transformation approach that we used in Section 5.2 did not perform well in modeling these data; thus, we needed to use an alternative approach. We opted to utilize a Generalized Linear Model [20] because it allowed us to model the variance of each measurement separately as a function of predicted values and, thus, perform appropriate statistical tests for significance.

We first modeled changes in participant proxemic preferences ($post - pre - \theta$) vs. distance from pre-interaction proxemic preference to the actual location of peak performance ($peak - pre$). In the ideal situation (for the robot), these match one-to-one—in other words, the participant meets the needs of the robot entirely by changing proxemic preferences to be centered at the peak of robot performance. Unfortunately for the robot, this was not the case. We detected a strongly correlated and statistically significant relationship between participant proxemic preference change and distance from pre-interaction preference to the peak location ($R^2 = 0.5474, \beta = 0.5361, t(98) = 9.71, p < 0.001$), but participant preference change only got the robot approximately halfway ($\beta = 0.5361$) to its location of peak performance [Figure 7]. Why is this?

Recall that results reported in Section 5.2 suggested that, while people do perceive a relationship between robot performance and distance, their ability to accurately identify the location of robot peak performance diminishes based on the distance to it as governed by a Power Law. Were participants *trying* to maximize robot performance, but simply adapting their preferences to a suboptimal location?

We investigated this question by considering changes in participant proxemic preferences ($post - pre - \theta$) vs. distance from pre-interaction proxemic preference to the perceived location of peak performance ($perc - pre$). If the participant was adapting their proxemic preferences to accommodate the needs of the robot, then these

should match one-to-one. A Generalized Linear Model was fit to these data, and yielded a strongly correlated and statistically significant relationship between changes in proxemic preferences and perceptions of robot performance ($R^2 = 0.5421, \beta = 0.9275, t(98) = 9.61, p < 0.001$) [Figure 8]. Thus, our hypothesis **H3** is supported.

The near one-to-one relationship ($\beta = 0.9275$) between post-interaction proxemic preferences and participant perceptions of robot peak performance is compelling, suggesting that participants adapted their proxemic preferences almost entirely to improve robot performance in the interaction.

5.4 Discussion

These results have significant implications for the design of social robots and autonomous robot proxemic control systems, specifically, in that people's proxemic preferences will likely change as the user interacts with and comes to understand the needs of the robot.

As illustrated in our previous work [16], the locations of on-board sensors for social signal recognition (e.g., microphones and cameras), as well as the automated speech and gesture recognition software used, can have significant impacts on the performance of the robot in autonomous face-to-face social interactions. As our now-reported results suggest that people will adapt their behavior in an effort to improve robot performance, it is anticipated that human-robot proxemics will vary between robot platforms with different hardware and software configurations based on factors that are 1) not specific to the user (unlike culture [3], or gender, personality, or familiarity with technology [24]), 2) not observable to the user (unlike height [7, 21], amount of eye contact [24, 18], or vocal parameters [29]), or 3) not observable to the robot developer. User understanding of the relationship between robot performance and human-robot proxemics is a latent factor that only develops through repeated interactions with the robot (perhaps expedited by the robot communicating its predicted error); fortunately, our results indicate that user understanding will develop in a predictable way. Thus, it is recommended that social robot developers consider and perhaps model robot performance as a function of conditions that might occur in dynamic proxemic interactions with human users to better predict and accommodate how the people will actually use the technology. This dynamic relationship, in turn, will enable more rich autonomy for social robots by improving the performance of their own automated recognition systems.

If developers adopt models of robot performance as a factor contributing to human-robot proxemics, then it follows that proxemic control systems might be designed to expedite the process of autonomously positioning the robot at an optimal distance from the user to maximize robot performance while still accommodating the initial personal space preferences of the user. This was the focus of our previous work [16], which treated proxemics as an optimization problem that considers the production and perception of social signals (speech and gesture) as a function of distance and orientation. Recall that an objective of the now-reported work was to address questions regarding whether or not users would accept a robot that positions itself in locations that might differ from their initial proxemic preferences. The results in this work (specifically, in Section 5.3) support the notion that user proxemic preferences will change through interactions with the robot as its performance is observed, and that the new user proxemic preference will be at the *perceived* location of robot peak performance. An extension of this result is that, through repeated interactions, user proxemic preferences will further adapt and eventually converge to the *actual* location of robot peak performance, a hypothesis that we will investigate in future work.

6 Future Work

Our experimental conditions (described in Section 4.2) were specifically selected to strongly expose a relationship (if one existed) between human proxemic preferences and robot performance—the robot achieved perfect success rates (100%) at "peak" locations and perfect failure rates (0%) at other locations, and these success/failure rates were distributed proportional to a Gaussian distribution with constant variance. Now that we have identified that a relationship exists, our next steps will examine how the relationship changes over time or with other related factors. A longitudinal study over multiple sessions will be conducted to determine if changes in preferences persist from one interaction to the next, and if user proxemic preferences will continue to adapt and eventually converge to locations of robot peak performance through repeated interactions. Other future work will follow the same experimental procedure described in Section 4.1, but will adjust the **attenuated performance condition** (described in Section 4.2) to consider how the relationship changes with 1) distributions of lower or higher variance, 2) lower maximum performance or higher minimum performance, 3) more realistic non-Gaussian distributions, and 4) the interactions between distributions of actual multimodal recognition systems [16].

This perspective opens up a whole new theoretical design space of human-robot proxemic behavior. The general question is, "How will people adapt their proxemic preferences in any given *performance field*?", in which performance varies with a variety of factors, such as distance, orientation, and environmental interference. The follow-up question then asks, "How can the robot expedite the process of establishing an appropriate human-robot proxemic configuration within the performance field without causing user discomfort?" This will be a focus of future work, and will extend our prior work on modeling human-robot proxemics to improve robot proxemic controllers [16].

7 Summary and Conclusions

An objective of autonomous socially assistive robots is to meet the needs and preferences of a human user [4]. However, this can sometimes be at the expense of the robot's own ability to understand social signals produced by the user. In particular, human proxemic preferences with respect to a robot can have significant impacts on the performance rates of its automated speech and gesture recognition systems [16]. This means that, for a successful interaction, the robot has needs too—and these needs might not be consistent with and might require changes in the proxemic preferences of the human user.

In this work, we investigated how user proxemic preferences changed to improve the robot's understanding of human social signals (described in Section 4). We performed an experiment in which a robot's performance was artificially varied, either *uniformly* or *attenuated* across distance. Participants ($N = 100$) instructed a robot using speech and pointing gestures, and provided their proxemic preferences before and after the interaction.

We report two major findings. First, people predictably underestimate the distance to the location of robot peak performance; the relationship between participant perceived and actual distance to the location of peak performance is represented well by a Power Law (described in Section 5.2). Second, people adjust their proxemic preferences to be near the *perceived* location of maximum robot understanding (described in Section 5.3). This work offers insights into the dynamic nature of human-robot proxemics, and has significant implications for the design of social robots and robust autonomous robot proxemic control systems (described in Section 5.4).

Traditionally, we focus on our attention on ensuring the robot is meeting the needs of the user with little regard to the impact it might have on the robot itself; it is often an afterthought, or something that we, as robot developers, have to "fix" with our systems. While robot developers will continue to improve upon our autonomous systems, our results suggest that even novice users are willing to adapt their behaviors in an effort to help the robot better understand and perform its tasks. Automated recognition systems are not and will likely never be perfect, but this is no reason to delay the development, deployment, and benefits of social and socially assistive robot technologies. Robots have needs too, and human users will attempt to meet them.

ACKNOWLEDGEMENTS

This work is supported in part by an NSF Graduate Research Fellowship, the NSF National Robotics Initiative (IIS-1208500), and an NSF CNS-0709296 grant.

We thank Aditya Bhatter, Lizhi Fan, Jonathan Frey, Akash Metawala, Kedar Prabhu, and Cherrie Wang for their assistance in recruiting participants and conducting the experiment.

REFERENCES

[1] C. Breazeal, 'Social interactions in hri: The robot view', *IEEE Transactions on Man, Cybernetics and Systems*, **34**(2), 181–186, (2003).

[2] C. Breazeal, *Designing Sociable Robots*, MIT Press, Cambridge, Massachusetts, 2004.

[3] G. Eresha, M. Haring, B. Endrass, E. Andre, and M. Obaid, 'Investigating the influence of culture on proxemic behaviors for humanoid robots', in *22nd IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2013, pp. 430–435, (2013).

[4] D.J. Feil-Seifer and M.J. Matarić, 'Defining socially assistive robotics', in *International Conference on Rehabilitation Robotics*, ICRR'05, pp. 465–468, Chicago, Illinois, (2005).

[5] E. T. Hall, *The Silent Language*, Doubleday Company, New York, New York, 1959.

[6] E.T Hall, 'A system for notation of proxemic behavior', *American Anthropologist*, **65**, 1003–1026, (1963).

[7] Y. Hiroi and A. Ito, *Influence of the Size Factor of a Mobile Robot Moving Toward a Human on Subjective Acceptable Distance*.

[8] P.H. Kahn, *Technological Nature: Adaptation and the Future of Human Life*, MIT Press, Cambridge, Massachusetts, 2011.

[9] D. Koller and N. Friedman, *Probabilistic Graphical Models*, MIT Press, Cambridge, Massachusetts, 2009.

[10] R. Mead, 'Space, speech, and gesture in human-robot interaction', in *Doctoral Consortium of International Conference on Multimodal Interaction*, ICMI'12, pp. 333–336, Santa Monica, California, (2012).

[11] R. Mead, A. Atrash, and M. J. Matarić, 'Proxemic feature recognition for interactive robots: Automating metrics from the social sciences', in *International Conference on Social Robotics*, pp. 52–61, Amsterdam, Netherlands, (2011).

[12] R. Mead, A. Atrash, and M. J. Matarić, 'Recognition of spatial dynamics for predicting social interaction', in *6th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 201–202, Lausanne, Switzerland, (2011).

[13] R. Mead, A. Atrash, and M. J. Matarić, 'Representations of proxemic behavior for human-machine interaction', in *NordiCHI 2012 Workshop on Proxemics in Human-Computer Interaction*, NordiCHI'12, Copenhagen, Denmark, (2012).

[14] R. Mead, A. Atrash, and M. J. Matarić, 'Automated proxemic feature extraction and behavior recognition: Applications in human-robot interaction', *International Journal of Social Robotics*, **5**(3), 367–378, (2013).

[15] R. Mead and M. J. Matarić, 'A probabilistic framework for autonomous proxemic control in situated and mobile human-robot interaction', in *7th ACM/IEEE International Conference on Human-Robot Interaction*, HRI'12, pp. 193–194, Boston, Massachusetts, (2012).

[16] R. Mead and M. J. Matarić, 'Perceptual models of human-robot proxemics', in *14th International Symposium on Experimental Robotics*, ISER'14, p. to appear, Marrakech/Essaouira, Morocco, (2014).

[17] R. Mead, E. Wade, P. Johnson, A. St. Clair, S. Chen, and M. J. Mataric., 'An architecture for rehabilitation task practice in socially assistive human-robot interaction', in *Robot and Human Interactive Communication*, pp. 404–409, (2010).

[18] J. Mumm and B. Mutlu, 'Human-robot proxemics: Physical and psychological distancing in human-robot interaction', in *6th ACM/IEEE International Conference on Human-Robot Interaction*, HRI-2011, pp. 331–338, Lausanne, (2011).

[19] A Murata, 'Basic characteristics of human's distance estimation', in *1999 IEEE International Conference on Systems, Man, and Cybernetics*, volume 2 of *SMC'99*, pp. 38–43, (1999).

[20] J. Nelder and R. Wedderburn, 'Generalized linear models', *Journal of the Royal Statistical Society*, **135**(3), 370–384, (1972).

[21] I. Rae, L. Takayama, and B. Mutlu, 'The influence of height in robot-mediated communication', in *8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI-2013, pp. 1–8, Tokyo, Japan, (2005).

[22] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita, 'How to approach humans?: Strategies for social robots to initiate interaction', in *HRI*, pp. 109–116, (2009).

[23] S.S. Stevens, 'On the psychological law', *Psychological Review*, **64**, 153–181, (2007).

[24] L. Takayama and C. Pantofaru, 'Influences on proxemic behaviors in human-robot interaction', in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS'09, pp. 5495–5502, (2009).

[25] A. Tapus, M.J. Matarić, and B. Scassellati, 'The grand challenges in socially assistive robotics', *IEEE Robotics and Automation Magazine*, **14**(1), 35–42, (2007).

[26] Elena Torta, Raymond H. Cuijpers, and James F. Juola, 'Design of a parametric model of personal space for robotic social navigation', *International Journal of Social Robotics*, **5**(3), 357–365, (2013).

[27] Elena Torta, Raymond H. Cuijpers, James F. Juola, and David van der Pol, 'Design of robust robotic proxemic behaviour', in *Proceedings of the Third International Conference on Social Robotics*, ICSR'11, pp. 21–30, (2011).

[28] M.L. Walters, K. Dautenhahn, R.T. Boekhorst, K.L. Koay, D.S. Syrdal, and C.L.: Nehaniv, 'An empirical framework for human-robot proxemics', in *New Frontiers in Human-Robot Interaction*, pp. 144–149, Edinburgh, (2009).

[29] M.L. Walters, D.S. Syrdal, K.L. Koay, K. Dautenhahn, and R. te Boekhorst, 'Human approach distances to a mechanical-looking robot with different robot voice styles', in *The 17th IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2008, pp. 707–712, (2008).

4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

A new biomimetic approach towards educational robotics: the Distributed Adaptive Control of a Synthetic Tutor Assistant

Vasiliki Vouloutsi[1], Maria Blancas Munoz[1], Klaudia Grechuta[1], Stephane Lallee[1,2], Armin Duff[1], Jordi-Ysard Puigbo Llobet[1] and **Paul F.M.J. Verschure[1,3]**

Abstract. Many fields can profit from the introduction of robots, including that of education. In this paper, our main focus is the advancement of the Synthetic Tutor Assistant (STA), a robot that will act as a peer for knowledge transfer. We propose a theory of a tutoring robotic application that is based on the Distributed Adaptive Control (DAC) theory: a layered architecture that serves as the framework of the proposed application. We describe the main components of the STA and we evaluate the implementation within an educational scenario.

1 INTRODUCTION

Robots are now able to interact with humans in various conditions and situations. Lately, there has been an increased attempt to develop socially interactive robots, that is, robots with the ability to display social characteristics: use natural communicative cues (such as gestures or gaze), express emotional states or even establish social relationships, all of which are important when a peer-to-peer interaction takes place [20]. In fact, given the current technological advancements, we are now able to develop robotic systems that are able to deal with both physical and social environments. One of the greatest challenges in the design of social robots is to correctly identify all those various factors that affect social interaction and act in accordance [43]. Indeed, different studies have shown that the complexity in the behavior of robots affect how humans interact with robots and perceive them [30, 55, 7, 52].

There are many fields that can profit from the introduction of robots [13], they include health care [9], entertainment [18], social partners [8] or education [21, 41]. Here we focus on the latter, by advancing the notion of the Synthetic Tutor Assistant (STA) (see section 3) which is pursued in the European project entitled Expressive Agents for Symbiotic Education and Learning (EASEL). In this perspective, the robot STA will not act as the teacher, but rather as a peer of the learner to assist in knowledge acquisition. It has been shown that robots can both influence the performance of the learner [41] and their motivation to learn [29]. One of the main advantages of employing a robotic tutor is that it can provide assistance at the level of individual learners, given that the robot can have the ability to learn and adapt based on previous interactions.

[1] Laboratory of Synthetic Perceptive, Emotive and Cognitive Systems - SPECS, Universitat Pompeu Fabra
[2] Institute For Infocomm Research, A*STAR, Singapore
[3] Center of Autonomous Systems and Neurorobotics - NRAS, Catalan Institute of Advanced Studies - ICREA, email: paul.verschure@upf.edu

Through education, people acquire knowledge, develop skills and capabilities and consequently form values and habits. Although there exist several educational approaches that could be considered, here, we will focus on Constructivism [35]. Constructivism proposes an educational approach based on collaboration, learning through making, and technology-enhanced environments. Such approach aims at constructing social interaction between the participant and the STA as it implies a common goal for both learners-players [45].

We consider tutoring as the structured process in which knowledge and skills are transferred to an autonomous learner through a guided process based on the individual traits of the learner. Here we present an approach where both the user model and the STA are based on a neuroscientifically grounded cognitive architecture called Distributed Adaptive Control (DAC) [51, 47]. On one hand, DAC serves as the theory which defines the tutoring scenario: it allows us to derive a set of key principles that are general for all learning processes. On the other hand, it is the core for the implementation of the control architecture of the STA, the robotic application. Following the layered DAC architecture, we propose the STA that will deploy tutoring strategies of increasing levels of complexity depending on the performance and capabilities of the learner. The DAC theory serves as both the basis for the tutoring robotic application, user model as well as for the implementation of the STA. Such design guarantees a tight interplay between the robotic application, the user and their interaction.

The present study is organized as follows: first, we present the background theory of the tutoring robotic application, the DAC theory, and we describe the tutoring model applied. Furthermore, we introduce the key implementation features of the STA based on DAC. To assess the first implementation of our system, we devised a pilot study where the STA performs the role of a peer-teacher in an educational scenario. The proposed scenario consists of a pairing game where participants have to match an object to its corresponding category. The setup was tested with both children and adults. The game had three levels of increased difficulty. Questionnaires distributed after every interaction to the players were used to assess the STA's ability to transfer knowledge.

2 DAC COGNITIVE ARCHITECTURE AND LEARNING

To provide a model of perception, cognition and action for our system, we have implemented the DAC architecture. [51, 47]. DAC is a theory of mind and brain, and its implementation serves as a real-time

neuronal model for perception, cognition and action (for a review see [49]). DAC will serve both as the basis for the tutoring model as well as the core of the implementation of the STA.

2.1 Distributed Adaptive Control (DAC)

Providing a real-time model for perception, cognition and action, DAC has been formulated in the context of classical and operant conditioning: learning paradigms for sensory-sensory, multi-scale sensorimotor learning and planning underlying any form of learning. According to DAC, the brain is a layered control architecture that is subdivided into functional segments sub-serving the processing of the states of the world, the self, interaction through action [48], and it is dominated by parallel and distributed control loops.

DAC proposes that in order to act upon the environment (or to realize the *How?* of survival) the brain has to answer four fundamental questions, continuously and in real-time: *Why, What, Where and When*, forming the H4W problem [50, 49]. However, in a world filled with agents, the H4W problem does not seem enough to ensure survival; an additional key question needs to be answered: *Who?*, which shifts the H4W towards a more complex problem, H5W [46, 39].

To answer the H5W problem, the DAC architecture comprises of four layers: Somatic, Reactive, Adaptive and Contextual, intersected by three columns: states of the world (exosensing), states of self (endosensing) and their interface in action (Figure 1). The Somatic Layer represents the body itself and the information acquired from sensations, needs and actions. The Reactive Layer comprises fast, predefined sensorimotor loops (reflexes) that are triggered by low complexity perceptions and are coupled to specific affective states of the agent. It supports the basic functionality of the Somatic Layer in terms of reflexive behavior and constitutes the main behavioral system based on the organism's physical needs. Behavior emerges from the satisfaction of homeostatic needs, which are also regulated by an integrative allostatic loop that sets the priorities and hierarchies of all the competitive homeostatic systems. Thus, behavior serves the reduction of needs [25] controlled by the allostatic controller [42].

The Adaptive Layer extends the sensorimotor loops of the Reactive Layer with acquired sensor and action states, allowing the agent to escape the predefined reflexes and employs mechanisms to deal with unpredictability through learning [14]. The Contextual Layer uses the state-space acquired by the Adaptive Layer to generate goal oriented behavioral plans and policies. This layer includes mechanisms for short, long-term and working memory, formating sequential representations of states of the environment and actions generated by the agent or its acquired sensorimotor contingencies. The DAC architecture has been validated through robotic implementations[19, 42], expanded to capture social interactions with robots [52, 39] as well as providing novel approaches towards rehabilitation [47]. Here, the implementation of DAC serves two main purposes. On the one hand, it acts as the grounding theory for the pedagogical model: it allows us to derive and deduce a set of key principles that are general for all learning processes. On the other hand, DAC is the core for the implementation of the STA.

2.2 Phases of learning

Based on the formal description of learning from the DAC architecture which has been shown to be Bayes optimal [48], we will focus on two main principles as it has a dual role within EASEL. On one hand, DAC is the core for the implementation of the Synthetic Tutor

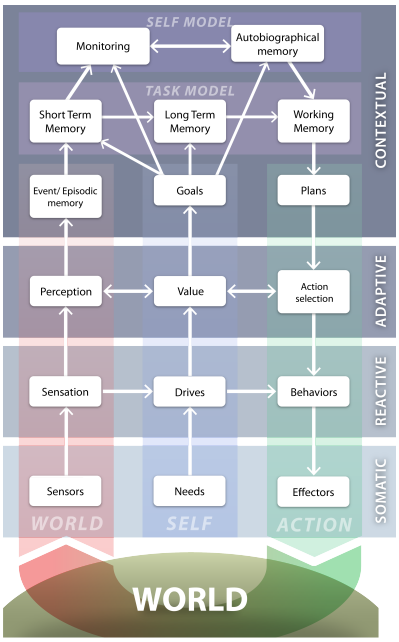


Figure 1. The DAC architecture and its four layers (somatic, reactive adaptive and contextual). Across the layers we can distinguish three functional columns of organization: world (exosensing), self (endosensing) and action (the interface to the world through action). The arrows show the flow of information. Image adapted from [49].

Assistant (STA). On the other hand, following the layered architecture, the STA deploys pedagogical strategies of increasing levels of complexity.

First, DAC predicts that learning is bootstrapped and organized along a hierarchy of complexity: the Reactive Layer allows for exploring the world and gaining experiences, based on which the Adaptive Layer learns the states of the world and their associations; only after these states are well consolidated, the Contextual Layer can extract consistent rules and regularities. We believe that the same hierarchy is applicable in the pedagogical context. Secondly, DAC predicts that in order to learn and consolidate new material, the learner undergoes a sequence of learning phases: resistance, confusion and resolution. *Resistance* is a mechanism that results from defending one's own (in)competence level against discrepancies encountered in sensor data. In DAC these discrepancies regulate both perceptual learning and the engagement of sequence memory. Consistent perceptual and behavioral errors lead to the second phase, namely *confusion*, the necessity to resolve the problem and learn through readapting. *Confusion* modulates learning as to facilitate the discovery and generation of new states to be assessed on their validity. In other words, to assist in performing abduction. Finally, *resolution* is the very process of stabilizing new knowledge that resolves the earlier encountered discrepancies and errors. This DAC-derived learning dynamics have been grounded in aspects of the physiology of the hippocampus [40] and pre-frontal cortex [32], and they reflect the core notions of Piaget's theory of cognitive development assimilation and accommodation through a process of equilibration [37, 56].

Human learners show a large variability in their performance and aptitude [16] requiring learning technologies to adjust to the skills and the progress of every individual. For learning to be efficient and applicable for as broad a range of students as possible, individual

differences need to be taken into account. The critical condition that has to be satisfied, however, is that the *confusion* needs to be controllable so that it adjusts to the skills and the progress of individual students. This is consistent with the classical notion of Vygotsky's Zone of Proximal Development which is the level of knowledge that the learner can acquire with external assistance of a teacher or a peer [54]. Individualization thus serves the identification of this epistemic and motivational level.

Monitoring, controlling and adjusting the phase of confusion is what we call *shaping the landscape of success*. This approach is consistent to the notion of scaffolding, a technique based on helping the student to cross Vygotsky's Zone of Proximal Development. The concept of controlled confusion, as well as of individualized training, has already been tested in the context of neurorehabilitation using DAC based Rehabilitation Gaming System (RGS) which assists stroke patients in their functional recovery of motor deficits [10, 11]. RGS indeed effectively adjusts to individual users in terms of difficulty, allowing for an unsupervised deployment of individualized rehabilitation protocols.

Within the DAC architecture, the processes of learning are not isolated within single layers but they result as the interplay among them and the external world [51]. Although both the processes of learning deployed in the current experiment (resistance, confusion, resolution) and the layers of the DAC architecture (Reactive, Adaptive, Contextual) constitute a specific order and initial dependencies, their relation is not fixed. Depending on the learning goal (learning a new concept, contextualizing new information within a broader scale, etc.) the tutoring may be focusing on one of the three layers. In order to systematically traverse the three phases of learning distinguished here, the user is guided through a goal-based learning.

By incorporating DAC within the educational framework, our aim is to be able to create the feeling of resistance and confusion to introduce new knowledge specific for every individual student. Adjusting to the skills and the progress of individual students may result in keeping the process of acquisition motivating; so it is essential that despite helping the student to overcome certain difficulties, the task remains challenging enough.

3 THE SYNTHETIC TUTOR ASSISTANT (STA)

The STA emerges as the interplay of the three layers of DAC architecture. It is the STA that provides individualized content, adapted to the needs and capabilities of each student. Here we layout the framework for the implementation of the STA within the DAC architecture. The Reactive Layer provides the basic interaction between the student, tutor and teaching material through a self-regulation system and an allostatic control mechanism. It encompasses the basic reaction mechanisms guiding the student through the learning material in a predefined reactive manner and is based on a self-regulation mechanism that contains predefined reflexes that support behavior. Such reflexes are triggered by stimuli that can be either internal (self) or external (environment) and are coupled to specific affective states of the agent.

The Adaptive Layer will adjust the learning scenario to the needs and capabilities of the student based on the user model that is online updated throughout the analysis of the interaction. To do so, the STA needs to assess the state of the student (cognitive, physical, emotional), learn from previous interactions and adapt to each student. This knowledge will support the rich and multimodal interactions based on a the DAC control architecture. Finally, the Contextual Layer will monitor and adjust the learning strategy over long periods of time and over all participating students through Bayesian memory and sequence optimization. In the pilot experiment reported here, we are assessing the properties of the Reactive Layer of the STA in an educational scenario.

3.1 Behavioral modulation

In case of the STA, the main purpose of the self-regulating mechanism of the Reactive Layer is to provide the tutor with an initial set of behaviors that will initiate and maintain the interaction between the STA and the student. Grounded in biology, where living organisms are endowed with internal drives that trigger, maintain and direct behavior [25, 38], we argue that agents that are endowed with a motivational system show greater adaptability compared to simple reactive ones [2]. Drives are part of a homeostatic mechanism that aims at maintaining stability [12, 44], and various autonomous systems have used self-regulation mechanisms based on homeostatic regimes [6, 3].

Inspired by Maslow's hierarchy of needs [33], Hull's drive reduction theory [25] and tested in the autonomous interactive space Ada [15], the robots behavior is affected by its internal drives (for example the need to socialize - establish and maintain interaction). Each drive is controlled by a homeostatic mechanism. This mechanism classifies the drive in three main categories: *under*, *over* and *within* homeostasis. The main goal of the STA is to maximize its effectivity (or "happiness") as a tutor assistant, by maintaining its drives within specific homeostatic levels. To do so, the STA will need to take the appropriate actions. These states are focusing on the level of interaction with the learner and its consistency. Coherence at the behavioral level is achieved through an extra layer of control that reduces drives through behavioral changes, namely the allostatic control. Allostasis aims at maintaining stability through change [34]. The main goal of allostasis is the regulation of fundamental needs to ensure survival by orchestrating multiple homeostatic processes that directly or indirectly help to maintain stability.

The allostatic controller adds a number of new properties of the STA-DAC architecture, ensuring the attainment of consistency and balance in the satisfaction of the agent's drives and foundations for utilitarian emotions that drive communicative cues [53]. This approach strongly contradicts the paradigm of state machines standardly employed in comparable approaches and, in general, within the robotics community. State machines provide a series of closed-loop behaviours where each state triggers another state in function of its outcome. Here, drives are not associated on a one-to-one basis with a specific behavior. Instead, each behavior is associated with an intrinsic effect on the drives and with the usage of the allostatic controller, drives, and therefore behavior, change as the environment changes. With such design, drives modulate the robot's behavior adaptively in the function of every learner and the learning environment in general. Although in our current implementation, the mappings are hard-coded as reflexes (Reactive Layer), according to the DAC architecture, the mappings should be learnt through experience to provide adaptation.

3.2 The setup (software and hardware)

The DAC architecture and framework proposed are mostly hardware independent, as it can be applied in various robotic implementations [19, 42, 53, 31]. Here, the implementation aims at controlling the behavior of the robot and it involves a large set of sensors and effectors, designed to study Human-Robot Interaction (HRI). The setup

(see figure 2) consists of the humanoid robot iCub (represented by the STA), the Reactable [23, 27] and a Kinect. The Reactable is a tabletop tangible display that was originally used as a musical instrument. It has a translucent top where objects and fingertips (cursors) are placed to control melody parameters. In our scenario, the usage of the Reactabe allows us to construct interactive games tailored to our needs. It furthermore provides information about the location of a virtual and physical object placed on the table and allows a precision that can hardly be matched using a vision based approach. In our lab, we have employed the Reactable in various interaction scenarios using the MTCF framework [28], such as musical DJ (cooperative game where the robot produces music with humans), Pong (competitive 2D simulated table tennis game) and Tic Tac Toe. The use and control of all these components allows the development of various interactive scenarios including educational games investigated here and allow the human and the robot to both act in a shared physical space. An extensive description of the overall system architecture can be found in [31, 52, 53]. The setup was designed to run autonomously in each trial, being the allostatic control the main component for providing the guidance for the learner/player during the task.

Figure 2. Experimental setup of the robot interacting with a human using the Reactable for the educational game scenario. In the image you can see the participant holding an object used to select an item from the Reactable (round table with projected images of countries and capitals). The participant is facing the iCub. The projected items are mirrored, so each side has the same objects.

4 TOWARDS ROBOTIC TEACHERS

In order to test the implementation of the STA-DAC as well as to evaluate the effectiveness of our scenario depending on different social features of the robot, we conducted a pilot study where the robot had the role of a tutor-peer.

The aim of the experiment focused on testing the effect of social cues (in this case, facial expression and eye contact) in HRI during an educational game. The goal was to test whether the variation of these social cues could affect the knowledge retrieval, subjective experience, and the very behavior towards the other player.

4.1 The educational scenario

The first question raised during the development of the STA is whether it can be an effective peer for the learner, both in terms of the social interactions and the impact on learning. Hence, the focus of this experiment is to study whether the modulation of certain behavioral parameters (based on the DAC architecture and the proposed behavioral modulation system), such as the use of eye contact and facial expressions, can change the acquisition of knowledge of a specific topic and the subjective experience of the user. On the one hand, eye contact can strengthen the interaction between the learner and the STA, for gazing can affect the knowledge transfer and the learning rate [36]. On the other hand, facial expressions can be used as a reinforcement of the participant's actions (the robot displays a happy face when the participant's choice is correct and a sad face when the matching is wrong), and could be considered as a reward.

The game-like scenario which we deployed is exercising Gagne's five learning categories [22]: verbal information, intellectual skill, cognitive strategy, motor skill and attitude. The game is based in a physical task, so the participants have to use their motor skills and, in order to solve the task, they have to develop a cognitive strategy to control their internal thinking processes. We also implemented three components of intellectual skill: concept learning, that is, learning about a topic; rule learning, used to learn the rules of the game; and, problem solving processes to decide how to match the pieces.

The educational scenario is a pairing game, where participants need to pair objects appearing on the Reactable to their corresponding categories. The pairing game is grounded in the premises of constructivism, where two or more peers learn together. Here the robot behaves similarly to a constructivist tutor: instead of just giving the information directly, it helps the student to understand the goal of the game (and, for example, reminding the subject the correct ways of playing) and it provides feedback regarding his actions (the robot only tells the correct answer to the subject when he has chosen a wrong answer, not before). For example, if the human selects a wrong pair, the robot indicates why the selection is wrong; it also comments on the correct selections. The players also receive visual information regarding their selection from the Reactable: if the selection is correct, the selected pair blinks with a green color and the object (but not the category) dissapears whereas the pair blinks with a red color if the selection is incorrect. The game was tested with both children and adults and the contents were adapted according their estimated knowledge. Therefore, for the children the game's topic was recycling, where the task was to correctly match different types of waste to the corresponding recycling bin. For the adults the topic was geography, where the task was to correctly match a capital with the corresponding country.

The learning scenario requires turn-taking and comprises three levels of increased difficulty. Both the human and robot had the same objects mirrored in each side. At each level, they had to correctly match the four objects to their corresponding category to proceed to the next level. The gradual increase of the difficulty allows for the scaffolding of the task, and consequently for the improvment of the learning process [4]. As mentioned earlier, the game was realized using the Reactable; the virtual objects were projected on the Reactable and object selection was achieved either with the usage of an object or with a cursor (fingertip). At the beginning of the interaction, the robot verbally introduces the game and is the first who initiates the interaction and the game.

4.2 Methods

We hypothesized that the combination of eye-contact and facial expressions strengthens the feedback between the player, the participant and the participant's choice, and affects the participant's subjective experience. As a result, we expected that when exposed to both behavioral conditions the participants would have a higher both knowledge transfer and the subjective experience.

To test our hypothesis and assess our architecture, we devised five experimental conditions (see Table 1) where we varied the gaze behavior and facial expressions of the STA. The experimental conditions are: Not-oriented Robot (NoR) (fixed gaze at a point - this way we are ensured that no eye contact is achieved); Task oriented Robot (ToR) (gaze supports actions, without making eye contact or showing facial expressions); Task and Human oriented Robot (T&HoR) (gaze supports actions, eye contact and showing facial expressions); Table-Human Interaction (THI), where the participant plays alone with the Reactable, and the Human-Human Interaction (HHI), where the participant plays with another human. Apart from the HHI, the behavior of the STA in terms of game play, verbal interaction and reaction to the participant's actions remained the same. The aim of the THI condition is to show the importance of embodiment of the STA during the interaction; the HHI condition acted as both the control group and a way of achieving a baseline regarding the interaction. The children were tested in the NoR, T&HoR and HHI conditions whereas the adults in all conditions.

Data were collected within three systems: knowledge and subjective experience questionnaires, behavioral data and the logs from the system. Participants had to answer pre- and post- knowledge questionnaires related to the pairing game. For recycling, the questionnaires had a total of twelve multiple-choice questions, including the same wastes and containers that the participants had to classify during the game. The information for creating this questionnaire came from the website "Residu on vas" (www.residuonvas.cat), property of the Catalan Wastes Agency. For geography, the questionnaires had a total of 24 multiple-choice questions (half of them, about the countries and capitals and the other half, about countries and flags). These questionnaires were given to the participants before and after the game, in order to evaluate their previous knowledge about the topic and later compare the pre- and post- knowledge results. The subjective experience questionnaire aims at assessing the STA's social behavior. It consists of 32 questions based on: the Basic Empathy Scale [26], the Godspeed questionnaires [5] and the Tripod Survey [17]. In the case of adults, there were 74 participants (age M = 25.18, SD = 7.55; 50 male and 24 female) distributed among five different conditions (THI=13, NoR=15, ToR=15, T&HoR=16, HHI=15). In the case of children, we tested 34 subjects (age M = 9.81, SD = 1.23; 23 male and 11 female) who randomly underwent three different experimental conditions (NoR=12, T&HoR=14, HHI=8).

Table 1. Table of the five experimental conditions.

	Embodiment	Action supporting gaze	Eye contact	Facial Expression
THI	No	No	No	No
NoR	Yes	No	No	No
ToR	Yes	Yes	No	No
T&HoR	Yes	Yes	Yes	Yes
HHI	Yes	Yes	Yes	Yes

Various conditions of robot behavior based on the interaction scenario

4.3 Results

First, we report a significant knowledge improvement in adults for all the conditions: THI, t(13) = 7. 697, p <0.001; NoR, (t(14) = 2.170, p = 0.048; ToR, t(14) = 3.112, p = 0.008, T&HoR, t(16) = 3.174, p = 0.006 and HHI,t(13) = 3.454 p = 0.004. In constrast, in children, there was no significance between conditions, although our results suggest a trend in improvement. We expected a difference among conditions, as we hypothesised that in the T&HoR condition, the knowledge transfer would be greater than the rest of the conditions. However this does not occur in neither the adult nor the children scenarios. In the case of children, we hypothesized that the associations were too simple; in the case of the adults, it seems that the knowledge transfer was achieved irregardless of the condition, suggesting that possibly the feedback of the Reactable itself regarding each pairing (green for correct and red for incorrect) might have been sufficient for the knowledge to be transferred.

Regarding the subjective experience, there was no statistical difference in the questionnaires data from children. We suspect that such result might be affected by the fact that both the Empathy and Godspeed questionnaires are designed for adults, and not children. In adults, although there was no significant difference among conditions for the Empathy and Tripod parts, there was a statistically significant difference between groups for the Godspeed part, as determined by one-way ANOVA (F(4,35) = 4.981, p = 0.003). As expected, humans scored higher (HHI, .06 $\pm$ 0.87), than the robot in two conditions (NoR, 2.84 $\pm$ 0.72, p = 0.003; ToR, 3.19 $\pm$ 0.46, p = 0.044, but surprisingly not in the T&HoR) and the table (THI,3.02 $\pm$ 0.56, p = 0.031) (Bonferroni post-hoc test). We can therefore hypothesize that the STA significantly scores lower than a human in all conditions but the one where its behavior is as close as possible to that of a human: gaze that sustains action (look at where the agent is about to point) and is used for communication purposes (look at human when speaking) and facial expressions as a feedback to the humans actions.

Regarding the behavioral data, there was a statistically significant difference between conditions for the mean gaze duration in children (one-way ANOVA (F(2,26) = 8.287, p = .0021)). A Bonferroni post-hoc test revealed that the time spent looking at the other player (in seconds) was significantly lower in the NoR (14.70 $\pm$ 8.81", p = 0.012) and the HHI conditions (11.74 $\pm$ 8.02", p = 0.003) compared to the T&HoR condition (30.97 $\pm$ 15.16")(figure 3). Our expectation regarding the difference between the NoR and T&HoR conditions was correctly met: people looked more at the agent who looked back at them. However, we were not expecting a difference between T&HoR and HHI condition. We believe that the reason why the difference in mean gaze duration occurs is because humans remained focused on the game and were mainly looking at table instead of looking at the other player. Furthermore, there were much less spoken interactions between them. In contrast, in the rest of the scenarios, the STA would comment on the actions of the participant, attrackting attention in more salient way.

In adults, a Kruskal-Wallis test showed that there was a high statistically significant difference in the time spent looking at the other player between the different conditions, x2(4) = 15.911, p = 0.003. The results of the Mann-Whitney Test showed significant differences between the THI (2.72 $\pm$ 5.53) and the NoR (16.37 $\pm$ 21.17) conditions (p = 0.026); the THI (2.72 $\pm$ 5.53) and the ToR (7.80 $\pm$ 7.76) conditions (p = 0.029); the THI (2.72 $\pm$ 5.53) and the T&HoR (19.87 $\pm$ 12.01) conditions (p <0.001); the ToR (7.80 $\pm$ 7.76) and the T&HoR (19.87 $\pm$ 12.01) conditions (p = 0.028); and the T&HoR

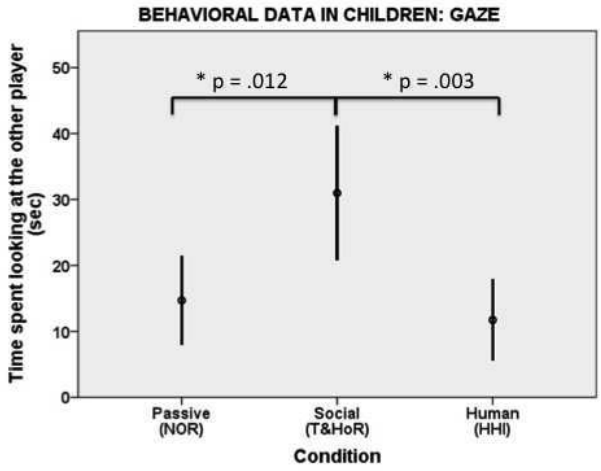


Figure 3. Time spent looking at the other player (in seconds) in children among conditions. Asterisks "*" depict significance.

(19.87 $\pm$ 12.01) and the HHI (3.66 $\pm$ 4.13) conditions (p = 0.002) (See figure 4). As expected, the more human-like the behavior of the STA, the more people would look at. The explanation regarding the difference between T&HoR and HHI in gaze duration is similar to that of children.

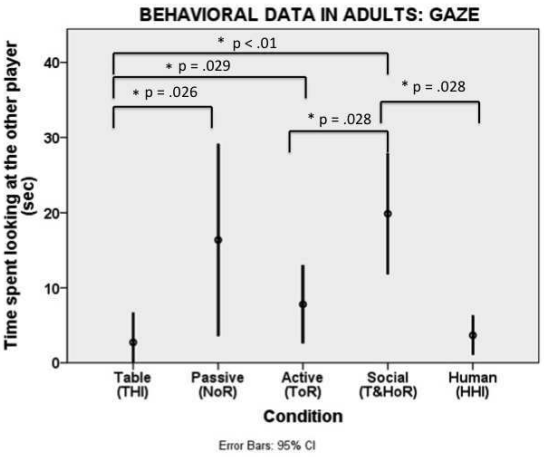


Figure 4. Time spent looking at the other player (in seconds) in adults among conditions. Asterisks "*" depict significance.

5 DISCUSSION AND CONCLUSIONS

The goal of the present study is to provide the key implementation features of the Synthetic Tutor Assistant (STA) based on the DAC architecture. Here, we propose the implementation of the STA within the DAC, a theory of the design principles which underlie perception, cognition and action. DAC is a layered architecture (Soma, Reactive, Adaptive and Contextual) intersected by three columns (world, self and actions), modeled to answer the H5W problem: Why, What, Where, When, Who and How. We explain the basic layers of DAC and focus on the Reactive Layer that constructs the basic reflexive behavioral system of the STA, as systhematically explained in section 3.1.

DAC predicts that learning is organized along a hierarchy of complexity and in order to acquire and consolidate new material the learner undergoes a sequence of learning phases: resistance, confusion and resolution. We argue that it is important to effectively adjust the difficulty of the learning scenario by manipulating the according parameters of the task (Adaptive Layer). This function will allow us for controlled manipulation of confusion, tailored to the needs of each student. Though it is not in the scope of the present study, in the future we plan to adjust the parameters of the learning scenario studied here on the basis of an online analysis of the learners' performance, interpreted both in terms of traditional pedagogical scales and the DAC architecture (Adaptive Layer). The learner's errors and achievements will be distinguished in terms of specific hierarchical organization and dynamics. Finally, the Contextual Layer will monitor and adjust the difficulty parameters for both individual students and bigger groups on a longer time scales. The motivational system presented is mainly focused on the Reactive Layer of the architecture, but our aim is to primarily adapt the Reactive Layer to the needs of STA and teaching scenarios and then extend the STA to include the Adaptive and Contextual Layers.

We devised an educational scenario to test the implementation of the STA-DAC as well as to evaluate the effectiveness of different social features of the robot (social cues such as eye contact and facial expressions). The task devised was a pairing game using the Reactable as an interface, where the robot acts as a constructivist tutor. The pairing consisted of matching different types of waste to the corresponding recycling bin (recycle game) for the children and matching the corresponding capital to a country (geography game) for the adults. The learning scenario was turn-taking with three levels of increased difficulty. The experiment consists of five different conditions, described in section 4.2: THI, NoR, ToR, T&HoR and HHI. Adults were tested in all conditions whereas children in NoR, T&HoR and HHI. To assess the interaction, the implementation as well as the effectiveness of the robot's social cues, behavioral data, logged files and questionnaires were collected.

In the results, we see that in adults, there are significant differences in knowledge improvement among conditions. On the other hand, there is a trend in knowledge improvement in children, but it is not significant. The results are not sufficient to draw any concrete conclusions about knowledge retrieval. Nevertheless, we can see that people scored higher in the post-experiment questionnaire, on the other hand, results are not enough to identify exactly the reason. It is possible that the task, though the difficulty increased on each trial, would still remain relatively easy. That is why we aim at devising a related experiment where we would exploit the Adaptive Layer that adapts the difficulty to each individual player.

Our results show that children looked more at the T&HoR robot than then ToR or HHI. Based on these results, we can conclude that the behavior of the Task and Human oriented Robot drew more the attention of the participant than the other human or the solely Task oriented Robot. The robot was looking at the participant when it was addressing him; its gaze followed both the player's and its own actions, meaning that it would look at the object that the participant had chosen or the object that it chose. Finally, it would show facial expressions according to each event: happy for the correct pair or sad for the incorrect one. Such cues may indeed be more salient and draw the attention of the player. In all conditions, the robot was speaking, so it seems that it was the implicit non-verbal communicative signals

of the robot that drew the attention of the participant. In the case of the adults, the results are also similar. Such behavior is important in the development of not only social but also educational robots, as gaze following directs attention to areas of high information value and accelerates social, causal, and cultural learning [1]. Indeed, such cues positively impact human-robot task performance with respect to understandability [7]. This is supported by results like the ones of [24], where the addition of gestures led to a higher effect on the participant only when the robot was also performing eye contact.

Finally, the results from the Godspeed questionnaire in adults show a significant difference in the overall score between HHI and THI, NoR, ToR but not the T&HoR. Such results were generally expected, as a human would score higher than the machine. In children, there was no significance in any of the conditions, however, it may be the case that the Godspeed questionnaire is not the optimal measurement for subjective experience, at it may contain concepts that are not yet fully understood by such a young age. Perhaps simpler or even more visual (with drawings that represent the extremes of a category) questionnaires would be more appropriate.

Though the knowledge transfer results are not sufficient to draw any concrete conclusions (as the knowledge transfer is not significantly different among conditions), the complex social behavior of the robot indeed attracts attention of the participant. As for the pilot study, the authors need to focus more on the evaluation of the system, and need to introduce a strong experimental design to derive more specific conclusions. Further analysis of the behavioral data can provide insight regarding eye contact in terms of error trials, decision time and task difficulty. In the upcoming experiments we will provide a better control in the HHI condition. A possible strategy is to deploy a specific person (an actor) as the other player, to normalize the characteristics of the scenario between all the subjects.

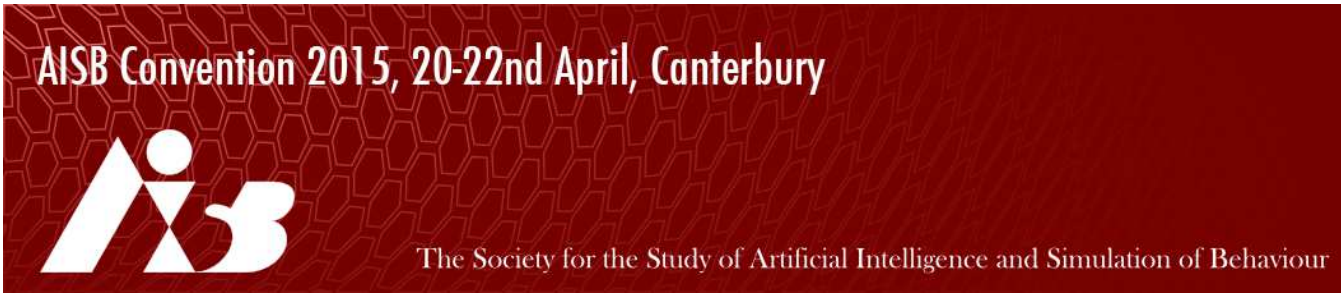
ACKNOWLEDGEMENTS

This work is supported by the EU FP7 project WYSIWYD (FP7-ICT-612139) and EASEL (FP7-ICT- 611971).

REFERENCES

[1] '"Social" robots are psychological agents for infants: a test of gaze following.', *Neural networks : the official journal of the International Neural Network Society*, **23**(8-9), 966–72, (2010).
[2] Michael A Arbib and Jean-Marc Fellous, 'Emotions: from brain to robot', *Trends in cognitive sciences*, **8**(12), 554–561, (2004).
[3] Ronald C Arkin, Masahiro Fujita, Tsuyoshi Takagi, and Rika Hasegawa, 'An ethological and emotional basis for human–robot interaction', *Robotics and Autonomous Systems*, **42**(3), 191–201, (2003).
[4] Roger Azevedo and Allyson F Hadwin, 'Scaffolding self-regulated learning and metacognition–implications for the design of computer-based scaffolds', *Instructional Science*, **33**(5), 367–379, (2005).
[5] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi, 'Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots', *International journal of social robotics*, **1**(1), 71–81, (2009).
[6] Cynthia Breazeal, 'Emotion and sociable humanoid robots', *International Journal of Human-Computer Studies*, **59**(1), 119–155, (2003).
[7] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin, 'Effects of nonverbal communication on efficiency and robustness in human-robot teamwork', in *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pp. 708–713. IEEE, (2005).
[8] Cynthia L Breazeal, *Designing sociable robots*, MIT press, 2004.
[9] Sylvain Calinon, Danilo Bruno, Milad S Malekzadeh, Thrishantha Nanayakkara, and Darwin G Caldwell, 'Human–robot skills transfer interfaces for a flexible surgical robot', *Computer methods and programs in biomedicine*, (2014).
[10] Mónica S Cameirão, Sergi B Badia, Esther Duarte Oller, PFMJ Verschure, et al., 'Neurorehabilitation using the virtual reality based rehabilitation gaming system: methodology, design, psychometrics, usability and validation', *Journal of neuroengineering and rehabilitation*, **7**(1), 48, (2010).
[11] Mónica S Cameirão, Sergi Bermúdez i Badia, Esther Duarte, Antonio Frisoli, and Paul FMJ Verschure, 'The combined impact of virtual reality neurorehabilitation and its interfaces on upper extremity functional recovery in patients with chronic stroke', *Stroke*, **43**(10), 2720–2728, (2012).
[12] Walter Bradford Cannon, 'The wisdom of the body.', (1932).
[13] Paolo Dario, Paul FMJ Verschure, Tony Prescott, Gordon Cheng, Giulio Sandini, Roberto Cingolani, Rüdiger Dillmann, Dario Floreano, Christophe Leroux, Sheila MacNeil, et al., 'Robot companions for citizens', *Procedia Computer Science*, **7**, 47–51, (2011).
[14] Armin Duff and Paul FMJ Verschure, 'Unifying perceptual and behavioral learning with a correlative subspace learning rule', *Neurocomputing*, **73**(10), 1818–1830, (2010).
[15] Kynan Eng, David Klein, A Babler, Ulysses Bernardet, Mark Blanchard, Marcio Costa, Tobi Delbrück, Rodney J Douglas, Klaus Hepp, Jonatas Manzolli, et al., 'Design for a brain revisited: the neuromorphic design and functionality of the interactive space 'Ada", *Reviews in the Neurosciences*, **14**(1-2), 145–180, (2003).
[16] RM Felder and R Brent, 'Understanding student differences', *Journal of engineering education*, **94**(1), 57–72, (2005).
[17] R Ferguson, 'The tripod project framework', *The Tripod Project*, (2008).
[18] Ylva Fernaeus, Maria Håkansson, Mattias Jacobsson, and Sara Ljungblad, 'How do you play with a robotic toy animal?: a long-term study of pleo', in *Proceedings of the 9th international Conference on interaction Design and Children*, pp. 39–48. ACM, (2010).
[19] Marti Sanchez Fibla, Ulysses Bernardet, and Paul FMJ Verschure, 'Allostatic control for robot behaviour regulation: An extension to path planning', in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pp. 1935–1942. IEEE, (2010).
[20] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn, 'A survey of socially interactive robots', *Robotics and autonomous systems*, **42**(3), 143–166, (2003).
[21] Marina Fridin, 'Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education', *Computers & Education*, **70**, 53–64, (2014).
[22] Robert Mills Gagné, *The conditions of learning and theory of instruction*, Holt, Rinehart and Winston New York, 1985.
[23] Günter Geiger, Nadine Alber, Sergi Jordà, and Marcos Alonso, 'The reactable: A collaborative musical instrument for playing and understanding music', *Her&Mus. Heritage & Museography*, (4), 36–43, (2010).
[24] J. Ham, B. René, R. Cuijpers, D. van der Pol, and J. J. Cabibihan, 'Making Robots Persuasive: The Influence of Combining Persuasive Strategies (Gazing and Gestures) by a Storytelling Robot on Its Persuasive Power', in *Third International Conference, ICSR 2011 Amsterdam, The Netherlands, November 24-25, 2011 Proceedings*, pp. 71–83, (2011).
[25] Clark Hull, 'Principles of behavior', (1943).
[26] Darrick Jolliffe and David P Farrington, 'Development and validation of the basic empathy scale', *Journal of adolescence*, **29**(4), 589–611, (2006).
[27] Sergi Jordà, 'On stage: the reactable and other musical tangibles go real', *International Journal of Arts and Technology*, **1**(3), 268–287, (2008).
[28] Carles F Julià, Daniel Gallardo, and Sergi Jordà, 'Mtcf: A framework for designing and coding musical tabletop applications directly in pure data', in *Proceedings of the International Conference on New Interfaces for Musical Expression*, volume 20011, (2011).
[29] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro, 'Interactive robots as social partners and peer tutors for children: A field trial', *Human-computer interaction*, **19**(1), 61–84, (2004).
[30] Cory D Kidd and Cynthia Breazeal, 'Robots at home: Understanding long-term human-robot interaction', in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pp. 3230–3235. IEEE, (2008).
[31] Stéphane Lallée, Vasiliki Vouloutsi, Maria Blancas Munoz, Klaudia Grechuta, Jordi-Ysard Puigbo Llobet, Marina Sarda, and Paul FMJ Verschure, 'Towards the synthetic self: making others perceive me as an other'.
[32] Encarni Marcos, Pierpaolo Pani, Emiliano Brunamonti, Gustavo Deco,

Stefano Ferraina, and Paul Verschure, 'Neural variability in premotor cortex is modulated by trial history and predicts behavioral performance', *Neuron*, **78**(2), 249–255, (2013).
[33] Abraham Harold Maslow, 'A theory of human motivation.', *Psychological review*, **50**(4), 370, (1943).
[34] Bruce S McEwen, 'Allostasis and allostatic load: implications for neuropsychopharmacology', *Neuropsychopharmacology*, **22**(2), 108–124, (2000).
[35] Seymour Papert, *Mindstorms: Children, Computers, and Powerful Ideas*, Basic Books, Inc., New York, NY, USA, 1980.
[36] Fiona G Phelps, Gwyneth Doherty-Sneddon, and Hannah Warnock, 'Helping children think: Gaze aversion and teaching', *British Journal of Developmental Psychology*, **24**(3), 577–588, (2006).
[37] Jean Piaget and Margaret Trans Cook, 'The origins of intelligence in children.', (1952).
[38] Robert Plutchik, *The emotions*, University Press of America, 1991.
[39] Tony J Prescott, Nathan Lepora, and Paul FMJ Vershure, 'A future of living machines?: International trends and prospects in biomimetic and biohybrid systems', in *SPIE Smart Structures and Materials+ Nondestructive Evaluation and Health Monitoring*, pp. 905502–905502. International Society for Optics and Photonics, (2014).
[40] César Rennó-Costa, John E Lisman, and Paul FMJ Verschure, 'A signature of attractor dynamics in the ca3 region of the hippocampus', *PLoS computational biology*, **10**(5), e1003641, (2014).
[41] Martin Saerbeck, Tom Schut, Christoph Bartneck, and Maddy D Janse, 'Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1613–1622. ACM, (2010).
[42] Marti Sanchez-Fibla, Ulysses Bernardet, Erez Wasserman, Tatiana Pelc, Matti Mintz, Jadin C Jackson, Carien Lansink, Cyriel Pennartz, and Paul FMJ Verschure, 'Allostatic control for robot behavior regulation: a comparative rodent-robot study', *Advances in Complex Systems*, **13**(03), 377–403, (2010).
[43] Brian Scassellati, 'Using social robots to study abnormal social development', (2005).
[44] John P Seward, 'Drive, incentive, and reinforcement.', *Psychological review*, **63**(3), 195, (1956).
[45] Gerry Stahl, 'What we know about cscl and implementing it in higher education', chapter Building Collaborative Knowing: Elements of a Social Theory of CSCL, 53–85, Kluwer Academic Publishers, Norwell, MA, USA, (2004).
[46] PAUL VERSCHURE, 'Formal minds and biological brains ii: from the mirage of intelligence to a science and engineering of consciousness', (2013).
[47] Paul FMJ Verschure, 'Distributed adaptive control: A theory of the mind, brain, body nexus', *Biologically Inspired Cognitive Architectures*, (2012).
[48] Paul FMJ Verschure and Philipp Althaus, 'A real-world rational agent: Unifying old and new ai', *Cognitive science*, **27**(4), 561–590, (2003).
[49] Paul FMJ Verschure, Cyriel MA Pennartz, and Giovanni Pezzulo, 'The why, what, where, when and how of goal-directed choice: neuronal and computational principles', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **369**(1655), 20130483, (2014).
[50] Paul FMJ Verschure, Cyriel MA Pennartz, and Giovanni Pezzulo, 'The why, what, where, when and how of goal-directed choice: neuronal and computational principles', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **369**(1655), 20130483, (2014).
[51] Paul FMJ Verschure, Thomas Voegtlin, and Rodney J Douglas, 'Environmentally mediated synergy between perception and behaviour in mobile robots', *Nature*, **425**(6958), 620–624, (2003).
[52] Vasiliki Vouloutsi, Klaudia Grechuta, Stéphane Lallée, and Paul FMJ Verschure, 'The influence of behavioral complexity on robot perception', in *Biomimetic and Biohybrid Systems*, 332–343, Springer, (2014).
[53] Vasiliki Vouloutsi, Stéphane Lallée, and Paul FMJ Verschure, 'Modulating behaviors using allostatic control', in *Biomimetic and Biohybrid Systems*, 287–298, Springer, (2013).
[54] Lev S Vygotsky, *Mind in society: The development of higher psychological processes*, Harvard university press, 1980.
[55] Kazuyoshi Wada and Takanori Shibata, 'Living with seal robots in a care house-evaluations of social and physiological influences', in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pp. 4940–4945. IEEE, (2006).
[56] Barry J Wadsworth, *Piaget's theory of cognitive and affective development: Foundations of constructivism* ., Longman Publishing, 1996.



AISB Convention 2015

University of Kent, Canterbury

Proceedings of the AISB 2015 Symposium From Mental "Illness" to Disorder and Diversity: New Directions in the Philosophical and Scientific Understanding of Mental Disorder

Edited by Joel Parthemore and Blay Whitby

ISBN: 978-1-908187-52-9

Introduction to the Convention

The AISB Convention 2015—the latest in a series of events that have been happening since 1964—was held at the University of Kent, Canterbury, UK in April 2015. Over 120 delegates attended and enjoyed three days of interesting talks and discussions covering a wide range of topics across artificial intelligence and the simulation of behaviour. This proceedings volume contains the papers from the symposium entitled *From Mental "Illness" to Disorder and Diversity: New Directions in the Philosophical and Scientific Understanding of Mental Disorder*, one of eight symposia held as part of the conference. Many thanks to the convention organisers, the AISB committee, convention delegates, and the many Kent staff and students whose hard work went into making this event a success.

—Colin Johnson, Convention Chair

Contents

Psychiatry and the poverty of subjectivity:

How phenomenology can contribute to the validation of categories of disorder

Anthony Vincent Fernandez[1]

Abstract. Psychiatry, and especially psychiatric classification, finds itself in a state of crisis. Recent criticisms have been leveled by patient advocacy groups, psychotherapists, and even psychiatrists (including the chairs of both the DSM-III and DSM-IV taskforces). Most notably, the National Institute of Mental Health (NIMH) announced—just weeks prior to the 2013 publication of the DSM-5—that it will primarily fund studies that *do not use* the DSM-5 categories of disorder. In light of the problems of classification plaguing the field of psychiatry, a number of phenomenologists (including Aho, Parnas, Ratcliffe, Sass, Stanghellini, and Zahavi) have argued that contemporary phenomenological research into psychopathology should be used to guide the project of reclassification. While I agree with this claim, I argue that these phenomenologists have failed to delineate among a number of domains of phenomenological research. And, in failing to make such distinctions, are unable to distinguish between those areas of research that can be used to validate categories of disorder, and those that cannot.

In order to remedy this issue in contemporary phenomenological psychopathology, I here propose three domains of phenomenological research—1) existential structures, 2) modes, and 3) traditions. The first is understood as the domain of phenomenology proper, and consists of the categorial characteristics of human existence (e.g. intersubjectivity, embodiment, situatedness, etc.). The second is understood as the study of the various modes of these categorial characteristics (the modes of Situatedness, for example, include anxiety, boredom, joy, etc.). The third is understood as the domain of hermeneutics proper, but is often included in phenomenological studies. It consists of the framework of meaning that sediments throughout cultural and biographical developments, shaping what we see things *as* (e.g. people from different religious backgrounds will experience different objects *as* sacred, without actively interpreting the meaning of these objects).

1 INTRODUCTION

Since the 1980s, psychiatric classification has been dominated by the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM). However, the DSM-5, released in May of 2013, was the target of searing criticism from patient advocacy groups, psychotherapists, and even psychiatrists (including Robert Spitzer [1], chair of the DSM-III taskforce, and Allen Frances [2], chair of the DSM-IV taskforce). However, the criticism with the greatest visibility and most significant ramifications came from the National Institute of Mental Health (NIMH). Just weeks prior to the publication of the DSM-5, Tom Insel, head of the NIMH, declared in a public announcement that NIMH funding will be largely reserved for studies that *do not use* the DSM-5 categories of mental disorders [3]. Instead, most funding will be awarded for studies that support the new Research Domain Criteria (RDoC) project in its

[1]Dept. of Philosophy, University of South Florida. Email: avf@mail.usf.edu.

attempt to develop scientifically (i.e. neurologically and behaviorally) validated categories of disorder.[2]

The major concern held by Insel is that psychiatric research has failed to correlate the diagnostic categories of the DSM with neurobiological mechanisms. In other words, the symptomatically delineated categories of the DSM, drawing primarily on references to patients' lived experience (e.g. cognitive distortions, emotional disturbances, delusions, or hallucinations) and expressions of subjective experience in behavior (e.g. insomnia/hypersomnia, tearfulness, or hyperactivity), have not been adequately correlated with relevant changes in the brain. In order to remedy this issue, the RDoC project seeks to delineate preliminary research categories of disorder using only third-person observable data—including neurobiological data and certain kinds of behavioral data [4][5].[3] As currently formulated, studies of the lived world of subjects with psychiatric disorders will play no role in the delineation of the preliminary research categories that will be drawn up by the RDoC project.

While I share Insel's concerns over the disutility of the DSM categories, especially in regard to their failure to map onto neurobiological mechanisms, I believe that the RDoC and other projects aimed as reclassifying psychiatric disorders have been too quick to dispense with a phenomenological orientation. I argue not only that references to lived experience are conducive to the preliminary delineation of abnormal phenomena for neurobiological research, but also that phenomenological psychopathology (with its roots in the tradition of 20th century continental philosophy) is an invaluable tool for obtaining just such data.

What this amounts to is an argument over which kinds of research can contribute towards the project of creating valid categories of disorder. Philosophers and psychiatrists such as Robins and Guze [6], and Jablensky and Kendell [7] have outlined at least four kinds of validity, including construct, content, concurrent, and predictive.[4] In following with Jablensky and Kendell's breakdown of the various kinds of validity, a category of disorder has construct validity when it "is based on a coherent, explicit set of defining features"; it has content validity when it "has empirical referents, such as verifiable observations for establishing its presence"; it has concurrent validity when it "can be corroborated by independent procedures such as biological or psychological tests"; and it has predictive validity when it "predicts future course of illness or treatment response" [7].

I argue that phenomenology can contribute directly to content validity by clearly describing the form of subjectivity and the lived world of a person with the disorder in question, and it can contribute directly to construct validity by differentiating one form of pathological subjectivity from another by clearly distinguishing essential from non-essential features of disorder. By offering rich descriptions of the disorders in question and by drawing clear boundaries around these disorders (at least in the cases where such boundaries exist), phenomenology can indirectly contribute towards the other forms of validity by supplying preliminary, symptomatically homogeneous categories that are more likely to correlate with specific psychological and neurobiological tests, as well as predict treatment response and course of illness.

My argument in this paper is presented in five parts. First, I review the work of the psychiatrist Gordon Parker and his colleagues in order to illustrate how close attention to subjective dimensions of disorder can lead to better systems of classification. Second, I review the recent literature on the role of phenomenology in psychiatric classification, focusing especially on the work of Josef Parnas and Dan Zahavi. In so doing, I bring to light some of the inadequacies in these accounts, showing that they fail to distinguish among a number of domains of phenomenological research, and thus among an array of different *kinds* of changes in subjectivity and human existence. Third, I draw on both historical and contemporary work in phenomenology and hermeneutics in order to delineate the three domains of phenomenological research. Fourth, I revisit each of these domains in light of the particular aims of phenomenological psychopathology, illustrating the kinds of pathological shifts that might be investigated in each domain. Fifth, and finally, I offer a preliminary sketch of how attention to these distinctions can lead to new psychiatric classifications with greater validity.

2 PSYCHIATRIC CLASSIFICATION AND THE POVERTY OF SUBJECTIVITY

In addition to the general criticisms leveled against psychiatry and the DSM, major depressive disorder (MDD) has found itself in the public spotlight following the publication of a number of popular books criticizing issues of classification, diagnosis, and treatment. Topics such as the pathologizing of normal kinds of sadness [8], the extremely low efficacy of anti-depressants [9], and the rapid rise in the number of people who meet the criteria for a diagnosis of MDD [10] have entered into public discourse, adding to the already marred reputation of contemporary psychiatry.

One researcher who has taken such criticisms to heart is the Australian psychiatrist, Gordon Parker. For over a decade, Parker has been pushing against what he calls the unitarian model of depression, which posits that depression is a single category of disorder that may differ along some dimensions (but in most cases is only considered to have one dimension—severity). His dissatisfaction with this model of depression led him to an article written by Kendell [11] that reviewed the historical ways of classifying depressive disorders. Drawing from these historical categories as well as his own research, Parker proposed three categories of depressive disorders (with the third category being

[2] It should be noted that the RDoC is not itself a system of classification. As Cuthbert and Kozack state, "It might better be termed 'an experiment toward classification.'"

[3] The place of behavior in this debate is a complex one, and I cannot say much about it here. Both the DSM and the new RDoC project rely heavily on observations of behavior. One important difference is that in the DSM-III and later editions, behaviors that show up exclusively—or at least primarily—in a single category of disorder are prioritized. In the RDoC, pathological or abnormal behaviors that show up across the boundaries of disorders drawn in the DSM are prioritized, primarily for the purpose of narrowing down avenues for further research on the neurobiological mechanisms behind such behaviors (rather than mechanisms behind certain categories of disorder, since it is these categories that the RDoC has put into question).

[4] It might be better to state that each of these aspects—rather than being independent kinds of validity—can be used to enhance or increase the validity of a category of disorder. However, this still leaves open the question of what validity itself *is*.

a catch-all for a diversity of depressive disorders that require further delineation) [12–14].

The classification he developed is hierarchical, with each subsequent level of the disorder incorporating the features of the level below it while including at least one additional feature. The three categories are, from the top to the bottom of the hierarchy, psychotic depression, melancholic depression, and non-melancholic depression. Non-melancholic depression is characterized simply by depressed mood (which is an admittedly ambiguous and likely heterogeneous symptom reference).[5] Melancholic depression is, in turn, characterized by "observable (and not merely reported) psychomotor disturbance" [14]. This characteristic, not being found in non-melancholic forms of depression, is an essential feature and a clear marker of melancholic depression. Psychotic depression, being the final category, includes depressed mood and psychomotor disturbance, as well as psychotic features, such as delusions or hallucinations.

Further research on the treatment efficacy and the neurobiological substrates related to these categories supplied evidence for their having a higher degree of validity than the DSM category of MDD [16]. For example, Parker and colleagues show that two-thirds of subjects who meet their criteria for melancholic and non-melancholic depression improve with anti-depressant drugs alone, while only one-quarter to one-third of subjects who meet their criteria for psychotic depression improve with the same treatment. Further data shows that anti-depressants have markedly higher efficacy for people with melancholic depression than for people with non-melancholic depression. Also, the addition of neuroleptics to anti-depressant treatment in the case of psychotic depression shows a marked increase in efficacy (beyond the rates for the same treatments when given to those with melancholic and non-melancholic depressions). Finally, psychotherapy proved beneficial only in non-melancholic forms of depression, having little primary effect on subjects with melancholic and psychotic depressions.

These findings, along with preliminary data pointing to distinct neurobiological substrates related to each category of depression, offer considerable evidence for the validity of Parker's hierarchical classification (at least when compared with the DSM category of MDD). However, what is most intriguing about this system of classification (at least for the purposes of this paper), is that its divisions and categorizations were originally made without reference to neurobiological causes, instead drawing primarily on subjective and experiential phenomena, such as depressed mood, delusions, and hallucinations.[6]

In spite of the success of Parker's categorization, it must still be kept in mind that these distinctions were drawn using a fairly superficial account of human subjectivity. This is not to say that the categories or divisions are illegitimate. Rather, I argue that such methods of categorization and classification can be markedly enhanced by traditions that have richer and more robust accounts of human subjectivity at their disposal.

3 CONTEMPORARY PHENOMENOLOGY AND THE RECLASSIFICATION OF DISORDERS

This basic line of argument has been offered by a number of contemporary phenomenologists and phenomenological psychopathologists [15,17–20]. While each of these authors has approached the possibility of using phenomenological research to inform psychiatric classification, I here focus primarily on a paper by Parnas and Zahavi entitled, "The Role of Phenomenology in Psychiatric Diagnosis and Classification," as it deals with the issue most directly.

In this work, Parnas and Zahavi aim to show how the tools and frameworks developed by the classical phenomenologists—including Husserl, Heidegger, and Merleau-Ponty—can help psychiatric researchers focus in on previously ignored (but often central) features of disordered subjectivity. They even go so far as to claim that "a search for a faithful *description of experience* must be considered as a necessary first step in any taxonomic effort, including attempts of reducing abnormal experience to its potential biological substrate" (2002, 137). They trace this idea back to Jaspers, who stressed the need for careful attention to experience, whether this is achieved by 1) observing "gestures, behavior, [and] expressive movements" in an attempt to perceive the meaning of such bodily engagements, 2) directly questioning or interviewing the subject, or 3) considering written first-person reports by the subject herself [21,22].

According to Parnas and Zahavi, phenomenology's major contribution towards the elucidation of psychiatric disorders stems from its account of the "essential structures" of subjectivity that were originally delineated by the classical phenomenologists. While there are numerous essential structures that might be discussed, they focus in particular on phenomenal consciousness and self-awareness; temporality; intentionality; embodiment; and intersubjectivity. These make up some of the core dimensions of phenomenological research, and the authors clearly illustrate how phenomenological research on each of these essential structures might contribute towards the project of re-classifying mental disorders.

However, because of the plethora of recent research in phenomenological psychopathology and the ensuing divergence of phenomenological frameworks and emphases among those working in the discipline, there is a different sort of clarification that is in sore need of attention. This is what I refer to as the *layers of phenomenological research*. The delineation of these layers does not amount to an alternative way of distinguishing among the essential features of subjectivity discussed by figures such as Parnas and Zahavi. Rather, all of these essential features are encompassed by just the first of three layers of phenomenological research.[7]

[5] See Stanghellini [15] for a phenomenological critique of the symptom of "depressed mood."

[6] Some kinds of psychomotor disturbance also fall into the category of experiential or subjective, but in this particular case Parker includes the qualification that it must be observable by someone besides the subject herself. As a result, this particular symptom does not technically count as experiential or subjective. Nonetheless, it does point, or refer, to an experiential phenomenon.

[7] The layers I sketch here were originally articulated in a paper with Giovanni Stanghellini. However, I here use slightly different terminology and draw the divisions in a slightly different manner. This is done in part because the original paper was written with a focus on Jaspers, while this paper focuses more directly on the philosophical tradition of 20th century transcendental phenomenology. However, it is also the case that I have realized that the characterizations of some of the layers in the earlier

4 LAYERS OF PHENOMENOLOGICAL RESEARCH

I refer to the three layers of phenomenological research as 1) existentials[8], 2) modes, and 3) tradition. These layers are related to each other in a particular manner, which is to say, they are not *merely* distinct. They are in some sense hierarchical, with the subject matter of each domain being a condition for the subject matter of the following domains (e.g. modes are modes *of* existentials). However, they can also be related in terms of degrees of particularity. The existentials that are discussed in the phenomenological canon are typically understood as applying to any and all human subjectivities, thereby being universal. Modes, on the other hand, tend to be available to all subjects, but at any time a subject finds herself only in particular modes. The term "tradition," on the other hand, may refer more directly to the subject matter of hermeneutics, but is also taken up in phenomenological studies of how culture and even personal narratives shape the way a world shows up to us. In this sense, existentials are the most universal, while tradition is the most particular.

4.1 Existentials

Existentials (sometimes referred to as "existential structures," "essential structures," or just "structures") comprise the first layer of phenomenological research, and are typically considered to be the subject matter of phenomenology proper. Phenomenology, with its Husserlian goal of discovering the *eidos*, or essence of the phenomenon under investigation, seeks out the necessary, universal, and invariant characteristics of human consciousness and existence. It is these characteristics that we call "existentials."[9]

Another important, but oft ignored characteristic of existential structures is that they are categorial. That is to say, existential structures are categories of characteristics of human existence. To take an example from Heidegger's *Being and Time*, the existential that he calls *Befindlichkeit*, typically translated as "situatedness," "affectedness" or even "sofindingness," refers to the fact that human beings always already find themselves situated in and attuned to the world. However, there are a variety of ways one can be situated in and attuned to the world. Situatedness, then, refers not to my particular way of being situated and attuned, but to the category that encompasses all the possible ways of being situated, such as through fear, anxiety, wonder, or boredom. It is this categorial characteristic that is considered an existential.

4.2 Modes

Modes make up the second layer of phenomenological investigation, but they are not, strictly speaking, the subject matter of phenomenology proper. This is because modes are, by their very nature, contingent and variable. They do not make up essential, categorial characteristics of human existence. To continue the example above, I can be attuned and situated by fear, anxiety, wonder, or boredom. But the very fact that I can be attuned through a variety of moods means that no particular mood is part of my essential, existential structure.[10]

There are at least two ways modes can be approached in phenomenological research. First, they can be approached for their own sake, which is to say, a particular mode can be studied with the express purpose of learning more about that mode. An example of this kind of study is found in Heidegger's lecture on boredom, in which he conducts a lengthy phenomenological investigation of this mood for the express purpose of understanding the ways we can be bored, and the ways boredom shapes the meaning and significance of our world. Second, modes can be investigated for the sake of discovering characteristics shared by all modes included in a particular category. For example, in this same lecture course, Heidegger distinguishes among three different kinds of boredom based on whether they are directed towards an object, a situation, or disclose the world as a whole. While these distinctions were derived from a study of boredom, they proved useful in understanding moods in general, and in this sense his investigations were able to shed light on the existential structure of situatedness as a whole [23].

4.3 Tradition

Along with existentials and modes, phenomenological research often involves the study of what may be termed "tradition." This term is used throughout the phenomenological canon, receiving considerable treatment in Heidegger's early lecture courses, as well as in *Being and Time*. It plays an important role in genetic and generative phenomenology more generally, especially in Husserl's later works, such as *The Crisis of European Sciences and Transcendental Phenomenology* and "The Origin of Geometry" [24]. The term is typically understood in a broad sense, referring to one's "totality of presuppositions." There is a range of terms that are related to, or sometimes used as synonyms for, tradition. Some of these are facticity, thrownness, hermeneutical Situation, history, culture, and prejudice.

In addition to the myriad ways of referring to tradition, there are at least two reasons it is made the object of phenomenological research. The first, which we perhaps see most often in Heidegger's early works (but also in the works of

paper were not adequate to the task at hand, and needed to be updated and revised. The divisions and definitions of these layers are still, to some degree, a work in progress.

[8] Existentials are typically understood as the subject matter of phenomenology proper. In some cases they are referred to as structures, rather than existentials, but the term "structure" [*Struktur*] is used in a variety of ways, both within and amongst the works of each phenomenologist. In light of the possibilities for confusion that are opened up by the sometimes loose definitions of "structure," I have decided to use the narrower term, "existential."

[9] Heidegger often speaks of "ex-sistence" as a standing outside of, transcending or, simply, openness. Understood in this way, we can take "existentials" as categorial characteristics of human existence that play a role in the openness of the lived world, or the way in which the lived world is opened up and articulated for us.

[10] Besides moods, there are a number of other modes that have been discussed in the phenomenological literature. However, most of the classical phenomenologists fail to offer clear and careful definitions of existential structures and modes, so I rely on *Befindlichkeit* (situatedness) and *Stimmung* (mood) here because they offer the clearest distinction between existentials and modes in Heidegger's texts.

Husserl and Merleau-Ponty), is the explicit interrogation of our (mostly) tacit presuppositions that shape our interpretation of whatever the phenomenologist is interested in investigating. In *Being and Time*, for example, Heidegger engages in explicit interrogations of our presuppositions with respect to our concepts of time, truth, and being. In order to approach these concepts as phenomena—which is to say, as the proper subject matter of phenomenological research—we need to first make explicit the presuppositions that are at play in determining our everyday, scientific, or even philosophical conceptions of these phenomena. In the absence of such an interrogation—taken as a pre-phenomenological investigation, or an investigation conducted for the purpose of preparing for a phenomenological investigation proper—the phenomenologist risks (or perhaps risks more severely) falling back into illegitimate conceptions of the phenomena at hand, thereby failing to return "to the matters themselves."

The second way in which tradition is approached in phenomenological investigations is simply for its own sake, or for the sake of better understanding the form of the lived, meaningful world within which a person (or a people) finds herself. In this sense, one's totality of presuppositions is not made explicit for the sake of escaping the presuppositions and developing our concepts anew. Instead, these presuppositions are made explicit in order to better understand the meaningfulness of the world one resides within. While such investigations are neither phenomenologically preparatory, nor phenomenology proper, they have held a central place in the canon since the advent of genetic phenomenology.

An example of this latter kind of phenomenological study of tradition is found in the work of Iris Marion Young. In her essay, "Throwing Like a Girl," [25] she discusses the modes of feminine embodiment, but she also discusses the fact that such modes are tied up with a kind of tacit cultural background that shapes the meaningfulness of certain entities within our world or the kinds of meanings things have for us. As she explains, many women in the contemporary, western, affluent world have a sense of their bodies as fragile, weak, or even as an obstacle. The body is not actively interpreted in these ways, but simply shows up *as* fragile or weak in everyday experience. Young speaks of some of the biographical and historical conditions that led to such senses of the body, but this genealogical aspect is not particularly important here. Rather, what I wish to stress in Young's work is that the various modes of feminine body comportment that she outlines cannot be adequately understood without reference to the traditions in and through which one is able to come into contact with the world. In other words, in order to actually understand the form of one's lived world, we need to include an account of one's existentials, modes, and traditions.

5 LAYERS OF RESEARCH AND PHENOMENOLOGICAL PSYCHOPATHOLOGY

With the distinctions among these layers of phenomenological investigation clarified, we can reexamine them within the explicit context of phenomenological psychopathology. First, investigations into existentials, because their aim is to discover those characteristics of human existence that are considered necessary and universal, seem to have no place in phenomenological psychopathology. Psychopathology is, by definition, concerned with those aspects of human existence that can and do change. If existential structures are, in fact, invariant, then rather than being the objects of study for phenomenological psychopathology, they might instead act as the background, or framework, within which phenomenological studies of disorders can be conducted.

However, taking such an orthodox stance ignores some of the major developments of 20th century phenomenology—specifically those of Merleau-Ponty [26]. Through his engagement with cases of subjects with severe psychiatric and neurological disorders, Merleau-Ponty came to doubt the absolute necessity of the existentials discovered and articulated by Husserl and Heidegger. By reassessing the case of Schneider, a WWI veteran who underwent profound changes in his perception and motility after being struck in his occipital lobe by a piece of shrapnel, Merleau-Ponty was able to show that phenomenology could not do justice to Schneider's disorder if it remained bound to the belief in absolutely necessary existentials, or structures of human existence. In order to adequately articulate Schneider's disorder, he had to appeal to changes in certain categorial characteristics of human existence that neither Husserl nor Heidegger would have allowed for.

Merleau-Ponty's insights fundamentally altered the kinds of investigations open to phenomenological psychopathologists. However, the distinction between this new layer of investigation and the layer in which we can examine modes is not immediately clear in light of Merleau-Ponty's work. In order to adequately express the difference between changes in an existential structure itself, and changes in the mode of an existential structure, I here briefly outline two ways in which phenomenologists might characterize certain forms of depression.

One account might characterize the affective dimension of depression as a severe change in ground-mood, which is a pre-intentional, world-disclosive affect or feeling. This account, because it refers primarily to certain kinds of moods, and the role that these particular moods play in the disorder, is a modal account of depression. That is to say, it portrays depression as a distinctive mode of finding oneself situated in and attuned to the world.

An alternative account might characterize the affective dimension of depression not as a particular mood, or mode of situatedness, but instead as an erosion of situatedness and attunement as a whole. In other words, depression can be characterized by a degraded or diminished capacity for being situated in and attuned to one's world at all. Such an account better explains the loss of meaning or significance in the world of the depressed person, as well as the lack of intense moods, degraded affect, emotional insensitivity to context, and even diminished capacity for sensory stimulation.

Both accounts seem to capture important features of the experience of being depressed. However, what is important to note here is that the former account characterizes depression as a particular mood, or mode of attunement, while the latter account posits a change in the category of moods as a whole or, in other words, a change in the existential of situatedness. This illustrates the difference between phenomenological studies of changes in existentials, and phenomenological studies of changes in the modes of these existentials.

With the distinction between existential and modal changes made, we can examine the role that tradition, or one's totality of

presuppositions, plays within the context of phenomenological studies of psychopathology. Cooper [27] offers one such example in the context of a criticism of phenomenology's role in psychiatric classification. In order to undermine the role of phenomenology in delineating categories of disorder, she considers the possibility of "masked depression," a condition that received considerable attention in the mid 20th century, but is still discussed to some extent today. These conditions are described as "depressions that do not make people feel depressed" [27]. As she explains, "Those who believe in masked depressions claim that cultural conditions can make it the case that certain individuals manifest depression in atypical ways. For example, in a society that sees sadness as unacceptable weakness, patients might instead report somatic complaints" [27].

Cooper argues that if psychiatric conditions such as masked depressions exist, then phenomenological investigations of disordered subjectivity are not particularly important for psychiatric classification (although she admits that there are a few cases in which such investigations might be useful). She is able to come to this conclusion because masked depression is meant to illustrate the possibility of disorders with a single cause manifesting—and being expressed—differently within different traditions or cultural contexts. In other words, the experiences of depression can differ in important respects (even to the extent that one might be said to *not* experience his own depression), and this is used to claim that phenomenology—understood broadly as any analysis of subjective experience—is of little to no use in such cases.

In contrast to such arguments, I believe the distinctions I have drawn among the layers of phenomenological research can be used to overcome such a criticism and show how phenomenology is sensitive, at least in principle, to the implications of cultural differences in the manifestation of psychiatric disorders. Insofar as phenomenologists are actually considered with making explicit and overcoming our traditional prejudices, or totality of presuppositions, they are not simply describing lived experience or offering an account of the way things seem or appear to us. In order to get at the changes in existentials and modes involved in a particular disorder, phenomenologists need to attend to the possible ways in which such a disorder might be misinterpreted. Such an investigations might involve detailed studies of cultural norms and prejudices, along with standard characterizations of disorders in the DSM and other psychiatric literature, as well as historical studies of the characterizations and classifications of disorders.

6 VALIDITY AND THE LAYERS OF PHENOMENOLOGICAL RESEARCH

In light of these distinctions, and the possibilities they open up for phenomenological research into psychopathology, we can return to the earlier discussions of phenomenology's role in the project of reclassifying psychiatric disorders with the intention of increasing validity. We can ask about which layers of phenomenological research contribute to the various kinds of validity, and especially towards the project of neurobiologically validating disorders. While it may be the case that research in all three layers can enhance validity, the primary contributions are likely to come from descriptions of the existential, and in some cases modal, changes that comprise a particular kind of disordered subjectivity.

For example, modal investigations, such as phenomenological studies of the features of anxious moods or feelings in generalized versus social anxiety disorders, might enhance construct validity by showing that the moods and feelings associated with these disorders do not differ in any important respect. In this case, the only relevant distinction between the two kinds of disorders may be that the population diagnosed with social anxiety interprets large groups or social events as threatening or imposing. Because this account would characterize these two anxiety disorders as analogous in terms of modes, but dis-analogous in terms of traditions or tacit presuppositions, scientific research into the neurobiological correlates of moods and feelings may not need to distinguish between the two disorders in their investigations. However, psychotherapists may still find accounts of traditions and presuppositions relevant in order to change how people with social anxiety interpret and experience large groups or social events.

The history of phenomenology offers us even more evidence for the role that distinctions based on existential changes might play in the neurobiological validation of psychiatric disorders. As mentioned above, the possibility that such existentials, or existential structures, might be capable of changing (or even being absent) was not broached until Merleau-Ponty's *Phenomenology of Perception*. In this work, Merleau-Ponty takes Husserl and Heidegger to task for their transcendental assumptions that prove to be unjustified in light of the case studies of subjects with severe neurological disorders that he reexamined.[11] The fact that our only examples of such existential changes come from case studies of subjects with severe neurological disorders gives us reason to believe that other existential changes might also have relevant neurobiological correlates.

In sum, I have argued that phenomenology, specifically in the form of phenomenological psychopathology, is capable of offering accounts of disordered forms of subjectivity that can offer us preliminary categories of disorder that are likely to have greater validity than the categories currently available in the DSM. However, in order to properly engage in such a task, phenomenologists must be clear about the layers of their research.

Studies such as those discussed above can contribute *directly* towards enhancing both content and construct validity by supplying rich descriptions of disordered subjectivity, and by clearly distinguishing one kind of disorder from another by pointing out essential versus non-essential features of each disorder. Such clarifications can contribute *indirectly* towards other kinds of validity by offering symptomatically homogeneous categories of disorder that can then be used in neurobiological research, drug trials, outcome studies, and even psychotherapeutic interventions. While phenomenology may not be where psychiatry should end, it is certainly where it should begin.

[11] The particular example I have in mind is the case of Schneider, considered in detail in Part I of *Phenomenology of Perception*. However, Merleau-Ponty considers a number of other cases throughout this part of the text that may also prove useful as a model for phenomenological research into psychopathology.

REFERENCES

[1] R. Spitzer, APA and DSM-V: Empty Promises | Psychiatric Times, (2009). http://www.psychiatrictimes.com/apa-and-dsm-v-empty-promises (accessed April 6, 2015).

[2] A. Frances, A Warning Sign on the Road to DSM-V: Beware of Its Unintended Consequences | Psychiatric Times, (2009). http://www.psychiatrictimes.com/warning-sign-road-dsm-v-beware-its-unintended-consequences (accessed April 8, 2015).

[3] T. Insel, Director's Blog: Transforming Diagnosis, (2013). http://www.nimh.nih.gov/about/director/2013/transforming-diagnosis.shtml.

[4] B.N. Cuthbert, T.R. Insel, Toward the future of psychiatric diagnosis: the seven pillars of RDoC, BMC Med. 11 (2013) 126.

[5] B.N. Cuthbert, M.J. Kozak, Constructing constructs for psychopathology: The NIMH research domain criteria., J. Abnorm. Psychol. 122 (2013) 928–937. doi:10.1037/a0034028.

[6] E. Robins, S. Guze, Establishment of Diagnostic Validity in Psychiatric Illness: Its Application to Schizophrenia, Am. J. Psychiatry. 126 (1970) 107–111.

[7] A. Jablensky, R.E. Kendell, Criteria for assessing a classification in psychiatry, Psychiatr. Diagn. Classif. (2002) 1.

[8] A.V. Horwitz, J.C. Wakefield, The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder, Reprint edition, Oxford University Press, Oxford ; New York, 2012.

[9] I., Kirsch Ph.D, Irving Kirsch Ph.D.'sThe Emperor's New Drugs: Exploding the Antidepressant Myth, Basic Books, 2010.

[10] R. Whitaker, Anatomy of an Epidemic: Magic Bullets, Psychiatric Drugs, and the Astonishing Rise of Mental Illness in America, 1 edition, Broadway Books, 2010.

[11] R.E. Kendell, The classification of depressions: A review of contemporary confusion.pdf, Br. J. Psychiatry. 129 (1976) 15–28.

[12] G. Parker, Beyond major depression, Psychol. Med. 35 (2005) 467–474. doi:10.1017/S0033291704004210.

[13] G. Parker, Classifying clinical depression: an operational proposal: Discussion paper, Acta Psychiatr. Scand. 123 (2011) 314–316. doi:10.1111/j.1600-0447.2011.01681.x.

[14] G. Parker, Classifying depression: should paradigms lost be regained?, (2000). http://ajp.psychiatryonline.org/doi/10.1176/appi.ajp.157.8.1195 (accessed April 8, 2015).

[15] G. Stanghellini, Disembodied Spirits and Deanimated Bodies: The Psychopathology of Common Sense, 1 edition, Oxford University Press, Oxford ; New York, 2004.

[16] G.S. Malhi, G.B. Parker, J. Greenwood, Structural and functional models of depression: from sub-types to substrates, Acta Psychiatr. Scand. 111 (2005) 94–105. doi:10.1111/j.1600-0447.2004.00475.x.

[17] K.A. Aho, Depression and embodiment: phenomenological reflections on motility, affectivity, and transcendence, Med. Health Care Philos. 16 (2013) 751–759. doi:10.1007/s11019-013-9470-8.

[18] J. PARNAS, L.A. SASS, Varieties of "Phenomenology," Philos. Issues Psychiatry Explan. Phenomenol. Nosology. (2008) 239.

[19] M. Ratcliffe, A Bad Case of the Flu? The Comparative Phenomenology of Depression and Somatic Illness, J. Conscious. Stud. 20 (2013) 198–218.

[20] L.A. Sass, J. Parnas, Explaining schizophrenia: the relevance of phenomenology, Reconceiving Schizophr. (2007) 63–95.

[21] K. Jaspers, General Psychopathology, Reprint edition, Johns Hopkins University Press, Baltimore, 1997.

[22] K. Jaspers, The phenomenological approach in psychopathology, Br. J. Psychiatry. 114 (1968) 1313–1223.

[23] M. Heidegger, The Fundamental Concepts of Metaphysics: World, Finitude, Solitude, Indiana University Press, Bloomington, 2001.

[24] E. Husserl, The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy, Northwestern University Press, Evanston, 1970.

[25] I.M. Young, On Female Body Experience: "Throwing Like a Girl" and Other Essays, 1 edition, Oxford University Press, New York, 2005.

[26] M. Merleau-Ponty, Phenomenology of Perception, 1 edition, Routledge, 2013.

[27] R. Cooper, Psychiatric Classification and Subjective Experience, Emot. Rev. 4 (2012) 197–202. doi:10.1177/1754073911430139.

The juice is in the detail: an affordance-based view of talking therapies

Mark McKergow[1]

Abstract. The burgeoning interest in enactive paradigms of perception and cognition offers an opportunity to reconsider how we conceive psychotherapy – 'talking cures' as functioning. In the past many therapy modes have focused on the over-riding importance of giving insight to the patient; knowing what caused the 'illness' provides a solid way to deal with it. Over the past half-century, more pragmatic forms of therapy focusing on behaviour change through adjusted thinking (cognitive behavioural therapy) have become commonplace.

But what does it mean to 'change our thinking' from an enactive perspective? If perception and cognition are direct engagement with the environment, what is changed by a therapeutic conversation? One answer lies in the idea of affordances [1] – the relationships between features of the environment and the abilities of the animal/person to interact with them. Recent views of affordances as dynamic [2] make even clearer the ways in which these factors may change and evolve.

The paper compares an affordance based view with practical examples from solution-focused brief therapy (SFBT), where recent developments have pointed to the power of developing detailed descriptions of 'better futures' and 'past instances' [3]. In such detailed conversations, everyday and overlooked events such as hugging a loved one when they return from work can become significant possibilities for building recovery. The paper will show examples and how such detailed descriptions can develop new affordances for clients.

One key aspect is how these features emerge and are developed during the therapeutic conversation. Do they come from the therapist or the client? How can the therapist help the client develop new affordances that are relevant without intervening with their own ideas about 'what ought to happen'? The ways in which conversations about affordances can be seen to connect to strong and modest ideas of narrative development will also be explored briefly.

1 INTRODUCTION

In a symposium entitled 'Reconceiving Mental Illness', we are invited to think broadly about the topic. I intend to take this invitation seriously and present a novel view of both mental illness and how to enhance mental health. These topics have been discussed for centuries, and I cannot hope to present the full historical discussion here. Rather, I intend to set out some key points and then present a philosophical and practical case for a new way to look at mental illness though affordances.

One of the great truths (and for some, mysteries) of the mental health profession is that most if not all forms of talking therapy have broadly similar effectiveness. The huge metastudy of Wampold [4] showed that not only do different therapy modes have similar effectiveness, and drew attention to the overall importance of 'common factors' (first listed by Lambert [5]). These include therapeutic relationship/alliance, hope/expectancy, client factors and extraneous events. Despite this, the therapy world has continued to debate different models and approaches. One shortcoming of the Wampold study (and of most outcome studies) is the lack of consideration of the duration of therapy as of key interest. If everything 'works', then what works faster? During the heyday of psychoanalysis this was an unasked question, since it was common knowledge that mental disorders took years to deal with. During the past decades, however, there has been a rise in 'brief therapies', where the focus is on helping the client using 'as few sessions as possible' [6]. Such therapies typically take a handful of sessions to work [7].

There has been a bizarre obsession relating effective treatment to long-term therapy over the years, mainly due to the assumptions of psychoanalytic practitioners in the first half of the twentieth century. Clients and practitioners have grown more pragmatic in recent times, and now brief therapies are more valued. In a system such as the UK National Health Service where limited numbers of practitioners are available, the impact of shortening treatment can be huge. Lord Layard and colleagues [8] showed the huge impact of depression and other mental health problems – over a million people off work on incapacity benefit, in some cases waiting years to see a therapist who could help them in relatively short order (Layard mentions Cognitive Behavioural Therapy and 16 sessions). If the duration of therapy can be reduced from 16 sessions (itself brief by many standards) to closer to 4 sessions as shown by the latest brief therapy research [3], then four times as many people can be helped – even without recruiting extra therapists..

2 WHAT IS MENTAL ILLNESS?

This is a much contested question, about which there is little space to go into detail here. It looks so obvious at first, but unpicking the issues leads to considerable complication and confusion. The usual contrast is with physical illness – nobody would say that a broken leg was a mental condition. A stroke – a blood clot in the brain – can lead to speech impediments that can appear 'mental' (but probably should not be treated as such. Is pain mental or physical? Kendler [9] lists some of the key issues as causation (what causes mental illness, and in particular can it all be reduced to the brain, as some reductionists hope), the role of phenomenology and personal experience (which demands contact with the first person client situation rather than the third person expert) and nosology (the way that mental illnesses are classified). At present a pluralist view – different kinds of explanation are relevant – seems in the ascendant.

1 HESIAN, Department of Philosophy, University of Hertfordshire.

In general terms, most people think of mental illnesses as 'in the head'. One typical quote from the BABCP website [10] says:

"During times of mental distress, people think differently about themselves and what happens to them. Thoughts can become extreme and unhelpful. This can worsen how a person feels. They may then behave in a way that prolongs their distress."

This shows the assumption that thoughts precede behaviour – typical of the cognitive school of thought. This is so ingrained in our society as to go almost unchallenged – something 'inside' the person then appears on the 'outside' as behaviour. It is this assumption that the enactive paradigmn seeks to challenge.

3 THE ENACTIVE PARADIGM – DIRECT ENGAGEMENT

The enactive paradigm of perception and cognition is probably the most radical of the '4Es' [11] (embodied, extended, embedded and enactive) cluster of approaches which stem from the original work of Varela, Thompson and Rosch [12]. Briefly, rather than organisms taking in information ('perception') and then using it to make decisions about behaviour ('cognition'), the entire perception/cognition process is seen as a direct engagement with the environment. It is easy to see how this might happen for a blind person exploring a fruit bowl with their fingertips (or a pavement with their stick), but there are also indications and theories about visual perception based around sensorimotor rather than image-building processes [13].

Whereas the cognitive paradigm sees mechanisms in the head – either physical or mental – the enactive paradigm sees no need to posit mental representations. The world is its own representation, and carrying another around 'in our heads' would seem to be an unnecessary assumption. Indeed, Radical Enactive Cognition (REC)[14], the most extreme variety of enactivism, does away with all mental content. Another key distinction is the position of experience – our first person experience and awareness of what is happening to us. From a cognitive standpoint, experience is an epiphenomenon – a by-product of cognitive activity in the mind or brain (which are routinely superposed). In enactivism, experience is a primary element of cognition and is to be taken seriously in any description of 'mental' activity[15].

4 ROLE OF THE BRAIN – THE TASK/TOOL METAPHOR

This switch in emphasis can lead some readers to think that enactivism posits no role for thinking or the brain. This is of course incorrect. The brain is a vital organ, and removing it will seriously impede the thinking of the subject involved! Rom Harré's task/tool metaphor [16], [17] is a key way to understanding a way to look at the role of the brain from an embodied/enactive perspective.

Imagine somebody using a spade to dig a ditch. The person is using the spade to dig the ditch. The task is digging, and the tool , used by the person, is a spade. The spade does not dig the ditch – the person digs the ditch, using the spade. We could (and should) study spades – after all, a well-designed spade will be a great help in digging the ditch. We can (and perhaps also should) study digging. Note the studying spades is not the same as studying digging, and to study digging we will need a person who is digging to make any progress in our study.

Now switch the task and tool to thinking and the brain. A person uses their brain to think. The person thinks, not the brain. We could (and should) study brains. However, to study thinking will require a person to do the thinking, in the same way that a study of digging requires a digger. To take on the idea that a brain thinks (as opposed to a person) is to commit what Maxwell Bennett and Peter Hacker call the 'meriological phallacy' [18] – applying to a part something which should only be applied to a whole. In this case the brain is a part of a person, and a person thinks (remembers, fears, loves, forgets, sees, etc), not a brain.

Memory can be treated the same way. Some people, including St Augustine [19] and Jerry Fodor [20] assume that memories must be treated like mental representations, carried around for reproducing at the desired moment. An enactive perspective makes clear that remembering is an activity of a person (not a brain), and involves an active constructive process – a re-membering, a putting together (as opposed to dis-membering, to pull apart). This view is being accepted in both scientific [21] and philosophical [22], [23] circles.

We might note that taking the task/tool metaphor seriously already offers a line on what constitutes a mental illness. One could imagine a separation between illnesses of the brain (for example brain tumours, strokes and even Alzheimer's disease) and diseases of the person (for example depression, anxiety). This is not to say that people are not incapacitated by brain diseases – far from it. It is interesting to note that Alzheimer's disease is formally classified as a mental illness in both the USA (within the DSM V [24]) and the UK (under the Mental Health Act 1983), which is probably a good thing in terms of sufferers getting practical help and protection under the law, but raises an interesting philosophical question.

5 IMPLEMENTATION

This paper promises an affordance based look at talking therapies. This section will take a look at affordances and the development of the idea over the past decades.

The term 'affordance' was originally introduced by ecological psychologist JJ Gibson [1], [25] in the late 1970s. Gibson's theory of direct perception, a precursor to the enactive paradigm, has three headlines:

- Perception is direct
- Perception is for action
- Perception is of affordances

Affordances are an interaction of an animal and its environment – what kind of opportunities for interaction the environment is offering the animal, relating to the animal's sensorimotor capacities. A small tree branch, for example, may offer a bird somewhere to perch and observe the surroundings, whereas the same branch might offer a person a handhold, a chance to gather kindling for a fire, a back scratcher, a drumstick, a subject for a sketch and so on. The affordance is neither a property of the animal or the environment, but in the interaction of both. Gibson himself defined affordances in this way:

[An] affordance is neither an objective property nor a subjective property; or it is both if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy. It is equally a fact of the environment and a fact of behaviour. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer. (Gibson, 1979, p 129)

Many people read Gibson as saying that the affordance is there to be discovered by the animal, in suitable ambient light. Varela, Thompson and Rosch [12] note that embodied perception is not 'direct detection' but is sensorimotor enactment, 'dependent on histories of coupling'. We might think of this as a learning process. Varela, Thompson and Rosch are also keen to emphasise the co-determination of animal and environment.

"A cognitive system is functioning adequately when it becomes part of an existing ongoing world (as the young of every species do." (p 207)

Anthony Chemero takes the idea of affordances on another level [2] with his 'affordances 2.0 model'. Having already refined his definition in an earlier publication [26] to be about the relationship between abilities of the animal and features of the enviroment (stressing further the learning element involved in developing affordances), he offers a dynamical model working on two timescales – developmental and behavioural. This shows even more clearly how abilities and affordances co-develop over both the life of an animal and over longer timescales.

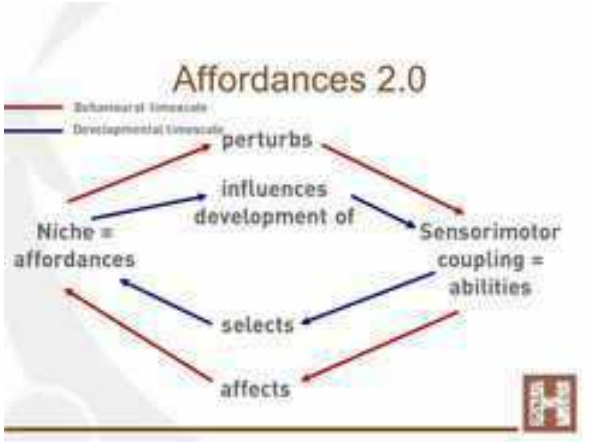


Figure 1: Affordances 2.0 (after Chemero, 2009)

Sanneke de Haan, Erik Rietveld and co-workers[27] have further developed these ideas by contrasting the 'landscape' of affordances with the narrower 'field' of affordances for an individual in a concrete situation.

"We distinguish between the landscape of affordances and a field of affordances. The landscape of affordances refers to all the possibilities for action that are open to a specific form of life and depend on the abilities available to this form of life. In our human case this notably includes socio-cultural practices. The landscape of affordances thus describes the so-called "ecological niche" of a form of life. A particular aspect of the environment, say a tree, can play a role in the landscape of affordances of multiple forms of life. Von Uexküll (Von Uexküll, 1920)[28] gives the famous example of an oak tree: for a rabbit it affords digging a hole between its roots, to a woodworm it provides food, for a person it could afford shelter from sun or rain, or cutting. The field of affordances refers to the relevant possibilities for action that a particular individual is responsive to in a concrete situation, depending on the individual's abilities and concerns. The field of affordances is thus a situation-specific, individual "excerpt" of the general landscape of affordances." (from De Haan et al, 2013)

The phrase 'form of life' in this paragraph is a nod back to Wittgenstein's [29] adoption of this phrase to signify a context where language has a (shared) meaning. The authors then develop a three dimensional model to describe the extent of a field of affordances. The three dimensions are:

- Width (broadness of scope and choice of options)
- Depth (temporal – now and in the future, with anticipatory affordance-responsiveness)
- Height (relavance/important of affordances, relating to motivation and 'affective allure'

De Haan et al, who are seeking a way to describe the changes produced by deep brain stimulation treatment on sufferers from Obsessive Compulsive Disorder (OCD), tentatively sketch out how a field of affordances may appear in three different cases:

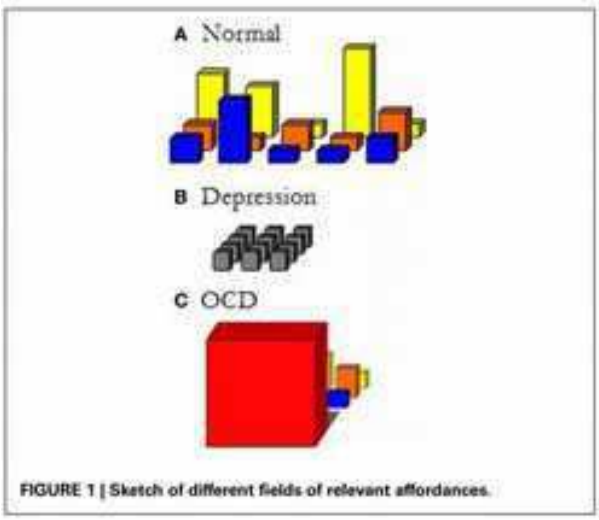


Figure 2: Sketch of different fields of relevant affordances (From De Haan et al 2013)

The 'normal' field of affordances show graspable variety in all three dimensions. The 'depression' verion shows everything looking the same – there is little in the way of meaningful choice or possibility of difference. The third diagram, reflecting the OCD case, shows one affordance (which may relate to washing hands or cleaning the house, for example) dominating the field in terms of importance. Note that these graphs are intended to be illustrative.

6 AN ENACTIVE VIEW OF MENTAL ILLNESS

German psychiatrist Thomas Fuchs offers an interesting way into a general discussion about enactivism and mental illness. In a paper [30] examining depression not as an inner and individual complaint, but as a detunement/disturbance ('Verstimmung') of 'the resonant body that mediates our participation in a shared affective' (which is very much stated in embodied and enactive terms), Fuchs harks back to phenomenologist psychiatrist Jan Hendrick van den Berg's pithy aphorism[31]: "*The patient is ill; this means, his world is ill."*

Fuchs elaborates on this position: "In this sense, the illness is not in the patient, but the patient is in the illness, as it were; for mental illness is not a state in the head, but an altered way of being in the world". (Fuchs 2013, p 222)

Taking the statement 'the world of the patient is ill', it is easy and tempting to fall back into a cognitivist picture that the world of the patient is inside the head of the patient. From an enactive perspective, the world of the patient is 'out here', in the interactions of the patient. The recent developments in the theory of affordances described above now offer a way to expand on this idea in more concrete terms.

The 'world of the patient' is the patient's field of affordances. Remember that this is an excerpt from the total landscape of affordances open to the patient's form of life. This is dynamic on many levels – including behavioural and developmental. So, if we take those mental illnesses best described as conditions of a person (as opposed to a brain disease), we can tentatively define this form of mental illness as:

A persistent Verstimmung (disturbance/detuning) of a field of affordancesa

These terms are carefully chosen:

Persistent: Not very temporary – we all have temporary disturbances in our worlds and deal with them by everyday actions. We feel a bit miserable and decide to go out for a walk and see some friends, for example. These are everyday ups and downs, and are dealt with routinely most of the time. Only if the 'ordinary' ways of dealing with something prove ineffective can we start thinking in terms of illness. This idea was first put forward by John Weakland and colleagues at the Mental Research Institute, Palo Alto in the 1970s [32], [33] and is still sound.

Verstimmung: This is a German word which has a number of meanings difficult to entirely sum up in English. These include disturbance, detuning, and leaving a bad mood. This is not a breakage – there is a sense in which the disturbance can be corrected. This is not, of course, referring to a bad mood which 'accompanies' the illness, the Verstimmung is key to the whole picture.

Field: This refers to the field of affordances relevant to this person in this context. This inevitably brings a first person perspective into action – different people will naturally have different fields of affordance, and in particular the therapist/practitioner will not be able to take on the client's field of affordance.

Of affordances: This is, again, not in the person or the environment (though it is hard to speak of them in those terms with the limitations of English grammer, as in the paragraph above) but in the relationship between the person and their environment, as shown in possibilities for action and engagement.

7 AN AFFORDANCE BASED VIEW OF TALKING THERAPIES

Psychotherapy has been characterised (and caricatured) as 'two people talking, trying to figure out what one of the wants'. All talking therapies have in common at least the talking element (though the topics of the conversation very dramatically between approaches). We can also recall the findings of Wampold [4] that all talking therapies are about as effective as each other in pure outcome terms.

What has never been done, as far as I know, is to look at talking therapy explicitly in the way it stretches and changes the client's field of affordances. On this basis, therapies which seek to address mental distress by a focus on long-passed causalities such as childhood trauma and familial relations might be expected to take a long time to work, whereas therapies focusing more on details of the a better future might be expected to bring more rapid progress.

If we are to look at talking therapy as helping to stretch the client's field of affordances in useful ways that connect to progress, we might expect to look for:

- The therapist taking the client as an active participant in the treatment
- The therapist taking the first person perspective/descriptions very seriously
- The therapist not attempting to discover what has caused the problem, but rather establishing a conversational narrative around progress in the past, present and future
- The conversation being focused on small details of a 'better world' – signs that things were improving.

One might expect that such a stretching of the field of affordances might have an emergent quality about it – sometimes neat, sometime messy, sometimes clear, sometimes confusing. To stretch a field of affordances is not the same as to provide key steps for action to the client.

Might such a therapy be effective? Well, there is already one that works in much the above fashion which is indeed effective – Solution Focused Brief Therapy (SFBT).

8 SFBT THROUGH AN AFFORDANCE LENS

Solution-Focused Brief Therapy (SFBT) was devised by Steve de Shazer, Insoo Kim Berg and colleagues at the Brief Family Therapy Center in Milwaukee WI in the 1980s [34], [35]. It has since spread around the world, being widely used in education, social work, organisational change as well as therapy, with a significant evidence base [7], [36]. The approach appeals to those who value a pragmatic and skilful approach to building progress, but it has not been widely supported by psychiatrists and medical professionals for whom it lacks proper 'theoretical' grounding. De Shazer, Berg and colleagues started with the interactional brief therapy approach devised by Weakland and others, and experimented with trying to make it both more minimal (in terms of the therapist's model and theory) and more efficacious (in terms of fewer sessions to help clients reach a position where they could carry on under their own steam,

without continuing therapy). In this way, the practice could be said to be pragmatically and empirically rooted.

The latest and most stripped down version of SFBT is that proposed and practiced by the BRIEF group in London [3]. In a typical first session, the therapist will:

- Discuss 'best hopes' of the client for the work together – a theme for the project
- Elicit a description of a 'preferred future' – with these best hopes realised
    - Tomorrow (usually)
    - Detailed and observable (referent)
    - From client's perception and relevant others' positions – spouse, colleagues etc
    - Suppose… all about how it could be, not how to get there
- Elicit 'instances' – in the past and/or present – of the preferred future happening already
    - Often using a scale from 1-10
    - Details, details, details…

In follow-up session(s), the therapist will ask about 'what's better?' since last time, seek more details about how the client managed to do that, and summarise progress so far. Using this model, Shennan and Iveson report (over an admittedly small number of clients) an average therapy duration of under four sessions.

It is generally found in practice (by me and others) that getting these conversations down into small tiny details is important. SFBT co-founder Insoo Kim Berg used to advise therapists learning the approach to value '$5 words' (very small concrete and everyday words) over the '$5000 words' of abstraction and professionalism typically used valued by self-important experts. I want to put forward the idea that these details are connected with stretching the field of affordances.

9 A REAL LIFE EXAMPLE: MANDY AND THE CUDDLE

To give a brief flavour of an SFBT session, I include here a very short excerpt from a real conversation. 'Mary' (not her real name) has been referred for treatment following long term depression and suicide bids. This is her first session. The therapist (Chris Iveson of BRIEF) is in the middle of helping Mary to describe a better tomorrow, when an imagined miracle has realised her self-defined hopes of 'the past not pulling her back any more'. After about 25 minutes, they reach a point in the day when Mary's partner Jeff will return from work.

Therapist: And what is the first thing he would notice when he got home, even before you spoke? What is the very first thing?

Mary: I would be… instead of a worried, stressed, anxious look on my face maybe a smile.

Therapist: Okay. And what would be the first thing you would notice about his response even before he spoke?

Mary: I think my body language would just be so… you know normally he has to come looking for me whereas I would imagine that I would be open to go and cuddle him instead. You know? So…

Therapist: Would he faint or…?

Mary: Possibly, yeah, absolutely. You might have to have the paramedics on standby, yeah. I think it would be shock, but pleasant shock rather than shock shock.

Therapist: So where would that be? Where would you be cuddling him?

Mary: I would imagine that… because I do almost always hear him pull up. I never go to the door. I let him come in through the door and come find me. Whereas I would probably go find him.

Therapist: Okay, so that would be a different…

Mary: Yeah.

Therapist: And what would you notice about the way you cuddled him that fitted with this sense of peace and pleasure, of being you?

Mary: He describes sometimes that when he asks me for a cuddle… he said 'When I ask you for a cuddle…' and I do give it to him, he goes 'You are rigid and you almost… you cuddle me but you are pushing me away.' So I would imagine that it would be a much more natural, open embrace where I felt relaxed and safe enough to do that. Not rigid and tight.

Therapist: And what would you notice about his response to your cuddling and that kind of relaxed…?

Mary: I think that he would be delighted with how it felt to have a cuddle that didn't feel like he was a) having to ask for or b) being pushed away from.

Therapist: And what would you notice about his arms?

Mary: I think they might be quite tight around me and probably hold me for longer than normal.

Therapist: Okay. And what would you notice about how you handled that?

Mary: I think it would be quite difficult because you get so rehearsed in how you do things. Whether that be good or bad, that's how you are. So I think it would be quite a new experience to have that.

Therapist: And if you are feeling like hugging him?

Mary: Not wanting to let go either rather than wanting to break that embrace.

Therapist: Okay.

Mary: Because at the moment it's like 'Okay, cuddle, quick, out of the way.' Whereas to actually enjoy the embrace and feel it rather than just do it and break away from it.

Therapist: And what would you notice about him as you do eventually break away from the embrace?

Mary: I think that he would possibly be very happy to have experienced a… not always having to want to ask. To find… you know, for me to acknowledge his needs and be able to actually do that for him.

Therapist: And how would he know that you are pleased to have had that embrace? What would he notice about you?

Mary: Because I wouldn't be rushing away from him, looking at the next task that has to be done. It's like hugging Jeff is on the list, I've got to do that and then I've got to get on and do this and do that. I probably would maybe just stand there with him maybe and chat about his day rather than rush off and try and do something different.

Therapist: Is that when you might suggest a walk or would that be…?

Mary: After dinner maybe.

Therapist: After dinner? Okay. So what might you have for dinner?

Note that the therapist is not himself contributing to the details. He is rather asking questions which help Mary come up with her own details. He asks questions such as:

- And what is the first thing he would notice when he got home, even before you spoke?
- And what would you notice about the way you cuddled him…
- What would he notice about you?
- And how would you respond, when he did that?

These questions are all in the context of Mary describing a future (tomorrow) that is both utterly mundane and yet transformed by the realisation of her own hopes. She is stretching and changing her world in response to the therapist's questions – and because the talk is of a better future, the stretching is in a potentially useful direction. (We might note that many therapeutic approaches take a lot of time talking about what happens when the problem occurs or started, which might be stretching the world in an unhelpful way.)

For clarity, some of the affordances discussed in the excerpt above might be:

- The sound of Jeff pulling up as an opportunity to go and meet him.
- Jeff's appearance as an opportunity for cuddling in a particular way.
- The cuddle as a longer engagement rather than something to be broken off.

I say these 'might' be affordances in the conversation. We cannot say from a third person perspectives what are new or important affordances - we would have to ask Mary herself. And I am not saying that it's now simply a matter of Mary going and doing these things – her world has been stretched, her field of affordances altered, and now life will go on. It is only later that the impact will be clarified.

Previous versions of SFBT have focused on the conversation as a route to the therapist being able to establish tasks or actions for the client to help them 'do more of what works'. The latest thinking from BRIEF, the author [37] and others is that such a direct interventionist approach is unnecessary – either asking the client what they are minded to do next, or even simply leaving that out of the conversation altogether seems even more effective. It is worth noting that when the client's description is as detailed as the example above, all sorts of tiny actions and reactions have become possibilities in a revised world. This supports my hypothesis that the world-stretching is the key, rather than any post-rationalising that may go on between client and therapist (though such further conversation may strengthen the new world in some way).

10 TALKING ABOUT AFFORDANCES AND BUILDING AFFORDANCES

We might legitimately ask about the connection between describing affordances and creating/using them. From a cognitive standpoint there is all the difference in the world between talk and action. From an enactive standpoint, the difference is considerably reduced. In order to describe something, the client has to somehow put themselves into a different world. And once it's been described it can't be undescribed – echoes of the social constructionist idea of Ken Gergen that we carry around all our previous interactions as potentials for action [38]. There is even a view that from the first person perspective of the client, there is no fundamental difference between information through language and through visual and corporeal channels [39], [40]. There is no space to go further into this fascinating position here.

One point worth making in closing – how this position relates to a narrative perspective, itself a popular strand of therapeutic thinking and practice with similarities and differences to SFBT[41]. There are some who hold 'strong narrative' views that everything in life should be viewed in narrative terms [42]. Others, with whom I would align my position [43] take a more modest view, embracing the idea that narrative offers a useful view rather than an overarching mechanism. This is consistent with the task/tool metaphor for the mind, where discourse is a key but not exclusive element.

11 CONCLUSIONS

This paper has covered a great deal of ground very quickly and lays out a potential agenda for investigation. The key points are:

- Affordances offers a new perspective for talking therapies
- There is intitial evidence that this perspective is useful on a practical basis
- This may go some way to show why some therapies take a lot longer than others
- This perspective offers a researchable hypothesis for even more effective forms of talking therapy.

ACKNOWLEDGEMENTS

This work was carried out as part of the HESIAN programme (Hertfordshire Enactive Solution-focused Interactional and Narrative) at the University of Hertfordshire. http://herts.ac.uk.hesian. The author would like to thank Dan Hutto, Brendan Larvor, Daniele Moyal-Sharrock and Zuzanna Rucinska for helpful discussions on the philsophy and Chris Iveson (BRIEF) for many useful conversations and use of the transcript example.

REFERENCES

[1] J. J. Gibson, The ecological approach to visual perception. Boston: Houghton Mifflin, 1979.
[2] A. Chemero, Radical Embodied Cognitive Science. Cambridge MA: MIT Press, 2009.
[3] G. Shennan and C. Iveson, "From solution to description: Practice and research in tandem," in Solution-Focused Brief Therapy: A Handbook ofEvidence-Based Practice, C. Franklin, T. S. Trepper, E. E. McCollum, and W. J. Gingerich, Eds. Oxford: Oxford University Press, 2011, pp. 281–298.
[4] B. E. Wampold, The Great Psychotherapy Debate: Models, Methods, and Findings, vol. 62. 2001.
[5] M. J. Lambert, "Implications of Outcome Research for Psychotherapy Integration," in Handbook of Psychotherapy Integration, J. C. Norcross and M. R. Goldstein, Eds. New York: Basic Books, 1992, pp. 94–129.
[6] S. de Shazer, Keys to solution in brief therapy. New York: WW Norton, 1985.
[7] C. Franklin, T. S. Trepper, E. E. McCollum, and W. J. Gingerich, Solution-focused brief therapy: A handbook of evidence-based practice. Oxford University Press, 2011.

[8] R. Layard, D. Clark, S. Bell, M. Knapp, B. Meacher, S. Priebe, L. Turnberg, G. Thornicroft, and B. Wright, "The depression report; A new deal for depression and anxiety disorders," London: LSE, 2006.
[9] K. S. Kendler, "Why does psychiatry need philosophy," in Philosophical Issues in Psychiatry: Explanation, Phenomenology and Nosology, Baltimore: Johns Hopkins University Press, 2008, pp. 1–16.
[10] "What is CBT?," British Association for Behavioural and Cognitive Psychotherapists. [Online]. Available: http://www.babcp.com/files/public/what-is-cbt-web.pdf. [Accessed: 04-Apr-2015].
[11]M. Rowlands, The New Science of the Mind: From Extended Mind to Embodied Phenomenology. MIT Press, 2010.
[12] F. Varela, E. Thompson, and E. Rosch, "The embodied mind," Cogn. Sci. Hum. Exp., 1991.
[13] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," Behav. Brain Sci., vol. 24, no. 5, pp. 939–973, 2001.
[14] D. D. Hutto and E. Myin, Radicalizing Enactivism: Basic Minds without Content. Boston: MIT Press, 2013.
[15] M. McGann, H. De Jaegher, and E. Di Paolo, "Enaction and psychology.," Rev. Gen. Psychol., vol. 17, no. 2, pp. 203–209, Jun. 2013.
[16] R. Harré, "Tasks, Tools and the Boundaries of the Discursive," Cult. Psychol., vol. 7, no. 2, pp. 145–149, Jun. 2001.
[17] R. Harré and F. M. Moghaddam, Psychology for the Third Millennium Integrating Cultural and Neuroscience Perspectives. London: Sage Publications, 2012.
[18] M. R. Bennett and P. M. S. Hacker, Philosophical Foundations of Neuroscience. Oxford: Blackwell, 2003.
[19] Augustine, Confessions, 2nd edition. Indianapolis: Hackett Publishing, 2006.
[20] J. Fodor, Hume Variations. Oxford: Oxford University Press, 2005.
[21] D. L. Schacter and D. R. Addis, "The cognitive neuroscience of constructive memory: remembering the past and imagining the future.," Philos. Trans. R. Soc. Lond. B. Biol. Sci., vol. 362, no. 1481, pp. 773–86, May 2007.
[22] D. Moyal-Sharrock, "Wittgenstein and the Memory Debate," New Ideas Psychol., vol. 27, no. 2, pp. 213–227, 2009.
[23] E. Myin and K. Zahidi, "The Extent of Memory: From Extended to Extensive Mind," in Mind, Language & Action, Proceedings of the 36th International Wittgenstein Symposium., D. Moyal-Sharrock, V. Munz, and A. Coliva, Eds. Ontos Verlag, 2014.
[24] American Psychologyical Association, Diagnostic and statistical manual of mental disorders (5th ed.). Washington, DC: Author, 2013.
[25] J. J. Gibson, "The Theory of Affordances," in Perceiving, Acting, and Knowing, R. Shaw and J. Bransford, Eds. New Jersey: Lawrence Erlbaum, 1977.
[26] A. Chemero, "An Outline of a Theory of Affordances," Ecol. Psychol., vol. 15, no. 2, pp. 181–195, 2003.
[27] S. de Haan, E. Rietveld, M. Stokhof, and D. Denys, "The phenomenology of deep brain stimulation-induced changes in OCD: an enactive affordance-based model.," Front. Hum. Neurosci., vol. 7, p. 653, Jan. 2013.
[28] J. Von Uexküll, Theoretische Biologie. Berlin: Paetel, 1920.
[29] L. Wittgenstein, Philosophical Investigations. Oxford: Blackwell, 1953.
[30] T. Fuchs, "Depression, Intercorporeality, and Interaffectivity," J. Conscious. Stud., vol. 20, no. 7–8, pp. 219–238, 2013.
[31] J. H. van den Berg, A Different Existence: Principles of Phenomenological Psychopathology. Pittburgh PA: Duquesne University Press, 1972.
[32] P. Watzlawick, J. Weakland, and R. Fisch, Change: Problem Formation and Problem Resolution. New York: WW Norton, 1974.
[33] J. H. Weakland, R. Fisch, P. Watzlawick, and A. M. Bodin, "Brief Therapy: Focused Problem Resolution," Fam. Process, vol. 13, no. 2, pp. 141–168, Jun. 1974.
[34] S. de Shazer, I. K. Berg, E. Lipchik, E. Nunnally, A. Molnar, W. Gingerich, and M. Weiner-Davis, "Brief therapy: focused solution development.," Fam. Process, vol. 25, no. 2, pp. 207–21, Jun. 1986.
[35] S. de Shazer, Clues: Investigating solutions in brief therapy. New York: WW Norton, 1988.
[36] A. J. Macdonald, Solution-focused therapy: Theory, research & practice (second edition). Sage Publications, 2011.
[37] P. Z. Jackson and M. McKergow, The Solutions Focus: Making Coaching and Change SIMPLE. Nicholas Brealey, 2007.
[38] K. Gergen, "Relating with Self and Others," Interact. J. Solut. Focus Organ., vol. 5, no. 1, pp. 9–25, 2013.
[39] L. L. Barrett, Beyond the Brain: How Body and Environment Shape Animal and Human Minds. Princeton University Press, 2011.
[40] A. D. Wilson and S. Golonka, "Embodied Cognition is Not What you Think it is.," Front. Psychol., vol. 4, p. 58, Jan. 2013.
[41] M. White and D. Epston, Narrative Means to Therapeutic Ends. New York: WW Norton, 1990.
[42] A. Rudd, Self, Value, and Narrative. Oxford: Oxford University Press, 2012.
[43] D. D. Hutto, "Narrative self-shaping: a modest proposal," Phenomenol. Cogn. Sci., Feb. 2014.

Are mental disorders illnesses? The boundary between psychiatry and general medicine

Valentina Petrolini[1]

Abstract. Are mental and physical disorders meaningfully comparable? Are we are entitled to characterize psychiatric disorders in terms of illnesses? Traditionally, most attempts to define what counts as an illness rely on some notion of *normal functioning* that has been altered or disturbed, where the "norm" is established from an evolutionary (Wakefield 1992; De Block 2008) or statistical perspective (Boorse 1977). In this sense, the substantial distinction between somatic and mental disorders may just reflect different stages of development within medical disciplines. In general medicine, clinicians have a clear idea of how organs normally function and thus can detect illnesses smoothly or with a small margin of error. The psychiatric case looks *prima facie* different: we currently lack an ideal model of brain functioning and the high variability among patients renders the diagnostic process particularly tricky. This argument reduces the distinction between psychiatry and general medicine to a *practical* matter. On this view, the high rates of misdiagnosis and disagreement among experts in the classification of mental disorders simply derive from a lack of knowledge about the brain (see Aboraya et al. 2006).

The main goal of this paper is to assess the argument above by showing that it stems from an overly simplistic conception of medical practice. On one hand, the diagnostic process in *general medicine* is not as straightforward as it initially appears, as some interesting studies on error and cognitive bias have recently shown. On the other, the core distinction between *psychiatry* and general medicine does not simply rest on practical issues: rather, the former exhibits some methodological peculiarities that are rejected by other disciplines within the medical field.

The paper is divided in four sections: in *§1* I motivate the need for more theoretical precision in defining the notion of illness, making the case particularly compelling for psychiatry. In *§2* I discuss some recent empirical studies on diagnostic error and cognitive biases in general medicine, and in *§3* I evaluate whether these results can be meaningfully applied to psychiatry. Finally, in *§4* I outline a *medical model* that aims at encompassing both somatic and mental disorders: in particular, I argue that in order to incorporate psychiatry within general medicine we need to adopt a *multi-level*, *holistic* and *dimensional* approach to illness.

1 THE NEED FOR THEORETICAL PRECISION

Within philosophy of psychiatry, the attempts to gain clarity from current definitions of *mental illness* have encountered a common difficulty. Psychiatry is a branch of medicine and thus a practical discipline whose main goals are to treat patients and alleviate suffering. As a result, not much work has been done to define concepts with theoretical precision, as suggested by the heated debate around classification and the DSM's new edition (see Cooper 2004 and Frances 2012). Consequently, among clinicians the question: "Is X a disease?" is often used as a shortcut for: "Should the person affected by X be subject to medical treatment?" This approach seems immediately problematic because doctors recognize that some conditions do not qualify as illnesses despite being treated (e.g. pregnancy or circumcision). Thus, the crude conditional: "*If* X needs to be treated, *then* X is a disease" should be discarded, at least because it does not reflect the common practice within medical sciences.

However, any attempt to define *mental illness* rests on having some conception of what counts as an illness in general: in this sense, the analogy between somatic and mental disorders becomes of paramount importance. On one hand, the two classes should be similar enough to be subsumed under the common label of "illness"; on the other, they should be different enough to motivate a principled distinction between the two sub-groups (see Brülde & Radovic 2006 and Brülde 2010). This network of similarities and differences between somatic and mental disorders has been extensively discussed both in the philosophical and psychiatric literature. For example, Culver & Gert (1982) attempt to draw the line by arguing that physical pain is "always *localized* to some part of the body" whereas mental suffering "is experienced by the *whole person*" (p. 89. Italics mine). Other authors – such as Boorse (1975) – adopt a more skeptical attitude by calling into question the validity of the analogy itself: "It seems an open question whether current applications of the health vocabulary to mental conditions have any justification at all" (p. 50). At the extreme of this spectrum, Szasz (1974) completely rejects the medicalization of mental disorders and argues that psychiatry should rather be concerned with "problems of living" – e.g. behaviors deviating from socio-cultural, moral or political norms.

1 Dept. of Philosophy, Univ. of Cincinnati, 2700 Campus Way, 206 McMicken Hall, Cincinnati (OH) 45221, USA. Email: petrolva@mail.uc.edu

Despite the difficulty to devise a precise definition, there are – at least – two reasons for advocating a more rigorous characterization of psychiatric illness:

a) The *social consequences* connected to a diagnosis of mental illness dramatically differ with respect to the ones connected to somatic ailments. Indeed, being classified as somatically ill presents a mixture of harmful and beneficial consequences for the patient (e.g. distress but also sympathy or support) whereas most mental disorders are still associated with various forms of stigma (e.g. shame, exclusion, discrimination). Since the personal and social implications of a psychiatric diagnosis may be highly disruptive for the patient, the highest level of precision would be needed in defining mental illness. This consideration becomes especially important in the light of Szasz's concerns about social control. For instance, equating "illness" with "in need of treatment" could allow psychiatrists to categorize all deviant beliefs and behaviors as mentally ill and thereby exercise some sort of coercive power over patients (see also Foucault 1964).

b) The identification of mental disorders also presents *legal* and *ethical* implications. For instance, most criminal systems do not rely on strict liability and thereby allow for excusing conditions (e.g. insanity). In the US, the *M'Naghten Rule* states that in order to successfully establish a defense on the grounds of insanity the party accused has to prove that – at the time of the crime – s/he was either not knowing the nature and quality of the act or s/he was not knowing that the act was wrong. Such a principle strongly connects legal and moral responsibility by acknowledging that no one should be punished for an action that was not committed voluntarily, but rather resulted from a "defect of reason" or a "disease of the mind" (see *M'Naghten Rule*). Again, these cases demand the highest level of precision: a sloppy characterization of mental illness runs the risk of unjustly punishing someone who should have been excused or applying the rule to someone who should have been convicted.

What a) and b) illustrate is that although instances of misdiagnosis in general medicine may have disruptive consequences (e.g. death of the patient), a lack of theoretical precision in psychiatry harbors implications that extend to the social, legal and ethical realm. Therefore, a more rigorous definition of illness that would comprise mental disorders is both desirable and called for.

2 DETECTING ILLNESS IN GENERAL MEDICINE: DIAGNOSTIC ERRORS AND COGNITIVE BIASES

From the discovery of bacteria to more recent microscopic and post-mortem techniques, diseases have come to be characterized in terms of "deranged biophysical structures, genes and molecules" (Kendell 1975, p. 306.). To this day, the most straightforward way to define somatic illness is by appealing to some form of *lesion* or structural *damage* of the body. This standard view raises three main issues: first, it relies on some notion of normal functioning that needs to be spelled out more or less precisely (e.g. prototypes). Second, it needs to account for individual variation while at the same time drawing a line to establish "where normality ends and abnormality begins" (Ibid., p. 308). Third, since not all deviations from the norm would be harmful – e.g. exceptionally high IQ – a distinction between positive, neutral and negative variations is needed.

Despite these potential problems with classification, general medicine seems to fare much better than psychiatry in terms of accuracy and reliability. The high rates of misdiagnosis and disagreement among psychiatrists support this point: for example, Kirk, Gomory & Cohen (2013) cite a recent estimate according to which the diagnostic error rate is 38% for ADHD and 21% for Oppositional Defiant Disorder (p. 170). The rationale behind this argument seems to be the following: *reliability* works as an indicator for the *validity* of a medical category, since a sound classification allows practitioners to distinguish between disorders and non-disorders in most circumstances. Due to the proliferation of false positives and false negatives, psychiatry's reliability appears tainted and consequently the whole classification of mental disorders is called into question. Yet, here I argue that the appeal to diagnostic unreliability *per se* fails to draw a meaningful distinction between psychiatry and other branches of medicine. To support this point, I discuss a growing body of literature focused on *error* and *accuracy* in various medical disciplines, showing that the diagnostic process – even for somatic disorders – is far from straightforward. These results are particularly interesting because they show that a complex array of factors – e.g. biases, modes of reasoning – can easily influence diagnosis. More specifically, *cognitive factors* are estimated to be responsible for the majority of errors: for example, in internal medicine 74% of the misdiagnoses appear to have such an origin (see Graber, Franklin & Gordon 2005).

In a recent study, Graber & Berner (2008) confirm that "diagnostic errors exist at non-trivial and sometimes alarming rates" (p. S6). The extent of incorrect diagnoses varies significantly according to the specialty, with perceptual disciplines – such as radiology – scoring lower (2-5%) and clinical ones higher (12-15%). Other important factors seem to be the context of stress or uncertainty that facilitates hasty decisions (e.g. emergency room), whereas the presence of a second opinion tends to increase accuracy. Yet, studies using standardized vignettes to enable comparisons across experts show that clinicians wildly *disagree* with one another, and sometimes "even with themselves when presented again with a case they have previously diagnosed" (p. S5). Another core issue seems to be the *lack of feedback*: most physicians regard diagnosis as a "'one-shot deal', […] a stand alone, discrete episode of judgement" rather than a process that stretches over time and can be refined or amended through multiple interactions with the patient (p. S34). In particular, doctors do not take advantage of autopsies as an opportunity to learn from past mistakes, although – on average – 25% of autopsies reveal new problems that were not suspected clinically (p. S5).

Graber & Berner also present a series of studies on the issue of *overconfidence*, arguing that it may significantly contribute to diagnostic error. The level of overconfidence can be measured through practical indicators such as the clinician's tendency to disregard decision-support resources even when they are available and easy to access (e.g. national clinical guidelines). Cognitive aspects – e.g. arrogance, excessive reliance on expertise – can instead be observed through the failure to elicit complete information from the patient and the biased interpretation of results. Interestingly, all the studies point to a systematic misalignment between the degree of confidence and the degree of correctness: "The level of physician confidence showed no correlation with their ability to predict the accuracy of their clinical diagnosis. […] The confidence level of the worst performers was actually higher than that of the top performers" (p. S8). Friedman et al. (2005) offer more results in support of the negative impact of overconfidence on diagnostic accuracy. This study measured the tendency to seek for external tools in the diagnostic process (e.g. computer-based support systems, advice from colleagues), finding again a correlation between high levels of confidence and errors. In a nutshell, overconfident physicians seem less likely to look for external sources to back up their decisions, thereby increasing the possibility of error.

Other studies focusing more specifically on *cognitive factors* (e.g. flawed reasoning, faulty data gathering, poor interpretation) have been carried out by Mamede and her collaborators (2008 & 2010). Great part of their work aims at drawing clearer distinctions between the modes of reasoning used by physicians when performing diagnoses. Apparently, doctors tend to switch back and forth between two alternative cognitive styles. On one side, *non-analytical reasoning* based on the recognition of similarities between "illness-prototypes" and the case under review; on the other, *reflective reasoning* based on the effortful and step-by-step analysis of specific features. Mamede et al. (2008) show that factors such as the perceived difficulty of a case can influence the way in which physicians approach diagnosis: for example, it is sufficient to tell them that other colleagues have failed to interpret the situation correctly to trigger the passage from non-analytic to reflective mode. In the experiment two groups of physicians were asked to work on the same case descriptions, but only one of them was primed to see the context as "problematic": as a result, this group spent more time on the diagnosis and displayed a significant increase in accuracy.

Another possible interpretation of this result – not discussed by Mamede – draws on the overconfidence studies just discussed: when cases are perceived as more difficult, the level of confidence may decrease and then lead to a more accurate assessment of the situation. In other words, knowing that a colleague has already failed in evaluating a case would attenuate overconfidence and force the physician to evaluate the context more carefully – e.g. spending more time on the diagnosis or taking alternative possibilities into consideration. This interpretation is consistent with the data presented by Mamede: in the contexts perceived as "non-problematic" the rate of confidence was higher and the level of diagnostic accuracy lower, whereas in the "problematic" cases the opposite occurred.

A more recent study (Mamede et al. 2010) uncovers the fact that experience with clinical cases similar to one another may trigger inaccuracy: indeed, physicians tend to perceive the diagnoses that come to mind more easily as correct even when they are not (*availability bias*). This bias also seems to get worse as expertise increases, suggesting again either a switch to non-analytical reasoning over time or the development of detrimental overconfidence. Like in the previous study, a combination of both factors might influence the diagnosis, since experience usually correlates with a greater number of cases encountered as well as with an increase in confidence.

These studies show that appealing to reliability to motivate a distinction between psychiatry and general medicine may be misguided. Indeed – contrary to most expectations – alarming rates of misdiagnosis and cognitive biases affect various medical disciplines in a similar way. Therefore, taking reliability *per se* as an indicator for validity does not create a meaningful contrast between psychiatry and other branches of medicine, since they all appear to have serious issues with diagnostic accuracy. Rather, it would be more fruitful to acknowledge that the lack of accuracy can be caused by different *kinds of factors*. Some of them may be mitigated or corrected without having to change the underlying structure of the discipline (e.g. biases, modes of reasoning); others may require a more profound revision of assumptions and methodology (e.g. faulty taxonomy). In this section I have shown that diagnostic issues in general medicine normally arise from factors of the first kind; in the next section I turn to psychiatry and argue that factors of the second kind are more pervasive.

3 DETECTING ILLNESS IN PSYCHIATRY: PRACTICING IN A MINEFIELD

The very idea of applying the results on cognitive biases and reasoning errors to psychiatry has generated a good deal of controversy. For example, Groopman's book on medical reasoning – *How Doctors Think* (2007) – purposefully excludes psychiatry from the discussion: "I quickly realized that trying to assess how psychiatrists think was beyond my ability" (p. 7). Moreover, despite the common complaint about the high rates of misdiagnosis in the field, the empirical literature on psychiatric errors is still quite small and the few exceptions tend to focus on other aspects of the practice (e.g. medication errors). Some researchers – such as Crumlish and Kelly (2009) – have attempted to counteract this tendency by arguing that the cognitive style employed by psychiatrists is not "esoteric" or "un-understandable" but rather similar to the one employed in other medical disciplines (p. 72). Others have defended a mixed approach, according to which psychiatric practice may commit errors that are common to other medical specialties but also faces a series of additional issues due to its unique patient population. For example, Cullen, Nath & Marcus (2010) point out that the peculiar features of psychiatric patients may have an impact on the "nature, prevalence and preventability" of the errors affecting them (p. 198). Interestingly, in this study *diagnostic* errors are the least commonly mentioned by

practitioners (9%), whereas *medication* errors account for approximately one-third of the total (34%) and *preventive* errors – e.g. failure to implement safety protocols – stand at the top (40%). Both medication and preventive errors are motivated by factors unique to the psychiatric setting, such as the lack of expertise in dealing with some extreme behavioral manifestations (e.g. violence, resistance to treatment) and various forms of stereotypes and stigma towards patients.

These data show that the topic of diagnostic reliability remains rather unexplored in psychiatry. Yet, the fact that diagnostic errors are both less reported and less investigated may indicate a more substantial difference between psychiatry and other medical disciplines. As Phillips (2014) put it: "You cannot detect error unless you have a reliable, valid method of making diagnoses. Since the diagnostic process is less certain in psychiatry than in general medicine, that will make the detection of error less confidant" (p. 75). One asymmetry arises from the fact that psychiatry does not avail itself of laboratory tests or biomarkers, and detects disorders almost entirely through clinical evaluations (e.g. structured interviews). Due to this unavailability of external resources to back up the diagnosis, psychiatry often lacks reliable methods to spot cases of under-reporting or over-reporting. For these reasons, the level of risk and uncertainty already connected to general medicine becomes higher in psychiatric practice, to the point that the diagnostic process "could be likened to a minefield" (Kapur 2000, p. 399). However, at this stage the problem might still be considered *practical*: for instance, the absence of laboratory tests and biomarkers may reflect the current *lack of knowledge* about brain functioning. Yet, reducing the difference between psychiatry and general medicine to a practical matter runs the risk of obscuring other important asymmetries. Most importantly, it assumes that psychiatry and general medicine already adopt a common *methodology* when approaching diagnoses.

According to Murphy (2006), this methodology can be summarized in a *medical model* exhibiting two characteristics: 1) The commitment to a view that sees disorders as *breakdowns* in normal processes of various kinds (e.g. biological, cognitive, affective, etc…). 2) The idea that any taxonomy of disorders should be constructed with the goal of uncovering underlying *causes*. In other words: "Diagnosis is causal. [It] is a matter of uncovering the causal antecedents of visible pathology" (p. 324). While this model accurately reflects what happens in most branches of medicine, in psychiatry neither 1) nor 2) are satisfied. With respect to 1), psychiatric classifications tend to characterize disorders in term of *distress* or *disability* but do not rely on normal human capacities that have been damaged or disrupted. Consequently, the recent editions of the DSM do not aim at uncovering malfunctioning mechanisms but rather at describing different forms of deviant behavior. As Kirk, Gomory & Cohen repeatedly stress, the symptoms that are supposed to guide clinicians in the diagnosis often re-state in different ways what the disorder is supposed to be about. The criteria for ADHD are a case in point: the attention-deficit part is spelled out in terms of "difficulty to sustain attention" or "easily distracted", while the hyperactivity part is characterized by actions such as "often leaves seat" or "often on the go" (2013, p. 167). With respect to 2), the DSM rejects any investigation on the causal underpinnings of mental disorders and advocates a *descriptive* approach that attempts to be "neutral with respect to etiology" (DSM-IV-TR, p. xxvi). In short, the rejection of 1) and 2) brings about a classification of mental disorders that neither focuses on the normal processes that are being *disrupted* nor attempts to understand what *causes* the disruption itself.

Psychiatry's disavowal of the medical model seems problematic for at least two reasons. First, it renders impossible to bridge the current gap between psychiatry and general medicine because the two disciplines are endorsing radically different *methodologies*. On one hand, the DSM defends a symptom-based approach based on the description of syndromes and completely divorced from theories or hypothesis about underlying causes. On the other, general medicine operates by constructing models of normal functioning and by grouping illnesses together via causal factors. In this sense, the problem appears more *epistemological* than practical: although our current understanding of the brain's functioning may be limited, the classificatory system in place prevents us from garnering more knowledge about mental disorders. Second, the adoption of a merely descriptive taxonomy creates *paradoxical situations* that become apparent once we re-apply a similar system to general medicine. If diagnoses were based on symptoms only, we would end up grouping together all the patients sharing similar clinical manifestations: "We would classify together everyone who coughs as sufferers from 'cough disorder' and thereby miss the fact that someone who coughs may be doing so for a number of very different reasons" (Murphy 2006, p. 312).

4 FITTING PSYCHIATRY INTO THE MEDICAL MODEL

Murphy's discussion on classification aims at uncovering the fact that psychiatry still remains distant from a full-fledged medical model. Here I expand on his proposal by suggesting a theoretical framework that would facilitate the inclusion of psychiatry within general medicine. In particular, I argue that a characterization of illness able to encompass somatic and mental disorders should be *multi-level*, *holistic* and *dimensional*.

Multi-level. The main barrier that prevents psychiatry from adopting a causal taxonomy consists in the fact that we are still quite ignorant with respect to the etiology of mental disorders. Many authors have highlighted the difficulty to reduce mental disorders to *brain pathologies*: for example, Kendell (1975) describes psychiatric patients as "behaving in ways that alarm of affront other people" and "believing things that other people don't believe" (p. 305). Broome and Bortolotti (2009) stress a similar point: "It does not take an expert to recognize that someone is mentally ill, but how would one decide whether dopamine quantal size, functional MRI activations, or repeats of genetic polymorphism were abnormal in the absence of a disordered person?" (p. 38). These passages point to the fact that – in order to diagnose someone as mentally ill – we often make use of norms that go beyond the somatic sphere to encompass

socio-cultural and epistemic factors. In this sense, most psychiatric explanations would appeal to the disruption of norms on *different levels*: for example, a patient suffering from the Capgras syndrome may present both a neurobiological abnormality (e.g. dopamine dysregulation) and an epistemic one (e.g. abnormal resistance to contrary evidence). Moreover, it would not always be possible to establish the correct level of explanation in advance: whereas for some disorders a fully biological account might suffice (e.g. Huntington's disease), for others we may need to appeal to socio-cultural factors (e.g. anorexia).

A multi-level approach could also be extended to *general medicine*: indeed, somatic illnesses are often the result of a complex array of factors ranging from faulty genes to unhealthy lifestyle. Obvious examples in this sense would be type-2 diabetes or lung cancer, where biological causes interact with environmental ones. Thus, both psychiatry and general medicine could benefit from a multi-level approach to illness. From a diagnostic viewpoint, taking a diverse group of factors into consideration would enhance our understanding of the *causes* behind diseases. For example, the social pressure to resemble women on commercials might matter more than genetic predisposition in the explanation of some eating disorders. Similarly, living in a culture where smoking has a particular social value may put a certain group of people at high risk of developing lung cancer (see Goldade et al. 2012). From a therapeutic viewpoint, a multi-level account allows to abandon a strictly pharmacological approach and to tackle diseases from different perspectives: e.g. cognitive behavioral therapy (CBT) in psychiatry; diet and exercise in general medicine.

Holistic. If somatic and mental diseases are the result of multiple factors and can be understood only by appealing to different levels of explanation, it would be important to explore the dynamics between them. For example, some recent studies have suggested a correlation between schizophrenia and dopamine regulation (see Kapur 2003 and 2004), while others have investigated the high incidence of this disorder within specific sub-groups of the population – e.g. immigrants in conditions of social defeat (see Cantor Graee & Selten 2005). By adopting a multi-level approach we grant that both factors may be useful to explain the onset of schizophrenia: on the biological level, a disrupted process of dopamine release; on the environmental level, risk factors such as migration history or adverse social conditions. Yet, the interaction between the two levels remains unspecified: Does the environmental condition of social defeat directly influence dopamine regulation (*state interpretation*)? Or rather, are the individuals already affected by this brain abnormality more likely to develop schizophrenia (*trait interpretation*)? The endorsement of a *holistic* approach takes advantage of both interpretations without having to consider them mutually exclusive. On one hand there is good evidence that social and cultural habits can shape neurological structures in meaningful ways: for example, taxi drivers appears to exhibit enlarged posterior hippocampal regions with respect to controls who are not experienced in spatial navigation tasks (see Maguire et al. 2000). On the other, chemical imbalances in the brain can affect behavioral manifestations in a variety of ways: the well-known correlation between serotonin levels and depressed mood is just an obvious example.

By adopting a holistic approach, we characterize illness as an emergent phenomenon in which biological and environmental factors are almost invariably influencing one another. More specifically, it may be possible to construct a spectrum indicating the degree of interaction between different kinds of factors in somatic and mental disorders. On one extreme we would find those diseases that emerge almost independently of environmental interaction (e.g. Down syndrome); on the other, those primarily caused by socio-cultural pressures (e.g. bulimia). An interesting consequence of this approach is that the distinction between somatic and mental disorders would somewhat collapse, because the unit of analysis would become the entire organism and its relationship with the environment. This proposal also allows considerable flexibility in classifying a condition as a disease: for example, sickle cell anaemia protects the organism from malaria and thus can be considered an adaptive trait in sub-Saharan Africa, and a serious illness in other environments. In other words, what is functional or dysfunctional cannot be established in a vacuum: "It is difficult to know whether a condition is pathological without considering the environment in which it occurs" (McGuire et al. 1992, p. 93).

Dimensional. According to Murphy, psychiatry can fit a medical model only by endorsing a categorical view of illness, where a condition results from multiple interacting causes but still qualifies as "a distinctive destructive process afflicting a system" (2006, p. 357). A couple of observations can be made in response to Murphy: first – although many illnesses are defined categorically – there are also conditions that arise as a consequence of meeting or exceeding a threshold (e.g. hypertension, diabetes or obesity). These processes are more or less "disruptive" but could hardly qualify as "distinctive": thus, sometimes general medicine treats illness as a condition diverging *quantitatively* – rather than qualitatively – from normal functioning. Second, there is good evidence that many psychiatric symptoms are widespread among the non-clinical population. For example, in a study conducted on 586 college students, 30 to 40% report to have experienced auditory hallucinations at least once in their lifetime, and almost half of these even once a month (see Johns & van Os 2001). Delusions are another interesting example, since they seem to lie on a continuum with other utterly irrational beliefs: thinking that your spouse has been replaced by an impostor does not seem distinctively different from believing that breaking a mirror would bring you seven years of bad luck.

Admittedly, regarding many mental disorders as dimensional would mean drawing the line between pathological and non-pathological with a certain degree of arbitrariness. Yet, it also allows a greater degree of flexibility and the opportunity to evaluate the context on a case-by-case basis. For example, we may want to be conservative in setting the threshold for psychopaths, due to the serious legal and ethical implications often connected to this condition. At the same time, we may decide to pay special attention to "high-risk" situations that need to be monitored or acted upon (e.g. students who regularly

experience auditory hallucinations). This last point seems consistent with what happens in dimensional somatic disorders: for example, if my blood tests report high cholesterol or high sugar level – even within the limits – the doctor may suggest a change in diet or life-style to avoid more problematic consequences. Therefore – despite Murphy's concerns – the endorsement of a dimensional approach sits comfortably with the medical model and promotes a more nuanced view of medical practice. Indeed, it shows that an important part of medicine consists in dealing with *chances* rather than *causes* and that the distinction between pathological and non-pathological may be a matter of degrees (see Gigerenzer 2008).

To sum up, I start by asking whether an analogy between somatic and mental disorders could be meaningfully defended. Then, I appeal to some recent studies on accuracy and cognitive biases to show that the core distinction between psychiatry and general medicine does not rest on the issue of *reliability*. Rather, the *symptom-based* approach currently endorsed in psychiatry is mostly responsible for distancing the discipline from the medical model, creating a gap between the ways in which mental disorders and other illnesses are diagnosed. Finally, I propose a *multi-level*, *holistic* and *dimensional* approach to illness that encompasses somatic and mental disorders.

REFERENCES

[1]Aboraya, A. et al. (2006). The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry (Edgmont) 3*(1): 41-50.

[2]Boorse, C. (1977). Health as a Theoretical Concept. *Philosophy of Science* 44: 542-573.

[3]Boone, C. (1975). On the Distinction between Health and Illness. *Philosophy and Public Affairs* 5: 49-68.

[4]Bortolotti, L. (2009). *Delusions and Other Irrational Beliefs.* Oxford: Oxford University Press.

[5]Broome, M. R., & Bortolotti, L. (2009). Mental Illness as Mental: in Defence of Psychological Realism. *HumanaMente 11*: 25-44.

[6]Brülde, B. (2010). On Defining 'Mental Disorder': Purposes and Conditions of Adequacy. *Theoretical medicine and bioethics 31*(1): 19-33.

[7]Brülde, B. & Radovic, C. (2006). What is Mental about Mental Disorders? *Philosophy, Psychiatry and Psychology 13*: 99-116.

[8]Cantor-Graae, E., & Selten, J. P. (2005). Schizophrenia and migration: a meta-analysis and review. *American Journal of Psychiatry*, 162 (1): 12-24.

[9]Cooper, R. (2004). What is wrong with the DSM? *History of Psychiatry15* (1): 5-25.

[10] Crumlish, N., & Kelly, B. D. (2009). How psychiatrists think. *Advances in psychiatric treatment 15*(1): 72-79.

[11] Cullen, S. W., Nath, S. B., & Marcus, S. C. (2010). Toward understanding errors in inpatient psychiatry: a qualitative inquiry. *Psychiatric quarterly 81* (3): 197-205.

[12] Culver, C. M. & Gert, B. (1982). *Philosophy in Medicine.* New York: Oxford University Press.

[13] De Block, A. (2008). Why Mental Disorders are just Mental Dysfunctions (and nothing more): Some Darwinian Arguments. *Studies in History and Philosophy of Biological and Biomedical Sciences 39*: 338-346.

[14] Foucault, M. (1964). *Madness and Civilization: A History of Insanity in the Age of Reason*. Random House LLC.

[15] Frances, A. (2012). "DSM-5 is a Guide, not a Bible: simply Ignore its 10 Worst Changes." *Huffington Post Science*.

[16] Friedman, C. P. et al. (2005). Do Physicians Know When Their Diagnoses are Correct? *Journal of General Internal Medicine 20* (4): 334-339.

[17] Gigerenzer, G. et al. (2008). Helping Doctors and Patients Make Sense of Health Statistics. *Psychological science in the public interest 8* (2): 53-96.

[18] Goldade, K. et al. (2012). Applying anthropology to eliminate tobacco-related health disparities. *Nicotine & Tobacco Research 14* (6): 631-638.

[19] Graber, M. L. & Berner, E. S. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine 121* (5): S2-S23.

[20] Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic Error in Internal Medicine. *Archives of internal medicine 165* (13): 1493-1499.

[21] Grasso, B. C. et al. (2003). What do we know about medication errors in inpatient psychiatry? *Joint Commission Journal on Quality and Patient Safety 29* (8): 391-400.

[22] Groopman, J. E. (2007). *How doctors think.* Mariner Books.

[23] Johns, L. C., & van Os, J. (2001). The continuity of psychotic experiences in the general population. *Clinical psychology review*, 21 (8): 1125-1141.

[24] Kapur, N. (2000). Evaluating risks. *Advances in Psychiatric Treatment 6* (6): 399-406.

[25] Kapur, S. (2004). Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia. *American Journal of Psychiatry* 160: 13-23.

[26] Kapur, S. (2003). How Antipsychotics Become Anti-'Psychotic' – from Dopamine to Salience to Psychosis. *Trends in Pharmacological Sciences* 25: 402-406.

[27] Kendell, R. (2001). The Distinction between Mental and Physical Illness. *British Journal of Psychiatry 178*: 490-493.

[28] Kendell, R. (1975). The Concept of Disease and its Implications for Psychiatry. *British Journal of Psychiatry 127*: 305-315.

[29] Kirk, S. A., Gomory, T., & Cohen, D. (2013). *Mad science: Psychiatric Coercion, Diagnosis, and Drugs*. Transaction Publishers.

[30] Maguire, E. et al. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Science of the United States of America 97*: 4398-4403.

[31] Mamede, S. et al. (2010). Effect of Availability Bias and Reflective Reasoning on Diagnostic Accuracy among Internal Medicine Residents. *Journal of American Medical Association 304* (11): 1198-1203.

[32] Mamede, S. et al. (2008). Influence of perceived difficulty of cases on physicians' diagnostic reasoning. *Academic Medicine 83* (12): 1210-1216.
[33] McGuire, M. T. et al. (1992). Evolutionary biology: a basic science for psychiatry? *Acta Psychiatrica Scandinavica 86* (2): 89-96.
[34] Murphy, D. (2006). *Psychiatry in the scientific image*. Cambridge, MA: MIT Press.
[35] Phillips, J. (2014). Detecting Diagnostic Error in Psychiatry. *Diagnosis 1*: 75-78.
[36] Szasz, T. (1974). *The myth of mental illness: Foundations of a theory of personal conduct*. Harper-Perennial.
[37] Wakefield, J. C. (1992). The concept of mental disorder: on the boundary between biological facts and social values. *American Psychologist* 47 (3): 373-388.

An encounter between Attachment Theory and 4e Cognition

Dean Petters [1] and **Everett Waters** [2]

Abstract.

A number of research questions arise from an encounter between the elements of 4e cognition and Attachment Theory. These include: (1) whether the Attachment Theory concept of Internal Working Models should be understood in terms of analogue representations more in line with embodied cognition, in addition to traditional cognitivist representations like linguistically mediated narrative measures of attachment meaning?; (2) are infant-carer dyads best thought of as environments of contextual embedding for infant cognition or as an arrangement where the carer can actually extend the infant mind?; and (3) are attachment phenomena best thought of in traditional representational terms or should the attachment control system be reframed in enactive terms where traditional cognitivist representations are: (3i) substituted for sensorimotor skill-focused mediating representations, (3ii) viewed as arising from autopoietic living organism and/or (3iii) mostly composed from the non-contentful mechanisms of basic minds?; A theme that cross-cuts these research questions is how representations for capturing meaning, and structures for adaptive control, are both required to explain the full range of behaviour of interest to Attachment Theory researchers. Implications are considered for future empirical and computational modelling research, and clinical interventions.

1 INTRODUCTION

The infant-caregiver relationship not only plays a central role in social and emotional development, but also in exploration and learning [3, 9, 10]. A traditional cognitivist approach to explaining these phenomena would emphasise internal information processing, located within the individual mind. So this approach in Attachment Theory would focus on what is or should be in the infant's head. A theoretical approach that keeps cognition within the infant is seductive because of its conceptual simplicity and because this approach is more easily implemented in cognitive models that focus on the creation and transformation of internal representations [18, 19, 20]. The elements of 4e cognition - viewing cognition as embodied, embedded, extended, and enacted - all reject or radically reconfigure traditional cognitivism [16]. Whilst the core ideas in Attachment Theory were set out by John Bowlby in a series of papers and books between 1958 and 1982 [2, 3, 5, 6], the elements of 4e cognition are more recently defined [16], but have many earlier conceptual antecedents [8, 11, 29].

How should Attachment Theory respond when viewed through the lense provided by 4e cognition approaches in cognitive science? And which elements of 4e cognition provide the best match for the requirements of a theoretical revision for Attachment Theory?

[1] Birmingham City University, UK, email: dean.petters@bcu.ac.uk
[2] SUNY, Stony Brook, USA.

Concepts from Systems Theory [8] as well as from Developmental Psychology, are key antecedents for contemporary Situated Cognition ([9] p 35). As Clarke notes:

> *"developmental psychologists were probably among the very first to notice the true intimacy of internal and external factors in determining cognitive success and change. In this respect, theorists such as Jean Piaget, James Gibson, Lev Vygotsky, and Jerome Bruner, although differing widely in their approaches, actively anticapted many of the more radical-sounding ideas now being pursued in situated robotics"* ([9] p 35)

The dialogic nature of the infant-mother relationship is exemplified by many types of interaction, including: the infant's active participation in co-operative games, the infant directing the mother's attention to acts by itself, use of objects as topics in infant-mother dialogues, and social and emotional referencing. The mutually contingent nature of these dialogues is demonstrated by experimental studies which perturb the contingency carergiver or infant responses, and in observational research of infant interactivity with depressed mothers [25]. Whilst Bowlby's formulation of Attachment Theory includes cognitivist constructs, like Internal Working Models (IWMs) and hierarchical plans, through which relationship patterns are represented internally, he was also inspired by Systems Theory [3], emphasising that an infant's main caregiver is the most salient part of the infant's environment. So Attachment Theory conceptualises infant-mother relationship as being between two active partners. Therefore, contemporary approaches from situated cognition can form a natural updating for Bowlby's systems approach, and may also help refocus cognitivist elements that Bowlby proposed within Attachment Theory.

The embodied approach views the body and physical world as the context or milieu" for cognition, rather than cognition conceived as the operation of disembodied algorithms [21]. So an encounter between Attachment Theory and embodied cognition asks how attachment representations should be conceptualised, and whether the cognitive component of Attachment Theory could then be *"augmented with the incorporation of bodily sensations, physiological responses, and analogue computations that rely on the physical substrate within the attachment control system"* [21]?

The hypotheses of embedded and extended cognition are competing theories in situated cognition that both give greater emphasis to the role that situations and context play in human cognition than traditional cognitivism. The extended approach is more radical, claiming that external supports become part of a person's cognitive apparatus. The embedded approach is still strongly anti-cognitivist, but sees cognition embedded in external support rather than constituted

of external structures. A key question is: whether attachment relationships can sometimes be conceived as extending cognition or are better thought of as embedding cognition?

The enactivist approach views psychological activity as occurring in the dynamic engagement between organisms and their physical and social context rather than within themselves [15]. The mind and subjective experience are not seen as inherent in, or arising from, the individual, but as emerging, from the interaction between organisms and their surroundings [15]. So another key question include: is enaction, rather than traditional forms of representation, a better way to think about how previous experiences mediates ongoing adaptive behaviour, and can the attachment control system be revised to act as an enactive "lived experiential structure" ([21, 28] p xvi)?

The intention in challenging Attachment Theory with recent ideas from 4e cognition is to revise rather than replace or reject it, and also see which diverse elements from 4e cognition can operate in 'joint purpose', motivating a progressive revision of a well established theory.

1.1 Bowlby formulated the attachment control system concept but did not specify it in detail

John Bowlby rejected psychoanalytic theory as a basis for explaining social and emotional development. Instead, he formulated a new explanatory framework by combining scientifically respectable ideas that originated across different disciplines. In his first presentation of Attachment Theory, in 1958, Bowlby provided an alternative motivational basis for attachment by replacing a psychoanalytic explanation based on Freudian instincts with a motivation framework based on ethological behaviours. Whilst this framework was too simple to explain different stages in attachment development it could be augmented further with other scientific concepts. In 1969, in the first volume of his Attachment Trilogy, Bowlby's theoretical revisionism included a much broader range of currently popular concepts, bound together in the attachment control system framework. So Bowlby's concept of an attachment control system replaced Freuds concept of psychical energy and its discharge ([3], p 18) and wove together constructs from: Piagetian theory; Cybernetics; Artificial Intelligence; and Systems Theory. He presented reflex behaviours and behavioral chaining of fixed action patterns as an example of a simple organizing principle for control systems, and hierarchical planning as much more complex and flexible ([3], p 76). Internal Working Models (IWMs) and natural language allowed higher level processes of integration and control. Then in the second and third parts of the attachment trilogy Bowlby invoked concepts from cognitive psychology. For example, he explained Freudian defensive processes in terms of selective attention ([6], chapter 4), and explained recall, reflection and potential internal conflict in self image in terms of the distinction between episodic and semantic memory ([6], p 61-64). Figure 1 shows illustrates how the 'theoretical borrowings' that Bowlby made changed with what were the prominent ideas of the day.

However, in none of his descriptions of the attachment control system did Bowlby set-down precise enough arrangements for how varied information processing elements might be organised in a running simulation. This is not a surprise, at the time that Bowlby formulated Attachment Theory, there existed no simulation technology to combine information processing elements such as ethological behaviours, IWMs and hierarchical plans within a single information processing architecture.

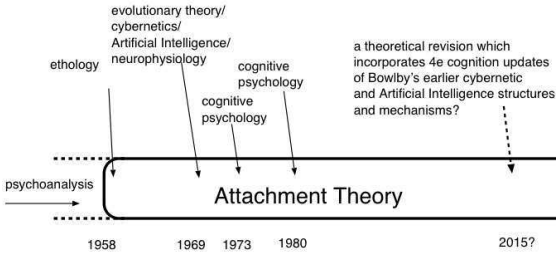


Figure 1. Diagram showing influences from other disciplines on Attachment Theory over time.

2 SHOULD INTERNAL WORKING MODELS BE VIEWABLE AS ANALOGUE IN ADDITION TO SYMBOLIC REPRESENTATIONS?

Internal Working Models are described by Bowlby as higher level representational forms which integrate and exert control over lower level control systems. Their principal information processing function is to allow predictions to be made about the likely outcomes of taking actions within a given environment. IWMs transmit, store and manipulate information and allow the individual to *"conduct small scale experiments within the head"* (([3], p 81). Their function, in terms of Bowlbys agenda of reforming psychoanalytic theory, was to take the place of the internal worlds of traditional psychoanalytic theory. Bowlby emphasizes the requirements for Internal Working Models to be updated. He also briefly observes that pathological sequelae of separation and bereavement can be understood in terms of out of date models or half revised models which may contain consistencies and confusions (Bowlby 1969 page 82). Bowlby invokes Internal Working Models at early stages in development also later on, when linguistic skills and conscious reflection can enable models to become more adequate ([3], p 84).

In contemporary reviews, IWMs are presented as transforming from sensorimotor representations in pre-linguistic infants to manipulable internal simulations in older children and adults that can enable short-term predictions, and conscious reflections on past, ongoing and future relationships ([7], p 102). Current research investigates IWMs through studies of memory talk, narrative completion, semi-projective measures and story-telling, with adults and children [7] - naturally linking IWMS to symbolic constructs from Artificial Intelligence like schemas and scripts. In his later writing Bowlby described IWMs in symbolic terms, for example:

"In reaching the decision to utilise certain actions rather than others the attachment system is conceived as drawing on the symbolic representations or working models, of the attachment figure, the general environment and the self, which are already stored and available to the system" ([4], p. 373).

However, links have also been drawn with IWMs and recent neuroscience research based upon mirror-neurons which presents IWMs as affording embodied simulation of the intentions of others ([7], p 109). Though research viewing IWMs as embodied simulations is very much in the minority in contemporary attachment research on IWMs ([7, 24], this section will argue that it is not only fully in the 'spirit' of Bowlby's original conception for IWMs, but also matches the 'word' of what he wrote about IWMs when he first introduced them. Bowlby did not use the term 'embodied simulation' but he did compare IWMs to analogue representations. For example, in his 1969 formulation of IWMs, Bowlby suggests that they can be used to

conduct 'small-scale experiments within the head' and notes that this notion would be an obvious possibility to electrical engineers familiar with analogue computers. Bowlby also refers to how anti-aircraft guns operate ([3], p 44) to exemplify how analogue control systems can set their own goals.

Looking back prior to 1969 to Bowlbys sources for the IWM concept provides added detail on how analogue representations can be conceived as mental models. Bowlby adopted the concept of Internal Working Models from the biologist J.Z. Young [31], whose treatment of Working Models is decidedly unambiguous in its preference for analogue over digital representations as a basis for Working Models in natural systems. As Young noted:

> "[In an analogue computer] *the pattern of connections that determines what computation is made is part of the structure or pattern of the machine. These features at once suggest to the biologist, and especially the anatomist, that the nervous system is likely to work at least in part on analogue principles. What we commonly call the structure of the nervous system determines what it does. It is not a general purpose computer at all, but consists of a number of analogues set up to perform a few particular tasks. [] One of the great advantages of an analogue machine is that it can receive information directly from particular environments. That is to say, the machine maybe itself a representation of the environment and its parts are pre-selected to perform certain calculations in relations to the latter."* ([31], p 39)

J.Z. Young acquired the working model concept from its original source - the cybernetician Kenneth Craik. In The Nature of Explanation [12], Craik first discussed how working models can be used in science. Physical systems can act as models which help scientists explain natural phenomena because their physical operation captures key aspects of how the target system operates:

> *"By a model we thus mean any physical or chemical system which has a similar relation-structure to that of the processes it imitates. By 'relation-structure' I do not mean some obscure non-physical entity which attends the model, but the fact that it is a physical working model which works in the same way as the process it parallels, in the aspects under consideration at any moment. Thus, the model need not resemble the real object pictorially; Kelvin's tide-predictor, which consists of a number of pulleys on levers, does not resemble a tide in appearance, but it works in the same way in certain essential respects"* ([12], p 51)

So in Craik's working models, although these systems can be argued to represent reality, when used by scientists to enable them to better explain and predict natural phenomena, it is by their physical properties rather than with abstract or arbitrary symbols that they represent other systems. Craik then made the significant leap to suggest that organisms can hold within their minds working models which operate in the same way. So living organisms can possess working models which represent their self and environment, and can run forward in time to make predictions or imagine the results of differing actions. Working models can also be configured to act as memories of past events.

The distinction between analogue and symbolic (discrete and digital) representations is important because analogue representations are much less flexible and are tied to the physical (embodied) properties of the medium in which they are implemented. Analogue systems carry out computational operations using continuously varying data. Data in analogue devices is also transferred around these machines from input to output in continuous form and is bound to the physical form of the computational medium. So analogue computation relies on a physical or embodied substrate in a manner in which discrete symbol processing computations do not. These distinctions certainly matter to the growing number of researchers engaged in computational modelling of attachment behaviour, who actually want to implement running simulations of the attachent control sytems. In addition, how IWMs represent self and environment will also be of interest to clinicians who are concerned to activate, de-activate or transform attachment representations as part of therapy.

That Bowlby would invoke analogue computation and representations in his first formulation of IWMs might seem surprising given the contemporary predominance of the linguistic/symbolic approach to IWMs in Attachment Theory. It is in part explained by the waining popularity of analogue computers. In the period between the end of the second world war and the late 1960s when Bowlbys initially adoption of the working models concept, analogue computing remained a significant alternative to digital computing and the rise and domination of digital computing in the post-war years was not viewed as a foregone conclusion [27]. In addition, the seeming change in emphasis from analogue representations in 1969 to symbolic in 1982 may not represent a completely radical change in Bowlby's conceptualisation because Bowlby was vague in the representational details he proposed. As Bretherton and Mulholland note, Bowlby's formulation of the representational basis for attachment *"was a promising conceptual framework to be filled in by others"* ([7], p 103). However, perhaps the key issue was that in the 1960s Artificial Intelligence was less prominent in comparison with Cybernetics than it would be in the future. So the cybernetic view on issues like meaning and control held greater sway. This was consequential because researchers in Cybernetics under-emphasized representational distinctions and the challenges arising from consideration of high level processes. As Boden notes:

> *"most cyberneticians seemed to see no difference between pure self-equilibration (as in homeostasis), purposive behaviour directed to some observable object (as in guided missiles), and goal seeking directed to some intentional end (as in human deliberation and planning)*([1], p 220)

The eclipse of Cybernetics by Artificial Intelligence may have led to Bowlby's switch from invoking an analogue basis for IWM in 1969 to symbolic basis for IWMs in 1982. More recent developments have shown movement towards an integrative approach which might guide the process of bringing diverse representational forms together in the attachment control system, bringing back together a cybernetic approach to adaptive control and an Artificial Intelligence approach to fully intentional thought and reasoning [24, 22].

3 ARE INFANT-CARER DYADS BEST DESCRIBED IN TERMS OF COGNITIVE EMBEDDING OR COGNITIVE EXTENSION?

The idea that infants, older children and even adult attachment partners all look to their carers as information sources about the broader world is a familiar one. For example, from the perspective of the socially situated mind, infant social referencing and joint attention between infant and carer may be seen as physical actions that make the infant's mental computations faster, more reliable or less effortful

by intimately linking internal infant cognition with external support [23]. So taking a situated cognition approach enrichs attachment theory by providing a more complete view of how infants gain information about environments from their caregivers.

Caregivers provide support to infant cognition in very many ways. They help to label, conceptualise, and structure information ([23, 10], p 44). Caregivers and infants are also situated within some of the same action loops that criss-cross close-coupled individuals and the environment [30, 9]. In such systems, caregivers can support 'soft assembly' of developing attachment competencies because secure attachment patterns are described in terms of response to set-goals rather than set actions ([9], p 44). Caregivers help scaffolding infant development by directing the child toward a correct/established outcome/solution/attitude or belief. When co-constructing they help the child take a course toward own-defined ends or end points. In addition, Bowlby describes how caregivers support infants by manipulating the environment and providing information directly through language use so that *"instead of each one of us having to build his environmental and organismic models entirely for himself, he can draw on models built by others"* ([3], p 82).

The hypotheses of embedded and extended cognition are competing and mutually exclusive explanations for how caregivers provide cognitive support. The hypothesis of extended cognition suggests that in some of the above examples, if the infant's ongoing computational needs are met by sensitive and timely support from his or her carer in such a way that the infant treats this support as part of their own cognitive processes then we might say that the carers cognitive support has become part of the infant's extended mind. For these examples to count as mind extension, caregiver cognitive support and information provision to the infant must be strongly trusted, relied upon and accessible. If these criteria are met then what is occuring is extension of mental states from an infant onto their caregiver. So in this view, the carer is actually extending the infant mind by incorporating the carer's help within the infant's cognitive operations - the carer's help becomes part of the infant's mind [3]. For these same examples of intimately integrated interactions between infant cognition and carer support, the hypothesis of embedded cognition views infant cognition and carer support of that cognition as clearly demarcated and separate. This hypothesis considers that *"cognitive processes depend very heavily, in hitherto unexpected ways, on organismically external props and on the structure of the external environment in which cognition takes place"* ([26] p 393). and that *"certain cognitive processes lean heavily on environmental structures and scaffoldings but not thereby include those structures and scaffoldings themselves"*([10], p 111).

We should be more accepting of claims to extended cognition in infants and younger children, because the caregiver's interactions are more long-lasting, they are relied upon more, and when there are less infant cognitive resources and routines for not believing [13]. So making acceptance of information from the carer as if it were an infant's own beliefs easier and more likely.

Two main reasons for preferring embedded explanations over extended explanations arise from considering non-social cognitive extension [10]. Most examples of extended cognition involve inorganic objects in the environment (such as a mathematician doing their 'working' on paper) providing the cognitive extension. The first criticism of extended cognition highlights the profound differences that appear to distinguish inner and outer contributions in extended cognition when cognition is extended onto such inorganic objects [10]. However, this criticism is much weaker when applied to the social case as it is a carer that does the extending. So there are not such profound differences in the supporting substrate for cognition between cognition inside the infant's brain and cognitive support originating from inside the carer's brain. A second criticism is the apparent scientific cost of any wholesale endorsement of extended cognition onto a motley collection of inorganic objects because it gives undue attention to transient external props and aids. In this view, following the extended mind hypothesis means scientists are not researching a suite of integrated persisting organismically grounded capacities [10, 30], and looking at developmental examples of cognitive extension onto inorganic objects is a series of separated developmental segments with external cognition onto different objects. So using a ball or balance beam may be a good example of mind extension at one age, but a year later the best example may involve a completely different object in a different task or action. Again, the social case of mind extension mitigates this criticism. Extended cognition does not only deal with transient external props and aids when the carer provides enduring support and continuity between otherwise disparate contexts.

If we accept the hypothesis of extended cognition over the hypothesis of embedded cognition this has important implications for computational modelling and in clinical interventions. Caregiving relationships are often very durable and reliable and if socially extended cognition occurs we can expect typical interactions and development to include micro and macro instances. Micro extension effects are described by Clark: *"The child is surrounded by exemplars of mind-reading in action, she is nudged by cultural interventions such as the use of simplified narratives, prompted by parental rehearsal of her own intentions, and provided with a rich palate of linguistic tools such as words for mental states"* ([10], p67). Macro effects occurs when children absorb complex ideas wholesale through the conduit of cognitive extension. Their caregivers can simply present beliefs which the children then adopt. Over the long-term caregivers attempt to socialise and indoctrinate infants in many ways that will impact the developing meaning a child gains of their attachment history. Two types of problems can occur: (1) relationships are not reliable or durable enough so infants and children do not gain the benefits of cognitive extension; and (2) pathological extension occurs, so instead of acting to scaffold or co-construct, a caregiver uses their power to extend an infant's mind to introduce (or put more strongly 'infiltrate' or 'hack' [17]) unhealthy or pathological beliefs about the infant's self and relationships into the infant's mind.

4 ENACTIVISING ATTACHMENT THEORY

Where the extended/embedded question highlighted the requirement for attachment structures and mechanisms that support narrative meaning making the three flavours of enactivism highlight different aspects of adaptive control and subjective experience in the attachment domain.

4.1 Attachment Theory encounters Sensorimotor Enactivism

Sensorimotor enactivism criticises the view that perception results in inner images or mental representations being produced. In the sensorimotor view, perception, action, and subjective perceptual experiences are all inescapably connected [14]. This approach allows that perceptual experience is grounded in knowledge and is therefore

[3] [23] presents a more detailed case that the infant carer dyad is an examplar of extended mind cognition, with the infant's cognition extended by their caregiver.

representationally contentful. But the kind of mediating knowledge in sensorimotor enactivist accounts is more like procedural or skill-based knowledge. It is 'know-how' rather than 'know-that', a kind of knowledge demonstrated by the skilled performance of its deployment rather than an independently queriable knowledge base [14].

Viewing attachent behavioural patterns in this enactivist manner - as social skills rather than arising as a result of internal representations - may provide a powerful spur towards new research hypotheses and clinical interventions. When individuals with insecure attachment gain secure status they can be viewed as gaining a skill which they can then use in other relationships.

4.2 Autopoeisis and representation from social interaction

According to autopoietic enactivism, cognition, mentality and subjective experience all emerge from the self-organising and self-creating activities of autonomous entities [14]. This activity is intimately spread between organism and environment. Enactivisits suggest that, because factors from 'within' and 'without' play equally important and necessary roles in creating cognition and behaviour, the distinction between organism and environment is viewed as only having a heuristic value rather than being a true metaphysical division [14].

Autopoiesis is a special case of homeostasis and it takes the position that metabolism and life is essential for grounding intentional categories like cognition, consciousness, and emotions [1]. In the second Volume of the Attachment Trilogy, Bowlby adopted the biological concept of homeostatis and applied it to behavioral as well as physiological control systems. In this view, physiological homeostasis which regulates food and sleep are an inner ring of control in the attachent control system. Attachment behavioural patterns constitutes an outer behavioral ring which is a complement to this inner physiological control system (Bowlby 1973, chapter 9). However, Bowlby did not set out how the intimate engagement of these two rings could give rise to phenomenoligical experience. He did describe attachment feelings, but within an emotional appraisal framework ([3], chapter 7). So viewing Attachment Theory through the lense of autopoietic enactivism can act as a spur for a more comprehensive approach that unifies behaviour, cognition, and subjective experience in a single explanatory framework.

4.3 A Radical Enactivist Manifesto for Attachment Theory?

Hutto and Myin propose the thesis of radical enactive cognition (REC) that is a variant of enactivism that states that only a small proportion of cognitive processing is mediated by contentful representations. In their view, the majority of human cognition is basic and non-contentful information processing that controls behaviour for adaptive purposes but does not possess truth bearing properties like reference, accuracy or implication. According to REC, contentful representations do mediate some cognition, but these representations play a minor role in cognition overall, *"emerging late in phylogeny and ontogeny, being dependent in special sorts of shared practices.* ([14], p 13). So what Hutto and Myin have proposed is a novel variant of a dual process approach to cognition, with linguistically mediated representations that can interpret or receive narrative meanings, and basic structures and mechanisms that carry out adaptive control [22]. However, whilst other dual process approaches make a distinction between self-reflective thought which is linguistically mediated and conscious, and processing which is not linguistically mediated and inaccessible to consciousness, REC 'carves things up' in a very different way [22]. As Hutto and Myin note, *"Enactivists are concerned to defend the view that our most elementary ways of engaging with the world and others - including our basic forms of perception and perceptual experience - are mindful in the sense of being phenomenally charged and intentionally directed, despite being non-representational and content-free* ([14], p 13). So according to a REC approach to Attachment Theory, an IWM that is formed early in ontogeny and has become inaccessible to linguistic self-reflection is not 'hidden', or at 'behind' or 'beneath' other more linguistically accessible IWMs. Instead, REC reframes inaccessibility - so in REC this is just linguistic inaccessibility - so such inaccessible structures are still at the forefront of mind and are phenomenally charged and conscious. This reframing can turn therapeutic ideas right around. Instead of therapy uncovering hidden structures it is about understanding how context and behavioural predispositions enact these structures in the moments they occur.

In addition, REC holds that an organism's current behavioural tendencies are not explained or structured by representations of the past but influenced more directly, just by its *"history of active engagement."* with the world ([14], p 11-12). So an organism's behavioural predispositions do *"not inherently "say" anything about how things stand in the world"* ([14], p 19). Rather, according to Hutto and Myin, *"a truly radical enactivism - REC - holds that it is possible to explain a creature's capacity to perceive, keep track of, and act appropriately with respect to some object or property without positing internal structures that function to represent, refer to, or stand for the object or property in question"* ([14], p 82)

So if Attachment Theory follows REC it might reconceive internal states like working models to be just control states and break the link with the reality they are supposed to represent. An attachment control system that proposes internal control states are not truthful representations of reality is a profound shift from current Attachment Theory. No longer would attachment interventions be concerned to assess how individuals represented their past relationships but instead they would be more focused on how to move towards more adaptive behaviour patterns.

5 Conclusion

In breaking from psychoanalysis Bowlby was a revolutionary, but at heart he was also a conservative, because he wanted to save the core and most valuable findings of Freud's psychoanalytic framework. These were insights about the highly active and interactive nature of social and emotional development in infancy. Since Bowlby was an eager 'borrower' of scientific concepts from the ideas which were popular at the time he formulated Attachment Theory, he might today look to incorporate the diverse insights of 4e cognition in a revised framework for the attachment control system. In section 2 we asked whether IWMs in adults are linked both to processes of shared meaning making and interpretation, and to processes of adaptive control, that is, whether they should not only be conceived in linguistic or symbolic form, but also conceived as analogue or embodied information processing structures [24]. In section 3 we showed how extended cognition provides a possible explanation for how infants derive narrative meaning about their attachment relationships from their caregivers. Then in in section 4 we considered how an enactivist approach can help explain subjective experiences in attachment interactions, and how internal control structures can direct future actions without a link to 'truthful' representations of past events. Con-

sidering issues of embodiment, cognitive extension, and enactivism together has a major benefit because these three approaches pull in different directions. So together they provide a balanced reformulation. Considering IWMs as analogue in addition to symbolic keeps the IWM construct tied to an individual. The extended cognition approach reminds us of the dialogic nature of attachment and the enactive approach forces us to question our representational assumptions. Taken together these three perspectives complement each other. We can never really know how Bowlby would have responded to the questions posed by 4e cognition but we can act to make revisions to Attachment Theory that conserve his key theoretical insights.

REFERENCES

[1] M.A. Boden, *Mind as Machine: A History of Cognitive Science*, Oxford University Press, Oxford, 2006.

[2] J. Bowlby, 'The nature of a child's tie to his mother', *International Journal of Psychoanalysis*, **39**, 350–373, (1958).

[3] J. Bowlby, *Attachment and loss: volume 1 attachment*, Basic books, New York, 1969.

[4] J. Bowlby, *Attachment and loss: volume 1 attachment*, Basic books, New York, 1969 | 1982. (Second edition 1982).

[5] J. Bowlby, *Attachment and loss: volume 2, Separation: Anxiety and Anger*, Basic books, New York, 1973.

[6] J. Bowlby, *Attachment and loss: volume 3 loss, sadness and depression*, Basic books, New York, 1980.

[7] I. Bretherton and K.A. Munholland, 'Internal working models in attachment relationships', in Handbook of Attachment, (Second edition *, eds. J. Cassidy & P.R. Shaver*, 102–127, Guilford Press, London, (2008).

[8] W.J. Clancy, 'Scientific antecedents of situated cognition', in Cambridge Handbook of Situated Cognition*, eds. P. Robbins & M. Aydede*, 11–34, Cambridge University Press, New York, (2008).

[9] A. Clark, *Being There: Putting Brain, Body and World Together Again*, MIT Press, Boston, 1998.

[10] A. Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford University Press,, New York, 2008.

[11] A. Clark and D. Chalmers, 'The extended mind', *Analysis*, **58**, 10–23, (1998).

[12] K. Craik, *The Nature of Explanation*, Cambridge University Press, London, New York, 1943.

[13] D.T. Gilbert, ' How mental systems believe', *American Psychologist*, **46**, 107–119, (1991).

[14] D. Hutto and M. Myin, *Radicalizing Enactivism: Basic Minds without Content*, MIT Press, Cambridge, MA, 2013.

[15] M. McGann, H. De Jaegher, and E. Di Paulo, 'Enaction and psychology', *Review of General Psychology*, **17**, 203–209, (2013).

[16] R. Menary, 'Introduction to the special issue on 4e cognition', *Phenomenological Cognitive Science*, **9**, 459–463, (2010).

[17] P. Paulconbridge, 'Hacking the Extended Mind', in *Proceedings of 'Re-conceptualizing Mental Illness: The View From Enactivist Philosophy and Cognitive Science', AISB Convention 2013*, 34–36, AISB Press, University of Sussex, Brighton, (2013).

[18] D. Petters, 'Simulating infant-carer relationship dynamics', in *Proc AAAI Spring Symposium 2004: Architectures for Modeling Emotion - Cross-Disciplinary Foundations*, number SS-04-02 in AAAI Technical reports, pp. 114–122, Menlo Park, CA, (2004).

[19] D. Petters, 'Building agents to understand infant attachment behaviour', in *Proceedings of Modelling Natural Action Selection*, eds., J.J. Bryson, T.J. Prescott, and A.K. Seth, 158–165, AISB Press, School of Science and Technology, University of Sussex, Brighton, (2005).

[20] D. Petters, 'Implementing a theory of attachment: A simulation of the strange situation with autonomous agents', in *Proceedings of the Seventh International Conference on Cognitive Modelling*, 226–231, Edizioni Golardiche, Trieste, (2006).

[21] D. Petters, ' Towards an Enactivist Approach to Social and Emotional Attachment.', in *ABSTRACTS. AISB50. The 50th annual convention of the AISB. Goldsmiths University of London.*, 70–71, AISB, Goldsmiths College, London, (2014).

[22] D. Petters and E. Waters, 'A.I., Attachment Theory, and Simulating Secure Base Behaviour: Dr. Bowlby meet the Reverend Bayes', in *Proceedings of the International Symposium on 'AI-Inspired Biology', AISB Convention 2010*, 51–58, AISB Press, University of Sussex, Brighton, (2010).

[23] D. Petters and E. Waters, 'Epistemic Actions in Attachment Relationships and the Origin of the Socially Extended Mind', in *Proceedings of 'Re-conceptualizing Mental Illness: The View From Enactivist Philosophy and Cognitive Science', AISB Convention 2013*, 17–23, AISB Press, University of Sussex, Brighton, (2013).

[24] D. Petters and E. Waters, 'From Internal Working Models to Embodied Working Models', in *Proceedings of 'Re-conceptualizing Mental Illness: Enactivist Philosophy and Cognitive Science - An Ongoing Debate', AISB Convention 2014*, AISB, Goldsmiths College, London, (2014).

[25] V. Reddy, D. Hay, L. Murray, and C. Trevarthen, 'Communication in infancy: Mutual regulation of affect and attention', in Infant development: recent advances*, eds. J.G. Bremner, A. Slater & G. Butterworth*, 247–273, Psychology Press, Hove, (1997).

[26] R. Rupert, 'Challenges to the hypothesis of extended cognition', *Journal of Philosophy*, **8**, 389–428, (2004).

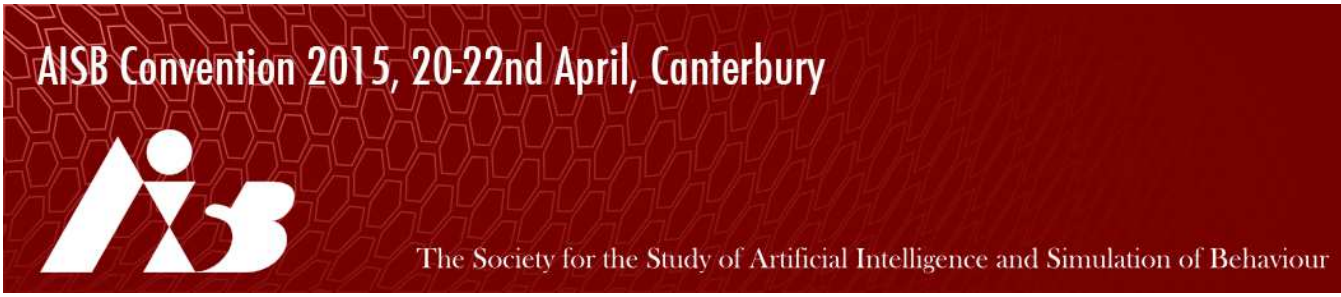
[27] J. Small, *The Analogue Alternative: The Electronic Analogue Computer in Britain and the USA, 1930-1975*, Routledge, London, 2000.

[28] E. Thompson, *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, MIT Press,, Cambridge, Mass, 2007.

[29] F. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience.*, MIT Press,, Cambridge, Mass, 1991.

[30] R.A. Wilson and A. Clark, 'How to situate cognition: Letting nature take its course', in Cambridge Handbook of Situated Cognition*, eds. P. Robbins & M. Aydede*, 55–77, Cambridge University Press, New York, (2008).

[31] J.Z. Young, *A Model of the Brain*, Oxford University Press, London, 1943.



AISB Convention 2015
University of Kent, Canterbury

Proceedings of the AISB 2015 Symposium on Social Aspects of Cognition and Computing

Edited by Gordana Dodig-Crnkovic, Yasemin J. Erden and Raffaela Giovagnoli

ISBN: 978-1-908187-53-6

Introduction to the Convention

The AISB Convention 2015—the latest in a series of events that have been happening since 1964—was held at the University of Kent, Canterbury, UK in April 2015. Over 120 delegates attended and enjoyed three days of interesting talks and discussions covering a wide range of topics across artificial intelligence and the simulation of behaviour. This proceedings volume contains the papers from the *Symposium on Social Aspects of Cognition and Computing*, one of eight symposia held as part of the conference. Many thanks to the convention organisers, the AISB committee, convention delegates, and the many Kent staff and students whose hard work went into making this event a success.

—Colin Johnson, Convention Chair

Introduction to the Symposium

This Symposium falls into the relatively new area of the intersection of computer science and social sciences. Known as social computing, this intersection has far reaching consequences for many fields including AI and philosophy. In order to have a fruitful discussion we intend social computing in a broad sense to explore different levels of social behavior in computational systems, both natural and artifactual. The following topics are considered:

I. Social computing in relation to cognitive computing and affective computing;

II. Strategies for analyzing the problem of representation from a philosophical perspective that implies the comparison between human and machine capacities and skills;

III. The relations between knowledge and categorization, and the promotion of communication among experts and users;

IV. Social computing and online relationships;

V. The rise of social computing and ethical issues.

Danielle MacBeth discusses the problem of mathematical logic and mechanical reasoning, which have turned out to be largely irrelevant to the practice of mathematics, and to our philosophical understanding of the nature of that practice. Her aim is to understand how this can be. We will see that the problem is not merely that the logician formalizes. Nor even is it, as Poincaré argues, that logicians replace all distinctively mathematical steps of reasoning with strictly logical ones. Instead, as will be shown by way of a variety of examples, the problem lies in the way the symbolic language of mathematical logic has been read. Rodger Kibble explores the idea that human cognition essentially involves symbolic reasoning and the manipulation of representations, which is central to cognitivistic approaches to AI and cognitive sciences. The very idea of representation has been problematized by philosophers such as Davidson, McDowell and Rorty. Along this line, the paper discusses Robert Brandom's thesis that the representational function of language is a derivative outcome of social practices rather than a primary factor in mentation and communication. The philosophical approach of Analytic Pragmatism (introduced by Robert Brandom) is at the center of Raffaela Giovagnoli's contribution. It represents a fruitful point of view to isolate what capacities and abilities are common to human and nonhuman and what capacities and abilities are typical of human beings. They give rise to different sorts of autonomous discursive practices (ADPs) which offer a new conception of AI and open interesting spaces for new forms of computation. One fundamental issue in social computing is the question of "digital identity" analyzed by Yasemin J. Erden.Identity is neither simple nor static, and in many ways the multiplicity of identity that this paper will consider is not in itself either novel or controversial. Our everyday roles and experiences contribute to the complex nature of our identity, and we are both defined by (and define ourselves according to) the actions, choices, beliefs and emotions that we either choose or deny. In these respects it seems likely that what we might call a digital identity would merely add to the multiplicity of our otherwise complex picture of ourselves. Colette Faucher moves from the observation that in modern asymmetric military conflicts the Armed Forces generally have to intervene in countries where the internal piece is in danger. They must make the local population an ally in order to be able to deploy the necessary military actions with its support. The paper focuses on the Intergroup Emotion Theory that determines from characteristics of the conveyed message the emotions likely to be triggered on info-targets.

It also simulates the propagation of the message on indirect info-targets that are connected to direct info-targets through the social networks that structure the population. Gaurav Misra and Jose Such notice that social computing revolutionized interpersonal communication. However, the major Online Social Networks (OSNs) have been found falling short of appropriately accommodating their relationships in their privacy controls, which leads to undesirable consequences for the users. The authors highlight some of the shortcomings of the OSNs with respect to their handlings of social relationships and present challenges to promote truly social experience. Another very interesting topic is related to the theory of social action. Leon Homeyer and Giacomo Lini concentrate on behaviourism and materialism in AI and agency in

general. They analyze a specific utility-based agent, the ps model presented first in (Briegel and De Las Cuevas 2012) which has in its capability to perform projections its key feature. This analysis allow the authors to present a feature-driven concept of agency that allows a comparison of different agents which is richer than solely behaviouristic or materialistic approaches in virtue of the shift from a theory-driven stance to a process- driven one. Giles Oatley, Tom Crick and Mohamed Mostafa introduce the goal of their long-term research on the development of complex (and adaptive) behavioural modeling and profiling a multitude of online datasets. They look at suitable tools for use in big social data, on how to "envisage" this complex information. They present a novel way of representing personality traits (using the Five Factor Model) with behavioural features (fantasy and profanity).

Searching for the fundamental mechanisms of rationality of social behaviour, Andrew Schumann offers an analysis of a remarkable organism, cellular slime mould which spends parts of its life as unicellular eukaryotic microorganism, but under specific circumstances of scarcity of food, it communicates chemical signals among its cells, and they gather into a cluster that acts as one single social organism. The interesting phenomenon discussed by Schumann is the behaviour of Physarum polycephalum as the individual- collective duality.

Another kind of duality, that Daniel Kahneman characterizes as fast vs. slow thinking is in focus of David C. Moffat's contribution. The author argues that the essential difference between the two is that the emotions (fast thinking) are unplanned and that rational/slow thinking requires planning. Immediateness of emotive response brings unpredictability, which is considered irrational. The priority of the emotional thinking comes as a result of it preceding the other cognitive processes.

The third dual aspect approach is taken by Judith Simon based on individual human agents perspective and the societal one used in political decision-making with regard to emerging big data. The governance of big data require, as Simon aptly emphasizes, taking into account not only political but equally importantly epistemological and ethical and aspects and preventing widespread and unjustified "trust in numbers".

Alexander Almér, Gordana Dodig-Crnkovic and Rickard von Haugwitz describe collective cognition as distributed information processing, taking the view that all living organisms posses certain level of cognition, the idea first proposed by Humberto Maturana and Francisco Varela. Authors argue, looking at social networks from bacteria to humans that social cognition brings new emergent properties that cannot be found on the individual level. Information processing range from transduction of chemical signals such as "quorum sensing" in bacteria, simple local rules of behaviour that insects follow leading to "swarm intelligence", up to human-level cognition based on human languages and other communication means.

In the search for distributed computational intelligence, Joseph Corneli and Ewen Maclean focus on computational blending that represents distributed development of ideas in social settings, which they modeled by cellular automata. Authors define and explore by simulation a large-scale system dynamics that emerges driven by local behavior, where local rules, unlike in standard cellular automata, are adaptive. This research anticipates a future computational search for rules that may lead to "intelligent" behavior of a distributed computational system.

One of the interesting questions is the character of social coordination. Taking cognitive agents to be humans, Tom Froese presents the enactive theory of social cognition describing the steps from theory to experiment. In the enactive approach to social cognition, which is the recent variety of embodied and extended theories of social cognition, it is possible to make specific predictions of behavior that can be experimentally evaluated. Understanding another person is studied as primarily as a direct perceptual interactive engagement. A second-person perspective is seen as co-constituted by the mutual coordination of bodily interactions. Preliminary results of this study show the social awareness increase over time, notwithstanding the lack of explicit feedback about task performance.

With thanks to all our authors for their contributions, we are convinced that our symposium provides a valuable contribution to the understanding of social aspects of cognition and its relation to computing.

—Raffaela Giovagnoli and Gordana Dodig-Crnkovic

Contents

The enactive theory of social cognition: From theory to experiment

Tom Froese[1, 2]

Abstract. For over a decade I have been working on applying an evolutionary robotics approach to gain a better understanding of the dynamics of social interaction. At the same time I have been developing the enactive theory of social cognition by drawing on the phenomenological philosophy of intersubjectivity. Recently I was able to test the predictions deriving from this research on the basis of a psychological experiment using a new variation of the perceptual crossing paradigm. The empirical results support a genuinely enactive conception of social cognition as primarily grounded in embodied intersubjectivity.

EXTENDED ABSTRACT

I argue that the enactive approach to social cognition is the most promising contender among the recent variety of embodied and extended theories of social cognition. It has the virtue of making specific predictions that can be evaluated experimentally.

The upshot of this theory is that the process of understanding another person is best studied as primarily consisting of a direct perceptual interactive engagement, whereby this genuinely second-person perspective is co-constituted by the skillful mutual coordination of bodily interaction.

There are many theoretical reasons for accepting this position, and a series of agent-based models of embodied interaction show that a dynamically extended embodiment spanning two agents is possible in principle [1,2]. In fact, the mathematics of nonlinear interactions leads us to expect that such mutual incorporation should be found empirically.

We studied this hypothesis using the perceptual crossing paradigm, in which the embodied interaction of pairs of adults is mediated by a minimalist virtual reality interface [3]. As predicted, movements became entrained during their interaction, and there was a positive correlation between objective measures of coordination and subjective reports of clearer awareness of each other's presence. Intriguingly, there was a tendency for coordinating participants to report independently yet within seconds of each other that they had become aware of the other, suggesting the emergence of a genuinely shared experience.

Since participants had to implicitly relearn how to perceive the other's presence in the virtual space, we hypothesized that there would be a recapitulation of the initial developmental stages of social awareness [4].

We analyzed trial-by-trial objective and subjective changes in sociality that took place during the experiment. Preliminary results reveal that, despite the lack of explicit feedback about task performance, there was a trend for the clarity of social awareness to increase over time.

We discuss the methodological challenges involved in evaluating whether this tendency was characterized by distinct developmental stages of objective behavior and subjective experience.

REFERENCES

[1] T. Froese and T. Fuchs. The extended body: A case study in the neurophenomenology of social interaction. *Phenomenology and the Cognitive Sciences,* 11(2): 205-235 (2012).

[2] T. Froese, C. Gershenson and D. A. Rosenblueth. The dynamically extended mind: A minimal modeling case study. In: *2013 IEEE Congress on Evolutionary Computation* (IEEE CEC 2013), IEEE Press, pp. 1419-1426 (2013).

[3] T. Froese, H. Iizuka and T. Ikegami. Embodied social interaction constitutes social cognition in pairs of humans: A minimalist virtual reality experiment. *Scientific Reports,* 4(3672). doi: 10.1038/srep03672 (2014).

[4] T. Froese, H. Iizuka and T. Ikegami. Using minimal human-computer interfaces for studying the interactive development of social awareness. *Frontiers in Psychology,* 5(1061). doi: 10.3389/fpsyg.2014.0106 (2014).

[1] Departamento de Ciencias de la Computación, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México (IIMAS-UNAM), C.U., D.F. 04510, Mexico. E-mail: **t.froese@gmail.com**

[2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México (C3-UNAM), C.U., D.F. 04510, Mexico. E-mail: **t.froese@gmail.com**

The dual sociality of big data practices: epistemological, ethical and political considerations

Judith Simon [1,2]

Abstract. Big Data, especially if assessed in its societal context, is a contested term and topic. Proponents emphasize its promises for economic prosperity, technological and societal advances, skeptics are alerting us to ethical and societal dangers of big data practices. In line with the symposium's focus on the social aspects of cognition and computing, I will investigate the dual sociality of data practices by focusing on a) big data related to human agents and b) the usage of these big data practices in political decision-making processes affecting societies and the lives of human agents therein. Given this framing, I will argue that any critical assessment of such big data practices requires a combination of epistemological, ethical and political considerations. More precisely, understanding the epistemology of big data is essential for any ethical and political assessment and intervention. On the one hand, numerous ethical problems, for instance those related to anonymity and privacy, can only be targeted if their epistemological premises, such as the possibilities of re-identification, are properly understood. On the hand other, using big data for political decision-making requires an understanding of the epistemic quality of big data analyses, of their premises, potential biases and limits, in order to prevent an unwarranted "trust in numbers" (Porter 1995), just as much as it requires an understanding of the potential ethical and political consequences that come with using big data for governance. Finally, these relationships between epistemology, ethics and politics need to be figured out for any effective governance of big data itself.

ACKNOWLEDGEMENTS

This research has been supported by the Austrian Science Fund (P23770).

REFERENCES

[1] T. Porter, Trust in Numbers. The Pursuit of Objectivity in Science and Public Life, Princeton University Press, Princeton, US, (1995).

[1] Technologies in Practice Group, IT University Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen DK. Email: jusi@itu.dk.

[2] Department of Philosophy, University of Vienna, Universitaetsstr. 7, 1010 Vienna AT. Email: judith.simon@univie.ac.at.

Reasoning In Mathematics and Machines: The Place of Mathematical Logic in Mathematical Understanding

Danielle Macbeth*

Abstract. Mathematical logic and mechanical reasoning have turned out to be largely irrelevant to the practice of mathematics, and to our philosophical understanding of the nature of that practice. My aim is to understand how this can be. We will see that the problem is not merely that the logician formalizes. Nor even is it, as Poincaré argues, that logicians replace all distinctively mathematical steps of reasoning with strictly logical ones. Instead, as will be shown by way of a variety of examples, the problem lies in the way the symbolic language of mathematical logic has been read.

What has mathematical logic to do with mathematical understanding?[1] One would have thought quite a lot. Mathematics is a paradigm of rational activity, of rigorous reasoning; and rigorous reasoning is a central concern of mathematical logic. So, one would think, any adequate understanding of mathematical practice would essentially involve appeal to mathematical logic. One would think. And yet it is by now *clear* that mathematical logic, together with its formalized, mechanistic proofs in which every step conforms to a recognized rule of that logic, is of *no* mathematical interest. Such proofs do not advance mathematical understanding; they are not more rigorous than the informal proofs that mathematicians actually produce; and very often they are simply unintelligible.[2] Mathematical logic, it has turned out, is irrelevant to the practice of mathematics—and to our philosophical investigations into the nature of that practice. [3] Where mathematical logic *has* proved exceptionally fruitful is, of course, in computing. Indeed, according to Kriesel ([2], 143-4), "the clear recognition of just how much reasoning—that is, as far as results are concerned, never mind the processes—can be mechanized is surely the most outstanding contribution of 20th century logic *sub specie aeternitatis*." I think that we should be *very* puzzled by this. Mathematical logic—which, as Burgess ([9], 9) points out, "was developed . . . as an extension of traditional logic mainly, if not solely, about proof procedures in mathematics"—provides the foundations for computer science, mechanical reasoning, but seems to be altogether irrelevant to mathematical reasoning. *How can this be?*

For much of the twentieth century the received view was that mathematical logic and rigorous, mechanical reasoning are less relevant to mathematical practice than one might initially have expected because fully rigorous, formalized proofs are simply too long and tedious to be bothered with in mathematical practice. On this view, mathematicians in their practice take for granted myriad little steps of logic, focusing instead on the mathematically significant steps of a proof. Because in a formalization of a mathematician's proof there are no jumps or gaps in the chain of reasoning, because every step conforms to a small number of antecedently specified rules of logic, what is mathematically interesting about a proof tends, so it is claimed, to get buried in the logical detail of a fully formalized proof.[4] But this is not right. The relationship between a mathematician's proof and a fully formalized proof is not in general that between a gappy and a gap-free proof. In fact, "the translation from informal to formal is by no means a mere matter of routine [as it would be were one only filling in missing steps of logic]. In most cases it requires considerable ingenuity, and has the feel of a fresh and separate mathematical problem in itself. In some cases the formalization is so elusive as to seem impossible" (Robinson [5], 54). Formalizing a mathematician's proof is not so much a matter of formalizing *that* proof (by filling in all the steps) as it is giving a completely different proof, indeed, a different kind of proof. A mathematician's proof is, for example, often explanatory; a formalized proof is not.[5] Mathematicians' proofs are not sketches of formal proofs, essentially like them save for omitting some steps, but instead something quite different.

* Haverford College, USA, email: dmacbeth@haverford.edu

[1] It perhaps needs to be emphasized that my concern here is with mathematical *understanding*, not with mathematics as such. That mathematical logic, for example, model theory, has made useful contributions to the discipline of mathematics seems clear—though even here mathematical logic has contributed less to mathematics as a discipline than one might have anticipated it would.

[2] All these points are well documented in the literature. See, for example, [1, 2, 3, 4, 5], and [6].

[3] That "mathematical logic cannot provide the tools for an adequate analysis of mathematics and its development" is, according to Mancosu [7], 5, one of the three main tenets of the "maverick" tradition in the philosophy of mathematics. It is also a main theme in Grosholz [8].

[4] For a logician's account see, for example, Suppes [10], 128. Mac Lane [11], 377, gives a mathematician's slant on the claim.

[5] As Robinson [5], 56, notes, "formalizing a proof has nothing whatsoever to do with its cognitive role as an *explanation*—indeed, it typically destroys all traces of the explanatory power of the informal proof".

Mathematical reasoning, the reasoning that mathematicians actually engage in, and logical reasoning as understood in mathematical logic, as essentially mechanical, are very different.[6] Most obviously, mathematical reasoning is focused on mathematical ideas while logical reasoning takes account only of logical form. Whereas a fully rigorous proof, in the logician's sense of rigor, is one each step of which conforms to some antecedently specified rule of pure logic and is thoroughly machine checkable, a rigorous proof in the mathematician's sense of rigor is instead one that a mathematician can see to be compelling by focusing on the relevant mathematical ideas and their implications. The two notions of rigor are different and often they are incompatible insofar as the logician's formalizations can undermine the rigor—in the mathematician's sense of rigor—of a chain of reasoning. As Detlefsen explains: "we're most certain to avoid gaps in reasoning when premises *explain* conclusions . . . The greater such explanatory transparency, the more confident we can be that unrecognized information has not been used to connect a conclusion to premises in ways that matter. To the extent, then, that formalization decreases explanatory transparency, it also decreases rigor" ([13], 19).

And there are other differences between the two sorts of proof as well. For example, although the mathematical logician focuses on the logical consequences of given axioms or other starting points, actual mathematical practice is more correctly described as problem solving: one starts not with axioms but instead with a conjecture and working backwards one seeks the starting points that would enable one to prove that conjecture.[7] Finished proofs are, furthermore, of interest to mathematicians not primarily because they establish the truth of their conclusions, which is and must be the primary focus of the mathematical logician, but because they are explanatory, or because they introduce proof techniques that can be brought to bear on other problems.[8] Similarly, what is for the mathematical logician merely a means of introducing an abbreviation can, for the mathematician, constitute a very significant mathematical advance. Although in logic definitions merely abbreviate, in mathematics good definitions, definitions that are fruitful, interesting, and natural, can be exceptionally important, both in themselves, for the understanding they enable, and for what they equip one to prove. For example, it is, as Tappenden [15], 264, argues, "a mathematical question whether the Legendre symbol carves mathematical reality at the joints". Given that the answer to this mathematical question has proved to be an unequivocal "yes", the Legendre symbol cannot be merely an abbreviation. It signifies something mathematically substantive, something of real and enduring mathematical interest.

It is unquestionable that mathematical practice is very different from what the logician and computer scientist would lead one to expect. But to know this is not yet to know *why* it is. Interestingly, the problem is *not* merely that the logician formalizes. "A formal proof," we will say following Harrison (2008, 1395), "is a proof written in a precise artificial language that admits only a fixed repertoire of stylized steps." The logician's formalized proofs clearly fit this characterization. But so do myriad proofs that *anyone* would deem properly mathematical. Consider, for example, this little proof of the theorem that the product of two sums of integer squares is itself a sum of integer squares. We begin by formulating the idea of a product of two sums of integer squares in the familiar symbolic language of arithmetic and algebra:

$$(a^2 + b^2)(c^2 + d^2).$$

Now we rewrite as licensed by the familiar axioms of elementary algebra, omitting obvious steps that could easily be included:

$$a^2c^2 + a^2d^2 + b^2c^2 + b^2d^2$$
$$a^2c^2 + b^2d^2 + a^2d^2 + b^2c^2$$
$$a^2c^2 + 2acbd + b^2d^2 + a^2d^2 - 2adbc + b^2c^2$$
$$(ac + bd)^2 + (ad - bc)^2.$$

This last expression is a sum of two integer squares, which is what we were to show, and so we are done. Our proof is, or could be made to be, fully formal in Harrison's sense: it is "written in a precise artificial language that admits only a fixed repertoire of stylized steps". And yet it is clearly mathematical. It follows directly that being formal is compatible with being of mathematical significance.

The symbolic language of arithmetic and algebra together with the familiar rules governing the use of its symbols is a paradigm of a formal language in Harrison's sense; it is "a precise artificial language that admits only a fixed repertoire of stylized steps". And proofs in this language are, or can easily be made to be, completely gap-free, fully rigorous. But even so the symbolic language of elementary algebra with its rules of use is not destructive of mathematical understanding but instead an enormous *boon* to mathematical understanding. As Grabiner once remarked [16], 357, *that* language has been "the greatest instrument of discovery in the history of mathematics"—of *discovery*. Why is it, then, that in the case of the symbolic language of elementary algebra, the formalization is *transformative* of mathematical practice, whereas in our case, the case of mathematical logic and machine reasoning, the formalization is utterly irrelevant to mathematical practice? What is the difference that is

[6] Again, this is a point that is often made in the literature. See, for example, Devlin [12], Rav [4], and Detlefsen [13].

[7] Cellucci has long emphasized this point. See, for instance, [14]. See also Rav [4], 6: "the essence of mathematics resides in inventing methods, tools, strategies and concepts for *solving problems*".

[8] That is why mathematicians so often reprove theorems. If all they cared about were the truth of theorems this would be inexplicable.

making the difference in the two cases if it is not the mere fact of formalization?

The problem of mathematical logic is not merely that one formalizes in it. Perhaps, then, the problem is that, as Poincaré argues, the logician *replaces* distinctively mathematical reasoning with purely logical, that is, mechanical, reasoning. After all, in our example of products and sums of integer squares we were still working with mathematical ideas, with sums, products, and so on. So, perhaps the real problem with the logician's formalization is not that it is a formalization, but that it is a strictly *logical* one. Perhaps, again as Poincaré argues, to reduce a step of reasoning that mathematicians can see to be valid to a series of little logical steps that anyone, or even a machine, can see to be valid is to destroy the mathematics; perhaps it is to *replace* mathematical knowledge—which constitutively involves one's grasping mathematical ideas and having the ability to see what follows on the basis of those ideas—with merely logical knowledge. Certainly it is true that having the ability to manipulate symbols according to rules, which is what machines can do and what is needed to do mathematical logic, is *not* to be able to do mathematics. So maybe Poincaré is right: to formalize a proof, replace all its distinctively mathematical steps with strictly logical ones is to destroy it, at least as a piece of mathematics.[9]

Poincaré's thought is that mathematical reasoning and understanding are grounded in grasp of mathematical ideas. Because they are, to reduce those ideas, and reasoning and understanding to logic, which is not about anything in particular, is irretrievably to lose the mathematics. This is not clearly right. Consider, first, the case in which what the mathematician takes to be a distinctively mathematical mode of reasoning is shown by the logician to consist in fact in a series of little steps all of which are purely logical. To show that seems clearly to show that what the mathematician had taken to be a distinctively mathematical step of reasoning is at bottom purely logical, strictly deductive. This would seem, furthermore, to be an interesting *mathematical* result: what the mathematician had taken to be a non-logical and presumably ampliative step of reasoning has been revealed to be strictly logical and hence merely explicative. In sum, to discover that some step of reasoning that we had assumed was distinctively mathematical is after all strictly logical is to discover something important *about mathematics*. But if that is right then, in at least some cases, the reduction is not destructive of mathematics but instead a contribution to it.

On the other hand, it does seem right to say, with Poincaré, that there is a crucial difference between the person who can only follow all the little logical steps and the person who can *also* discern the mathematical ideas at work in a proof. As Detlefsen explains: "even perfect *logical* mastery of a body of axioms would not, in his [Poincaré's] view, represent genuine mathematical mastery of the mathematics thus axiomatized. Indeed it would not in itself be indicative of any appreciable degree of mathematical knowledge at all: knowledge of a body of mathematical propositions, plus mastery over their logical manipulation, does not amount to mathematical knowledge either of those propositions or of the propositions logically derived from them" ([18], 210). According to Poincaré, replacing all mathematical modes of inference with a series of purely logical little steps destroys the mathematical unity of the proof that is essential to any adequate understanding of it. But why, and how, does it do that? Again, if what we had thought was a distinctively mathematical mode of reasoning turns out to be reducible to a series of strictly logical steps then that is an important, and importantly mathematical, discovery. So the cases of concern must be ones in which, paradigmatically, steps that are mathematically motivated are made explicit in conditionals, so that the conclusion can now follow as a matter of pure logic.[10] And now someone not in the know might well understand the step merely as a matter of logic: if A then B (which here formulates a mathematical rule), but A, therefore B. But is there any reason to think that the *mathematician* could not still see that what is crucial mathematically is that if A then B, that it is this mathematical rule that is licensing the move from A to B? If there remains a discernable difference between cases in which some *mathematical* rule is being followed and cases that merely involve some truth-function, either not-A or B, then there will remain a difference between what the mathematician can discern in the proof and what the non-mathematician will discern.

Suppose, for example, that we took our little proof that the product of two sums of integer squares and made it strictly logical, that is, every step in conformity with a rule of logic. Where before we had drawn a mathematical inference, we now write down the relevant conditional and justify the step by modus ponens. Once we have done this for all the steps of the proof, it might well be much harder to discern the important steps of the proof, as well as its key ideas—to order the summands in a certain way and then add and subtract the same thing so as to be able to factor—but those steps and ideas would still be there to be discerned. The formalized proof would not in that case destroy the mathematics—though it also would no longer highlight it. But if that is right then Poincaré's claim that replacing distinctively mathematical forms of reasoning with strictly logical ones destroys the mathematics cannot be quite right. The complete and utter lack of interest mathematicians show for formalized proofs strongly

[9] This, the mathematical logician is likely to respond, is merely a matter of psychology, and irrelevant to our philosophical understanding of what is going on in a piece of mathematical reasoning. See Goldfarb [17].

[10] Detlefsen [19] considers this sort of case.

suggests that, just as Poincaré charges, the mathematics *is* being lost in the formalization. But given that this loss is not a necessary result of formalizing in the language of logic, we have yet to understand what is really going on here, *why* the mathematical logician's formalized, mechanical proofs are so completely irrelevant to mathematical practice.

Mathematicians do not need to study logic and they do not use the signs of logic except here and there as abbreviations for everyday words: "the everyday use of logical symbols we see [in mathematical practice] today closely resembles an intermediate 'syncopation' stage in the development of existing mathematical notation, where the symbols were essentially used for their abbreviatory role alone" (Harrison [1], 1398). And so, it is sometimes claimed, the signs even of a mathematical language such as the symbolic language of elementary algebra similarly do nothing more than to provide abbreviations of words of natural language. But this is simply (and really rather obviously) not true: mathematical languages such as the symbolic language of algebra, as they are actually used, function in a fundamentally different way from the way natural languages function. In particular, one can reason *in* a mathematical language in a way that is simply impossible in natural language. Although one cannot, for instance divide the words 'six hundred and seventy-three' by the word 'seventeen', one *can* divide the Arabic numeral '673' by the numeral '17'. In the latter case one works out the answer on paper, through a chain of paper-and-pencil reasoning (or else one imagines oneself doing this). Even more obviously, although one cannot bisect the word 'line' one *can* bisect a Euclidean (drawn) line.

But not all mathematical reasoning is a matter of scribbling in a specially devised system of written marks. Is the reasoning in other cases instead done in natural language? It is not, at least not in the way that it *is* done in a specially devised written mathematical language. Where there is no system of written marks within which to work, the reasoning is instead performed *mentally*, by reflecting on ideas in ways that can then be *reported* in natural language.[11] The ancient proof that there is no largest prime is a familiar example of such a report of mental mathematics. Lacking any means of displaying what it is to be a prime number, or even what it is to be a product of numbers, ancient Greek mathematicians could nonetheless work mentally with the idea of a prime number, and with the idea of a product of a finite list of primes plus one, and could recognize that such a product of primes plus one must either be prime or have a prime divisor larger than any hitherto considered. And having determined this, they could report their reasoning in just the way Euclid in fact does in the *Elements*. Al-Khwarizmi, a ninth century Islamic algebraist, similarly can tell us in natural language how to find a particular root. What he cannot do is *show* us how to determine that root by performing a calculation.[12]

Sometimes we can work out the solution to a mathematical problem by paper-and-pencil reasoning. In other cases, we instead must reflect on the relevant ideas in order to solve the problem by a chain of mental reasoning. It can also happen that a piece of mathematical reasoning that at first can only be reported in natural language can later have a counterpart displayed in a mathematical language. A very simple example is this from Euclid's *Elements*, Proposition IX.21: if as many even numbers as you like are added together, the whole will be even. The crucial step in the reasoning, as reported by Euclid, is that since each of the numbers added together is even, each has, by the definition of *even*, a half part; thus it follows that the whole has a half part, and hence (by definition) is even. That is, we are simply to *see*, as it were with the mind's eye, that if each summand has a half part then the sum does as well. And this is, admittedly, very easy to see; but it is not by logic alone, or any antecedently specified step of mathematical reasoning, that we see this. It is an intuitively obvious step of reasoning but nevertheless one that is *not* justified by any rule. The inference is only reported, and either one gets it or one does not. But a comparable step *can* be shown in the symbolic language of algebra, and in that case, the conclusion *does* follow by an antecedently specified rule. First, we display in the language what it is to be even, that is, the *form* that even numbers take in the language, namely, $2n$, for natural number n. Now we display an arbitrarily long finite sum of such numbers: $2a + 2b + 2c + \ldots + 2n$.[13] Because there are explicitly formulated rules governing the use of such signs, we can apply a rule to transform the expression thus: $2(a + b + c$

[11] There are also a wide variety of intermediate cases, cases involving systems of written marks together with some mental mathematics. Leaving these out of consideration does not affect the points at issue here; what matters for our purposes is the two extremes, the case in which one has a system of written marks within which to reason and the case in which one instead engages in purely mental reasoning, the results of which can be reported in natural language.

[12] al-Khwarizmi writes: "*Roots and squares are equal to numbers*: for instance, 'one square, and ten roots of the same, amount to thirty-nine dirhems'; that is to say, what must be the square which, when increased by ten of its own roots, amounts to thirty-nine? The solution is this: you halve the number of roots, which in the present instance yields five. This you multiply by itself; the product is twenty-five. Add this to thirty-nine; the sum is sixty-four. Now take the root of this, which is eight, and subtract from it have the number of the roots, which is five; the remainder is three. This is the root of the square which you sought for; the square itself is nine" ([20], 229). The correctness of the implicit rule would have been demonstrated geometrically.

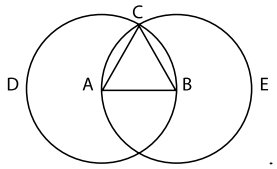
[13] It is worth noting in this context that our symbolic expression is arbitrary along two different dimensions. First, each of the letters 'a', 'b', 'c', and so on stand in for some natural number not further specified. The letter 'n' is different insofar as it is *also* arbitrarily large. My thanks to Jean Paul Van Bendegem for making this explicit.

+ . . . + n). This is manifestly an even number; we have our proof.

As this little example of the sum of even numbers shows, and Whitehead [21], 34, explicitly says, "by the aid of symbolism, we can make transitions in reasoning almost mechanically by the eye, which otherwise would call into play the higher faculties of the brain". Once we have symbolized our problem we do not have to *think* about what follows from the fact that each number in the sum has a half part. We simply have to apply a rule that enables us to *show* that the sum is even. Of course, we do need to be able to see the mathematical ideas in the symbolism, for example, that the expression '2($a + b + c +$ . . . $+ n$)' designates an even number; but it is the symbolism, not the ideas, that enables us to operate as we do. "In mathematics, granted that we are giving any serious attention to mathematical ideas, the symbolism is invariably an immense simplification. It is not only of practical use, but is of great interest. For it represents an analysis of the ideas of the subject and an almost pictorial representation of their relations to each other" (Whitehead [21], 33). Again, when one is working in a written mathematical language such as the symbolic language of arithmetic and algebra one does not have to *think* about the relevant mathematical ideas in the way one *does* have to think about them in the absence of such a language. And *that* is just our problem: we have in mathematical logic as in, say, the symbolic language of elementary algebra, a "precise artificial language that admits only a fixed repertoire of stylized steps," a formal language "designed so that there is a purely mechanical process by which the correctness of a proof in the language can be verified" (Harrison [1], 1395). But unlike the symbolic language of elementary algebra, the language of mathematical logic is of no mathematical interest or utility. *Why*?

Although it might have been expected to, the language of mathematical logic and mechanical reasoning has not proved to be a mathematically tractable language, a language within which to reason in mathematics. Mathematicians working today do not display their reasoning in the formal language of mathematical logic but only report it.[14] We need, then, to think about what is required of a language within which to display mathematical reasoning. The short answer, explicit already in Leibniz, is that the language must exhibit mathematical content in a mathematically tractable way, that is, in a form that enables reasoning in the guise of a series of rule-governed manipulations of signs. It must be, as Frege also saw, at once a *lingua characteristica* and a *calculus ratiocinator*. There is, however, a hitch: it is possible to read one and the same notation *either* as formulating content in a mathematically tractable way *or* as merely recording information in a way enabling mechanical reasoning. And because one and the same notation can be read in either of these two very different ways, it is impossible to show what is needed in a mathematical language by appeal only to a system of signs. One must also take into account *how expressions in the system are understood*. Some examples will help to clarify this essential point.

Consider, first, the familiar distinction between a mathematical and a mechanical proof, which we here apply to the first proposition of Euclid's *Elements*: to draw an equilateral triangle on a given straight line. The diagram for both the mechanical and the mathematical proof is this:



But it is drawn with very different intentions in the two cases. Because, in a mechanical proof, the aim is to construct an actual, *empirical* triangle, one with, as far as possible, sides that are actually equal in length, one is well advised, in that case, to use a compass to draw the required circles and a straight-edge to draw the lines that are radii of the circles and form the sides of the triangle. One could then measure the lines to determine how closely they approximate lines equal in length. In a *mathematical* demonstration no such precautions are necessary because the drawn circles are not intended in this case to be *instances* approximating as far as possible the ideal of a mathematical circle. Instead they are drawn to formulate or display the *content* of the concept of a circle, *what it is* to be a circle, namely, a plane figure all points on the circumference of which are equidistant from a center.[15] As formulating such content, the drawn circles license inferences: if one has two radii of one circle then one can infer that they are equal in length—whether or not they *look* equal in length in one's drawing. What in the mechanical proof is treated as a means of constructing some *particular* triangle (with its particular spatial location, and particular size and orientation) is in the mathematical proof a way of solving a strictly mathematical and hence constitutively general problem, the problem of the construction of *an* equilateral triangle—not any equilateral triangle in particular—on a given straight line. As Shabel [25], 101, puts the point in a

[14] Avigad [22] makes this point. It is also the basis for Azzouni's [23] derivation-indicator account of mathematical proofs. Rav [4], 13, makes the point in an especially graphic way: "The argument-style of a paper in mathematics usually takes the following form: ' . . . from so and so it follows that . . . . , hence . . . .: as is well known, one sees that . . . ; consequently, on the basis of Fehlermeister's Principal Theorem, taking into consideration α, β, γ, . . . , ω, one concludes . . . , as claimed'."

[15] See my [24], Chapter 2.

discussion of Kant on pure and empirical intuition in mathematical practice, "the mechanical demonstration is not distinguished from the mathematical demonstration by virtue of a distinction between an actually constructed figure and an imagined figure, but rather by the way in which we operate on and draw inferences from that actually constructed figure". One and the same drawing is regarded in two systematically different ways in a mechanical and a mathematical proof.

A second example is this. Suppose that, having not yet learned various simple sums (but knowing how to count), one wished to determine how many seven things and five things make when taken together. One might proceed by marking out seven strokes and then five more and counting how many that is. This is a mechanical reading of the display of seven and five strokes. One thinks of it as presenting two collections of things, namely, strokes that taken together make a collection of twelve things—as one discovers by counting the whole collection. The proof is mechanical insofar as one is actually putting things together in order to see empirically, by counting, what totality they make. That one is working with a system of written marks is irrelevant; one could have worked as easily with pebbles, or peaches, or puppies. (Well, maybe not *as* easily.)

Now we regard the strokes differently, not merely mechanically but as signs of a Leibnizian language within which to formulate content and to reason. In this case we do not regard each stroke as standing in for a thing to be counted, or indeed as itself a thing to be counted. Instead we regard each stroke as expressing something like a Fregean sense, as contributing to the sense of a whole collection of signs that together, as *one* complex sign, designates a number, say, the number seven, or the number five. So regarded, the collection of seven strokes exhibits *what it is* to be the number seven, namely, a certain multiplicity. The collection of seven strokes is not in this case a collection of seven things; it is a *single* complex sign for *one* number, a sign that, by contrast with a simple sign such as the Arabic numeral '7', displays what it is to be seven in a mathematically tractable way. Given the display of five and seven using the Leibnizian stroke language, one can progressively reconfigure the whole display, adding strokes from the sign for the number five to the strokes making up the sign for seven in such a way that one eventually achieves a complex sign for the number twelve. Much as in Euclid's system one shows (mathematically) that an equilateral triangle can be constructed on a given straight line, so here one shows that (a sign for) the number twelve can be constructed from (signs for) the numbers seven and five. And the result in both cases is synthetic a priori insofar as what one has to begin with provides everything one needs in order to perform the required construction through a course of mathematical reasoning. In the mathematical demonstrations, the triangle, and the number twelve, are *not* contained already implicitly in one's starting points, but the *potential* for achieving them is there in the starting points. They *can be* produced, which means that the result is synthetic rather than analytic. But they are not produced mechanically, that is, empirically, as in a mechanical proof. They are produced mathematically. The result is a priori.

Notice further that in both the Euclidean diagram and the Leibnizian stroke language, the signs are taken to function in a very distinctive way. In the case of the Euclidean diagram, what are at first seen as two radii of a circle (required in order to determine that they are equal in length) are later seen as sides of a triangle. One and the same sign, namely, a drawn line, is in the context of one collection of signs a radius of a circle and in the context of another collection of signs a side of a triangle. We can take it either way. What we cannot do, of course, is take that same line as, say, an angle or the circumference of a circle. The drawn line expresses a sense that completely and perfectly delimits its possibilities for designating in this or that use in a diagram. Similarly, and even more simply, for the strokes: a stroke that I first see as a part of the sign for five, as contributing a sense to the complex sign designating five, I later see as part of the sign for, say, the number eight constructed out of the original seven strokes plus one more. There is nothing like this in the mechanical proofs. In mechanical proofs, the marks are simply material things that are constrained by the physics of the case. The expressive intentions of a thinker are irrelevant when one is proving mechanically.

We have seen that in a mechanical proof one pictures or records something, for instance, a particular circle or how many in a collection. In a mathematical proof one instead *formulates content*, what it is to be, say, a circle or the number seven; and one does so in a way that enables reasoning in the system of signs. We can similarly read a complex sign of Arabic numeration in either way, *either* as recording how many (how many units, tens, hundreds, and so on), that is, mechanically, *or* as formulating the arithmetical or computational content of numbers. If one sees the numeral the former way then one will take it that a calculation in Arabic numeration is merely a mechanical expedient for arriving at a desired result, not in any essential way different from the sort of mechanical manipulations that can be made on Roman numerals.[16] If one instead sees the Arabic numeral as expressing arithmetical content, one will think of the calculation as a bit of *mathematical reasoning*, as an episode of mathematical thought rather than as something mechanical, and hence as something quite different from the manipulations that can be made on Roman numerals.[17]

[16] See Schlimm and Neth [26] for such a view.

[17] I am of course assuming that the signs of Roman numeration are being read mechanically, and this is certainly the most natural way to read

In the examples we have so far considered one has a system of written marks that can be conceived in either of two fundamentally different ways, either mechanically, as providing an instance or record of something that can then be operated on in some way to yield the desired result, or mathematically. And in the mathematical case, we have seen, one formulates the content of some mathematical notion—the content of the concept of a circle, say, or that of some particular number—and one does so in a way that enables reasoning *in* the system of signs. Now we need to consider how things stand with systems of signs of logic.

Consider, first, Peirce's system of alpha graphs. Shin [27] has shown that although we can take the primitives of the system directly to picture or record, we can also take them only to express senses independent of their involvement in a proposition, to contribute a sense to the whole thought expressed, which thought can then be variously analyzed.[18] In Peirce's system considered the first way, that is, mechanically, to enclose a propositional sign in parentheses just is to negate it; the concatenation of signs serves similarly as conjunction.[19] The complex sign '((A)(B))', then, is to be read as recording the fact that it is not the case that not-A and not-B. But we can also read this same complex sign as an expression of a *Leibnizian* language, as exhibiting a thought that can be variously regarded, for instance, as the disjunction of A and B, or as the conditional 'if not-A then B', or as the conditional 'if not-B then A'. Much as a line in a Euclidean diagram is a radius or side of a triangle *only* relative to a way of regarding that diagram, so here on the Leibnizian reading, the collections of signs is a disjunction or conditional only relative to a way of regarding it. And of course just this same point can be applied to the standard notation of mathematical logic and as well to Frege's *Begriffsschrift* notation. Expressions in all these various systems of notation can be read *either* as picturing some state of affairs, say, that if not-A then B, *or* as displaying logical content in a way that can be regarded in turn *either* as, say, a conditional with a negated antecedent *or* as a disjunction, depending on whether one takes the tilde (negation stroke) to attach to the content A or to function together with the horseshoe (conditional stroke) to designate disjunction.

Read mechanically a notation such as that of mathematical logic or Frege's *Begriffsschrift* records information, and the rules governing the manipulation of the signs enable one to show that other information is also contained therein. Manipulating the signs according to the rules can thus make explicit what is contained already implicitly. The deduction is merely explicative. Much as making seven strokes and then making five more is already to have twelve strokes, so that counting up the resultant number of strokes is a mechanical means of determining how many, so manipulating the signs of some premises expressed in the language of mathematical logic, as it is generally conceived, is a mechanical means of showing that certain information is contained already in one's starting points. But, we now know, we can also read the notation differently, as a notation of what I have been calling a Leibnizian language. Furthermore, we know that in general, because the signs of a Leibnizian language only express senses independent of a context of use, those signs can be used to formulate the contents of concepts. Can the signs of a logical language, read as a Leibnizian language, similarly be used to formulate the contents of concepts and to do so in a way that enables reasoning in the system of signs? They can.

In Euclidean diagrammatic reasoning, the content of the concept of, say, a circle is conceived diagrammatically, that is, as something that can be exhibited in a drawn circle. In Descartes's analytic geometry, the content of that same concept is conceived instead arithmetically. It is given in the equation '$x^2 + y^2 = r^2$'. We have further seen that although the content of the concept of an even number, or of an odd number, cannot be displayed in a Euclidean diagram, those contents *can* be displayed in a mathematically tractable way in the language of arithmetic and algebra, the notion of an even number as '$2n$' and that of an odd number as '$2n + 1$'.

Different mathematical languages can thus involve very different conceptions of what are in fact the same mathematical concepts, very different analyses of those concepts. What sort of analysis is needed, then, for the sort of reasoning from concepts that is characteristic of contemporary mathematical practice? Given that the mathematical practice we are concerned with is that of *deductive* reasoning from concepts, the answer is clear: a logical analysis. We need to be able to display the contents of concepts as they matter to inference.

What we are after is a way to formulate the contents of mathematical concepts that enables deductive reasoning in the system of signs. And we know by now that to achieve this it is not enough to introduce various signs together with rules governing their use because any such system of signs can be read either as a Leibnizian language or merely mechanically. To exhibit the contents of concepts in a mathematically tractable way, we need to read the system of signs as a *Leibnizian* language, its primitive signs as only expressing senses independent of any context of use, because only so can a whole complex of signs serve to designate a *single* concept, only so can we display content at all.

---

them. But our Leibnizian stroke language suggests that it may be possible, if difficult, to read signs of that language likewise as the signs of a Leibnizian language.

[18] Shin does not put the point this Fregean way, but could have done.

[19] In Peirce's system one encircles propositional signs rather than enclosing them in parentheses. The latter is, however, more convenient here and works in essentially the same way.

Think again of our simple stroke language or of the system of Arabic numeration. In both cases we can treat the primitive signs either as having their meaning or designation independent of any context of use or as having only a sense independent of a context of use. Taken in the former way, as having meaning (designation) independent of any context of use, the signs are signs of a mechanical language: a collection of five strokes is just that, a collection of five things, and an Arabic numeral such as '376' similarly denotes a collection, a collection of three hundreds and seven tens and six ones. A numeral such as '3' in the language so conceived invariably denotes some particular number, here the number three; its position serves only to indicate what is being so counted, whether ones or tens or hundreds or something larger. But we know that we can also read the language differently, the primitive signs of the language as only expressing senses independent of a context of use. In that case, the collection of five strokes is a complex sign that designates *one* thing (not five things), namely, the number five. And the Arabic numeral '376' similarly is a complex name of one number. The numeral '3' does not in this case designate three (of something) no matter what the context; instead it contributes a sense to a whole that only as a whole functions as a name for something, namely, in our example, for the number three hundred and seventy-six. In just the same way, we can regard a definition of a mathematical concept in a written system of logical and mathematical signs *either* as recording necessary and sufficient conditions, the state of affairs that obtains if the concept applies, *or* as exhibiting the content of the concept as it matters to inference.

In mathematical logic and computing, the definiens of a definition is understood to provide necessary and sufficient conditions for the application of the concept, and the definition as a whole is taken merely to introduce an abbreviation for those conditions. The definition has no philosophical or mathematical significance; it is a convenience. The defined concept is, in that case, reduced to, or replaced by, a set of conditions much as a number is reduced to, replaced by, a collection of things when it is represented mechanically by a series of strokes. But again, in actual mathematical practice, definitions—both those that stipulate a simple sign for some complex notion and those that provide a new and deeper analysis for some concept already in use—can constitute a significant mathematical advance, one that is just as important mathematically as a new proof. And the definition *is* mathematically important precisely because and insofar as it formulates mathematical content in a tractable way, in a way enabling new and better, more explanatory proofs. But in order to do that in a specially devised system of signs, the system of signs must be read as a Leibnizian language the primitive signs of which only express senses.

In a definition in a Leibnizian language the defined concept is not *reduced* to something else but instead designated. Indeed, it is designated twice, once by a simple sign, the definiendum, and again by a complex sign, the definiens. The two signs have the same designation or meaning. But although they designate one and the same concept, the two signs express two very different Fregean senses. And one can just see that they do insofar as the one sign is simple while the other is complex. Because the definiens is a complex sign that is made up of a variety of primitive signs of the language, the transformation rules of the language can be applied to it in a way that is manifestly impossible in the case of the simple sign that is the definiendum. The simple sign, the definiendum, is unequivocally a name for the relevant concept. The complex sign, the definiens, is also a name for that concept but because it is complex it can enable one to reason in light of the content it displays and discover thereby new truths about the concept in question. But, of course, one can see all this to be going on only if one understands the system of signs as we have done here, not merely mechanically but as a Leibnizian system the primitive signs of which only express senses independent of any context of use. In a fully formalized proof in a Leibnizian language the mathematics is not destroyed but instead displayed, and although superficially each step is the same as any other, one and all steps of logic, the knowledgeable reader can nonetheless distinguish those steps that are mathematically important from those that are trivial, and can discern as well the key mathematical ideas of the proof. The language functions, in other words, in much the way the symbolic language of arithmetic and algebra does, to extend our mathematical knowledge.

It has long been known that the reasoning mathematicians engage in is quite unlike reasoning as it is understood in mathematical logic and computer science. What has proved much harder to determine is why that is. The problem is not merely that the logician formalizes, either in the sense of producing proofs that are completely gap-free or in the sense of working in an artificial symbolic language the licensed moves of which are all specified in advance. Nor even is it, as Poincaré suggests, that logicians replace all distinctively mathematical steps of reasoning with strictly logical ones. We know that all these explanations fail because it is possible to find or develop examples of mathematical proofs in the formula language of arithmetic and algebra that exhibit some or all of the features that have been focused on and nevertheless *retain* their mathematical interest. The explanation for the irrelevance of mathematical logic to mathematics must, then, be something distinctive of that logic in particular. And so it is: the reason mathematical logic is irrelevant to mathematical practice is that its language is read mechanically. Because reasoning in mathematics is *not* merely mechanical, to formalize a mathematician's proof

in mathematical logic really does destroy it as a piece of mathematical reasoning—just as Poincaré thought. Because the language is read mechanically, all differences between mathematically significant steps of reasoning and merely trivial steps of logic are completely effaced. No one, not even the mathematician, can now discern what is mathematically important in the proof.[20]

I began with a question: what has mathematical logic to do with mathematical understanding? In particular, why is it that a fully formalized, mechanical proof in mathematical logic destroys the mathematical interest of the proof given that in other cases of formalizations, paradigmatically in the symbolic language of arithmetic and elementary algebra, the result is of clear and significant mathematical interest? The problem, we have found, does not lie in the language of mathematical logic conceived simply as a system of signs. The problem lies in the way that system of signs is conceived, in the fact that it is conceived mechanistically. Were it to be conceived instead as a Leibnizian language—that is, as a language within which to display the contents of concepts in a way enabling one to reason on the basis of those contents in the system of signs—then it could be used in formalizations in much the way the language of arithmetic and algebra is. It could be used, that is, to clarify and enrich both mathematical practice and our understanding of that practice. And *that* is to say that it could be used in just the way Frege envisaged the use of his *Begriffsschrift*, his concept-script—if only we had understood him.[21]

ACKNOWLEDGEMENTS

My thanks to Emily Grosholz for very thoughtful and useful comments on an earlier draft.

REFERENCES

[1] J. Harrison, "Formal Proof—Theory and Practice", *Notices of the AMS* **55** (11), 1395-1406 (2008).

[2] G. Kreisel, "Mathematical Logic: Tool and Object Lesson for Science", *Synthese* **62** (2), 139-51 (1985).

[3] K. Manders, "Logical and Conceptual Relations in Mathematics", In *Logic Colloquium '85*, Elsevier Science, North Holland, 1987.

[4] Y. Rav, "Why Do We Prove Theorems?', *Philosophia Mathematica* (III) **7**, 5-41 (1999).

[5] J. A. Robinson, "Informal Rigor and Mathematical Understanding", *Computational Logic and Proof Theory*, ed. Georg Gottlob, Alexander Leitsch, and Daniele Mundici, Springer, Berlin and Heidelberg, 1997.

[6] W. P. Thurston, 1994. "On Proof and Progress in Mathematics".,*Bulletin of the AMS* **30**: 161-77 (1994), reprinted in *18 Unconventional Essays on the Nature of Mathematics*, ed. Rueben Hersh, Springer 2006.

[7] P. Mancosu, "Introduction", *The Philosophy of Mathematical Practice*, Oxford University Press, Oxford and New York: 2008.

[8] E. R. Grosholz, *Representation and Productive Ambiguity in Mathematics and the Sciences*. Oxford University Press, Oxford, 2007.

[9] J. P. Burgess, "Proofs about Proofs: A Defense of Classical Logic. Part I: The Aims of Classical Logic", *Proof, Logic and Formalization*, ed. Michael Detlefsen. Routledge, London and New York, 1992.

[10] P. Suppes, *Introduction to Logic* [1957], Dover, Mineola, N.Y., 1999.

[11] S. Mac Lane, *Mathematics: Form and Function*, Springer, New York, 1986.

[12] K. Devlin, "When is a Proof?", http://www.maa.org.external_archive/devlin/devlin_06_03.html. Mathematical Association of America, Devlin's Angle, 2003.

[13] M. Detlefsen, "Proof: Its Nature and Significance", *Proof and Other Dilemmas: Mathematics and Philosophy*, ed. Bonnie Gold and Roger A. Simons, The Mathematics Association of America, Washington, D.C., 2008.

[14] C. Cellucci, *Rethinking Logic: Logic in Relation to Mathematics, Evolution, and Method*. Springer Science + Business Media, Dordrecht, 2013.

[15] J. Tappenden, "Mathematical Concepts and Definitions", *The Philosophy of Mathematical Practice*, ed. Paolo Mancosu, Oxford University Press, Oxford and New York, 2008.

[16] J. V. Grabiner, "Is Mathematical Truth Time-Dependent?", *The American Mathematical Monthly* **81**, 354-65, (1974).

[17] W. Goldfarb, "Poincaré Against the Logicists", *Minnesota Studies in the Philosophy of Science*, Vol XI: History and Philosophy of Modern Mathematics, ed. William Aspray and Philip Kitcher, University of Minnesota Press, Minneapolis, 1988.

[18] M. Detlefsen, "Brouwerian Intuitionism", *Proof and Knowledge in Mathematics*, ed. Michael Detlefsen, Routledge, London and New York, 1992.

[19] M. Detlefsen, "Poincaré Against the Logicians". *Synthese* **90** (3), 349-78, (1992).

[20] Fauvel, John and Jeremy Gray (eds.), *The History of Mathematics: A Reader*, Macmillan Press, London, 1987.

[21] A. N. Whitehead, *Introduction to Mathematics* [1911], Barnes and Noble Books, New York, 2005.

[22] J. Avigad, "Mathematical Method and Proof". *Synthese* **153** (1), 105-59 (2006).

[23] J. Azzouni, "The Derivation-Indicator View of Mathematical Practice", *Philosophia Mathematica* (3) **12**, 81-105, (2004).

[20] Mathematical logic is so named because and insofar as it is (as Boole explicitly urged it should be) a branch of mathematics; it is a mathematical investigation into (mathematically investigable) patterns of reasoning. The logic that one would need for the purpose of actually reasoning in the system of signs in mathematics would be a mathematical logic in a very different sense.

[21] Frege explicitly notes that his aim was different from Boole's, and different in just the way I have tried to bring out here. He writes in "On the Aim of the 'Conceptual Notation'": "I did not wish to present an abstract logic in formulas [as Boole did], but to express a content through written symbols in a more precise and perspicuous way than is possible with words. In fact, I wished to produce, not merely a *calculus ratiocinator*, but a *lingua characteristica* in the Leibnizian sense" ([28], 90-91). Or again in "Boole's logical Calculus and the Concept-script": "In contrast [to what Boole aimed for] we may now set out the aim of my concept-script. Right from the start I had in mind the *expression of a content*. What I am striving after is a *lingua characterica* in the first instance for mathematics, not a *calculus* restricted to pure logic" ([29], 12). See my [30] for an extended defence of this way of reading Frege's distinctive two-dimensional notation.

[24] D. Macbeth, *Realizing Reason: A Narrative of Truth and Knowing*. Oxford University Press, Oxford and New York, 2014.

[25] L. Shabel, *Mathematics in Kant's Critical Philosophy: Reflections on Mathematical Practice*. Routledge, New York and London, 2003.

[26] D. Schlimm, and H. Neth, "Modeling Ancient and Modern Arithmetical Practices: Addition and Multiplication with Arabic and Roman Numerals", *Proceedings of the 30th Annual Cognitive Science Society*, ed. B. Love, K. McRae, and V. Sloutsky, Cognitive Science Society, Austin Tex., 2008.

[27] S-J. Shin, *The Iconic Logic of Peirce's Graphs*, MIT Press, Cambridge, Mass. and London. 2002.

[28] G. Frege, "On the Aim of the 'Conceptual Notation'" [1882], *Conceptual Notation and Related Articles*, ed. T. W. Bynum, Clarendon Press, Oxford, 1972.

[29] G. Frege, "Boole's logical Calculus and the Concept-script" [1880], *Posthumous Writings*, ed. H. Hermes, F. Kambartel, and F. Kaulbach, and trans. P. Long and R. White, University of Chicago Press, Chicago, 1979.

[30] D. Macbeth, *Frege's Logic*, Harvard University Press, Cambridge. Mass., 2005.

Propagation of the Effects of Certain Types of Military Psychological Operations in a Networked Population

Colette Faucher[1]

Abstract. In modern asymmetric military conflicts, the Armed Forces generally have to intervene in countries where the internal peace is in danger. They must make the local population an ally in order to be able to deploy the necessary military actions with its support. For this purpose, psychological operations (PSYOPs) are used to shape people's behaviors and emotions by the modification of their attitudes in acting on their perceptions. PSYOPs aim at elaborating and spreading a message that must be read, listened to and/or looked at, then understood by the info-targets in order to get from them the desired behavior. A message can generate in the info-targets, reasoned thoughts, spontaneous emotions or reflex behaviors, this effect partly depending on the means of conveyance used to spread this message. In this paper, we focus on psychological operations that generate emotions. We present a method based on the Intergroup Emotion Theory, that determines, from the characteristics of the conveyed message and of the people from the population directly reached by the means of conveyance (*direct info-targets*), the emotion likely to be triggered in them and we simulate the propagation of the effects of such a message on indirect info-targets that are connected to them through the social networks that structure the population.

1 INTRODUCTION

Nowadays, when the Armed Forces have to intervene in the framework of asymmetric conflicts, it is essential for them to make the local population of the concerned country an ally. Operations of influence are then essential and take precedence over combat actions. SICOMORES (SImulation COnstructive et MOdélisation des effets des opérations d'influence dans les REseaux Sociaux) is a system that simulates the effects of some operations of influence (CIMIC, PSYOP and KLE operations) on the population structured within social networks underlied by diverse links (religious link, ethnic link, etc.). PSYOP operations are meant to spread a message that must be read, listened to and/or looked at, then understood by the info-targets [3]. A message can generate in them reasoned thoughts, spontaneous emotions or reflex behaviors. In this paper, we focus on the simulation of the effects of messages likely to trigger emotions, both on the direct info-targets and on the indirect ones due to propagation through the social networks that structure the population.

In section 2, we explain why the system SICOMORES is interesting and useful for the military. In section 3, we describe the state of the art of the systems dealing with the propagation of sentiments/emotions in a social network, then SICOMORES' theoretical bases are outlined in section 4, followed by the specification of the Human Terrain of the environment in section 5. The characteristics and the modeling of psychological operations, as well as the mechanism of effect generation of emotion-triggering psychological operations are then respectively detailed in section 6 and 7. Section 8 concludes the paper.

2 INTEREST OF THE SYSTEM SICOMORES

A military analyst who is in charge of conceiving psychological messages, is generally a person who knows very well the country to which the recipients belong, its language and the local culture through all its facets. When he intends to reach a given group of people being part of the population and characterized by their social, psychological and/or cultural features (the *direct info-targets*) and to have them feel a specific emotion, he knows how and what to say. He can be efficient without the help of a system. However, a major factor intervenes when a message is spread: the means of conveyance of this message, because it defines the scope of the message, that is the area within which the direct info-targets can be reached. What is to be taken into account is that, within this area, other people than the direct info-targets may be reached. When they get the message, they will have their own reaction, that the analyst has not thought about, but that can be very important and can play a great part. In that case, the system will be able to compute this reaction, because it has the knowledge that describes the characteristics of the Sociocultural Groups of people that were reached by the message in a non intended way. For that purpose, it will use the Intergroup Emotion Theory presented in section 4. What we are underlining is the superiority of the computer over a human being as regards the capacity of storing information such as all the types of people that can be found on a specific area and also its ability to compute an emotion felt by people characterized by social features when they get a given psychological message. There is still another aspect for which the computer will help the analyst. In the country where the conflict takes place, the population is structured within networks based on different links, political, religious, family links, for instance. When a direct info-target is reached by a message, they will probably propagate it, according to some rules we will explain in section 4, to the people they are connected to by the various links (the *indirects info-targets*) and those people will in turn do the same thing with their own connections and so on. Contrary to a human being, the computer can memorize the structure of the networks and then it can determine who will be the indirect info-targets and what will be the effect of the message on them.

[1] LSIS, Aix-Marseille University, 13397, Marseille, Cedex 20, FRANCE. Email: colette.faucher@lsis.org.

From these considerations, we can see how a system like SICOMORES can be precious to the military who use psychological operations, to predict the impact of a message on the whole population.

3 PROPAGATION OF SENTIMENTS OR EMOTIONS IN SOCIAL NETWORKS

To our knowledge, all the works that deal with the propagation of sentiments/emotions in a social network exclusively refer to online virtual communities.

In [14], the authors have developed an agent-based framework to model the emergence of collective emotions. A node is an individual called a Brownian agent which has emotions described by their valence and their arousal that change according to a stochastic dynamics. An individual's next emotional state is determined with a linear sum of psychological factors, including the feedback of the community, and a Gaussian error. In this work, a unique source of nformaton is supported, contrary to [7] where multiple sources of information are taken into account. In [6], the author generates a fully-connected polar social network graph from the sparsely connected social network graph in the context of blogs, where a node represents a blogger and the weight of an edge connecting two bloggers represents the sentiment of trust/distrust between them. The sign and magnitude of this sentiment value is based on the text surrounding the link. The author uses trust propagation models to spread this sentiment from a subset of connected blogs to other blogs to generate the fully connected polar blog graph. In [10], nodes represent posts in a directed graph and edges, hyperlinks connecting posts. Each post is analyzed using sentiment analysis techniques [8] and the goal is to determine how sentiment features of a post affect the sentiment features of connected posts and the structure of the network itself. In [19], the same approach is adopted, but specific questions are answered, like: how to identify features that lead to a sentiment propagation, how does the sentiment propagate, how fast, on the basis of which factors, how are the propagation speed variations connected to real world events, how does the role of the different individuals influence the propagation, etc. ?

4 SICOMORES' THEORETICAL BASES

4.1 Theories of Emotion

The Appraisal Theory of Emotion [13] postulates that, when a human being (or any living organism) lives, imagines or remembers a situation, they experience an emotion that results from the assessment of that situation according to a few cognitive criteria that can be classified into four families and answer specific questions:

- ***Relevancy***: Is the situation relevant to me, does it affect my well-being?
- ***Implications***: What are the implications of the situation and how do they affect my well-being and my short-term and long-term goals?
- ***Coping potential***: To what extent can I face the situation or adjust to its consequences?
- ***Normative significance***: What is the significance of the situation as regards my social norms and my personal values?

Scherer's version of the Appraisal Theory includes 16 specific criteria (Stimulus Evaluation Checks – SECs) that belong to the previous families. A combination of values of the criteria determines in a unique way a specific emotion, but the assessment of the different criteria is subjective. Thus, the same situation can trigger different emotions in people with different traits and coming from different cultures. Only the correspondence between a combination of values and a specific emotion is universal (*Universal Contingency Hypothesis*).

According to the Social Identity Approach [17], people categorize the others and themselves into social categories or groups defined by social criteria like age, religion or social status. The people who belong to the same category as an individual are called their ***ingroups*** and the others are called their ***outgroups.***

The Intergroup Emotion Theory [9] is defined in an intergroup context and suggests that the emotional experience of a person as a member of a social group is identical to the experience they live as an individual, as it is described in the Appraisal Theory. The only difference is that the Intergroup Emotion Theory implies the cognitive evaluation of a situation, that concerns the social identity of an individual (traits that connect the person to social groups) instead of involving their personal identity (the aspects that make the person unique). According to Garcia-Prieto and Scherer [5], the criteria that are sensitive to the social identity of a person are the ones that have a social connotation:

- Social goal conduciveness/obstructiveness (*Implications*),
- Agency/responsibility, action target (*Implications*),
- Control, power, adaptability (*Coping potential*),
- External standards (*Normative significance*).

For an individual to feel an intergroup emotion, the situation or the stimulus has to be relevant to the individual's social identity.

	Anger	*Guilt*
Social goal conduciveness	No	No
Action responsibility	Outgroup	Ingroup
Action target	Ingroup	Outgroup
Coping potential	High	Weak
Normative significance	Open	Immoral/Illegit.

Table 1. Examples of Emotion Definitions with Social Cognitive Criteria

4.2 Frijda's Laws of Emotion

An emotion is generally defined as "a subjective response to events that are important to the individual" [4]. Emotions are best characterized by two main dimensions: *arousal* and *valence*. The dimensions of valence ranges from highly positive to highly negative, whereas the arousal can be interpreted as the intensity. For Frijda, an emotional event generates a memory relative to the emotion felt by an individual during this event/situation, but here the situation itself is much less important than the emotion and the target of the emotion.

According to the *Law of habituation* [4], if one has often experienced an emotion towards someone during repeated

emotional events, then the next time an analogous emotion will occur, it will be less intense. It is the "repeated exposure to the emotional event" that accounts for habituation (*Law of Conservation of Emotional Momentum*). However, the *Law of Hedonic Asymmetry*, which highlights the different adaptation to pleasure or pain, states that the intensity of intense negative emotions seem not to diminish. The *Law of Comparative Feeling* expresses another interesting fact: "The intensity of an emotion felt during an event depends on the relationship between the event and some frame of reference against which the event is evaluated". The frame of reference is often the current situation, but it can also be an expectation, which is the case for relief and disappointment.

4.3 Propagation of Emotional Information in Social Contexts

According to [11], people are most willing to communicate social anecdotes that arise emotions and, as Rimé [12] reported, "The communicability of emotional social information is situated as some emotions are better able to increase communicability than others, and this varies with the identity of the audience". Several emotions selectively increase the communicability of social information: for instance, surprise and sadness only increase the communicability with friends (or ingroups), fear only with strangers (or outgroups). Guilt and shame are emotions that people keep to themselves and generally don't communicate.

4.4 Maslow's Pyramid and its Limitations

Maslow created a hierarchy of the human beings' needs, where the fundamental needs of a person (*physiological needs*: to eat, to drink, to sleep, etc., *security need*, *social needs*) are to be satisfied before the higher level ones (*need of esteem*, *need of self-accomplishment*). The fact that Maslow's pyramid was designed for Western countries has been underlined in several works, e.g. [15], where the author explains that both the hierarchy of priorities between the different needs and the needs themselves may differ between cultures. For instance, in an Asian country, interpersonal relationships and social interactions are more valued, on average, than self-accomplishment needs.

5 MODELING THE HUMAN TERRAIN IN SICOMORES

The Human Terrain consists of Social Agents. A Social Agent can be an individual who is part of one or several Sociocultural Groups (network(s)). A Social Agent can also be a Sociocultural Group like a Community Council, a religious network, an ethnic group, an NGO, a volunteer association, a group of interests, etc. Individuals and Sociocultural Groups are part of the population. The other Social Agents are local authorities, ONU Agencies, etc. Individuals are modeled as intelligent agents, Sociocultural Groups as groups of agents, whereas the other social agents are modeled as global social entities.

5.1 Individuals

Each individual is described by a set of attributes:
- *Social features*: age, gender, language, social status, religion, ethnicity, location, professional status, media (through which they can be reached: tracts, posters, newspaper ads, loudspeakers, radio, television, SMS and phone calls) and social goals.
- *Cultural features*: values, norms, artifacts, rituals, institutions, symbols.
- *Psychological features*: interests, vulnerabilities, types of needs, satisfaction degrees (in [-10, 10]) for each type of needs (according to Maslow's terminology). We will explain these notions in detail in the next section.

Cultural and some social and psychological features can be "factorized" in the description of Sociocultural Group(s) to which the individuals are linked.

Political, religious and other types of Sociocultural Group leaders are represented as particular individuals.

5.2 Sociocultural Groups

A sociocultural Group is a group of people recognized as such by its members and also by the other people, and is described by attributes specifying Social (including social goals), Cultural (Values, Norms, Artifacts, Rituals, Institutions and Symbols) and/or Psychological features. Let us specify the previous notions:

A *social goal* is any desired social reward (a positive outcome provided by and revered by a society) that one works toward, i.e. getting an education, obtaining a good job, getting married and having children, buying a nice car, even buying an Ipod can be considered a pop-culturally social goal.

A *norm* [20] is "a group-held belief about how members should behave in a given context. Sociologists describe norms as informal understandings that govern individuals' behavior in society, while psychologists have adopted a more general definition, recognizing smaller group units, like a team or an office, may also endorse norms separate or in addition to cultural or societal expectations. The psychological definition emphasizes social norms' behavioral component, stating norms have two dimensions: how much behavior is exhibited and how much the group approves of that behavior".

A *cultural artifact* is "an item that, when found, reveals valuable information about the society that made or used it. What is qualified as a cultural artifact? Burial coins, painted pottery, telephones or anything else that evidences the social, political, economic or religious organization of the people whom they belong to can be considered cultural artifacts" [21].

A culture's *values* are "its ideas about what is good, right, fair, and just. For example, American sociologist Robert K. Merton suggested that the most important values in American society are wealth, success, power, and prestige" [24].

A ritual "is a sequence of activities involving gestures, words, and objects, performed in a sequestered place, and performed according to set sequence". Rituals may be prescribed by the traditions of a community, including a religious community. Rituals are characterized by formalism, traditionalism, invariance, rule-governance, sacral symbolism and performance.

Rituals of various kinds are a feature of almost all known human societies, past or present. "They include not only the various worship rites and sacraments of organized religions and

cults, but also the rites of passage of certain societies, atonement and purification rites, oaths of allegiance, dedication ceremonies, coronations and presidential inaugurations, marriages and funerals and son on. Many activities that are ostensibly performed for concrete purposes, such as jury trials, execution of criminals, and scientific symposia, are loaded with purely symbolic actions prescribed by regulations or tradition, and thus partly ritualistic in nature. Even common actions like hand-shaking and saying hello may be termed rituals" [22].

Cultural *institutions* are "elements within a culture/subculture that are perceived to be important to, or traditionally valued among its members for their own identity. Examples of cultural institutions in modern Western society are museums, churches, schools, work and the print media. "Education" is a "social" institution, "post-secondary education" is a cultural institution, "high-school" is an instantiation of the institution within America [23]".

To the human mind, *symbols* are "cultural representations of reality". Every culture has its own set of symbols associated with different experiences and perceptions. Thus, as a representation, a symbol's meaning is neither instinctive nor automatic. The culture's members must interpret and over time reinterpret the symbol. Symbols occur in different forms: verbal or nonverbal, written or unwritten. They can be anything that conveys a meaning, such as words on the page, drawings, pictures, and gestures.

We intend the notion of *vulnerabilities*, as people's weaknesses regarding different aspects:
- *Commerce/Economy*: financial situation, commerce, industry, etc.
- *Resources*: food, arms, money, oil, etc.
- *Critical needs*: hunger, thirst, care, rest, security, etc.
- *Infrastructures*: health, communications, energy, water, transport, etc.
- *Emotional aspects*: frustration, isolation, fear, anger, etc.
- *Organisational aspects*: alliances, loss of an expert, international dissents, structural weaknesses, limitations, etc.

For each Sociocultural Group is defined a particular Maslow's-like pyramid with specific types of needs to which is associated a given respective importance.

For a given Sociocultural Group, to each specific value of the attributes mentioning Cultural features, social goals and types of need is associated the quantified (between 0 and 10) importance/typicality of that particular element for the Sociocultural Group.

The different Sociocultural Groups are organized within a hierarchy of power. Sociocultural Groups are networks, as long as their members interact with each other.

Various links may connect the members of a Sociocultural Group (e.g. religious link or family link). Some Sociocultural Groups are temporary, for example the group of people working on a Civil-Military project or the group of people gathered together at a periodic market.

6 PSYOP CHARACTERISTICS AND MODELING IN SICOMORES

For a PSYOP, a group of individuals called the direct *info-targets* is defined by means of social and/or psychological criteria which allows to find out their membership Sociocultural Group(s) and assign them cultural features and social goals. A message is then spread out to them on a specific area, depending on the scope of the means of conveyance the Forces are using and the individuals' receptivity to this means (e.g. individuals must have a radio to be reached by a message broadcasted on the radio). After the message has reached the direct info-targets, the latter will propagate to the *indirect info-targets* the content of the message. Given that SICOMORES is meant to simualte the propagation of PSYOP effects through the population structured within Sociocultural Networks, the user of the system must provide some general information concerning the PSYOP that is the input: date, effect desired by the military, direct info-targets, used mean of conveyance, means of conveyance scope, theme of the message (religious, political, etc.). Moreover, given that we don't use image recognition, nor spoken language or text semantic analysis, we expect the user to directly give some characteristics of the information conveyed by the message whatever its form (video clip, radio or television program, speech, image, text) and we assume that it is the description of an action or an event such that the agent and the target of the action are Social Agents. This action/event gives rise to a situation described as follows:

- *Relevancy:* list of the Sociocultural Groups to which the situation is relevant.

- *Goal facilitation/obstruction*: set of tuples (Social goal, "favored", concerned Sociocultural Group) or (Social goal, "obstructed", concerned Sociocultural Group).

- *Causal Agent, Action Target*: Social Agent who performs the action that gives rise to the situation and Social Agent who is the target of the action.

- *Coping potential*: set of tuples (Sociocultural Group or leader, value in {low, medium, high}). The Coping Potential of each Sociocultural Group is globally assessed by the user.

- *Sociocultural Elements*: set of tuples (Sociocultural Group, "flouted" or "accentuated" or "obstructed" or "favored", sociocultural characteristic) (see next section).

- *Need Satusfaction or Dissatisfaction*: set of tuples (type of needs, Sociocultural Group or leader, positive or negatve satisfaction degree). The types of needs are by default "Physiological Needs", "Security need", "Social Needs", "Need of Esteem", "Need of Self-Accomplishment" [15], but may be replaced by other types of needs specific to a given culture. The satisfation degree ranges between 0 and 10.

To provide these pieces of knowledge, the user is guided. For each Sociocultural Group concerned by the situation, they can display the name of its possible leader and all the social, cultural and psychological characteristics of the group as well as the hierarchy of power. The information provided by the user will

help the system assess the cognitive criteria mentioned in the section presenting the Intergroup Emotion Theory in order to determine the emotion triggered by the given message.

7 Effect Generation of a PSYOP

7.1 General Scheme

A direct info-target receives a message and feels an emotion related to the information conveyed by this message, according to the iNtergroup Emotion Theory. Their well-being may also be affected by the action/event described by this message, the notion of well-being representing the satisfaction/ dissatisfaction of the info-targets' needs. The direct info-targets then propagates the information to the indirect info-targets who, in turn, propagate it. An Info-target decides to propagate an information only if they judge it interesting enough. In that case, the choice of the people to whom it is propagated depends on that emotion generated in the emitter in accordance with what was mentioned in section III concerning the type of people to whom emotional information is propagated. It is the information that each info-target receives that determines their own emotion and well-being, not the emotion of the emitter of the information. It is important to notice that all the individuals who are members of the same Sociocultural Group experience the same emotions (as we will see later, their intensity may vary though) and feel the same well-being.

We will first explain how he arousal of an emotion determined by the Intergroup Emotion Theory is computed, then adjusted due to prior experiences and the strength of the concerned message. We will then show how the well-being of an info-target is computed. The notions of interest of an information and unexpectedness of a situation will be defined and quantified.

Finally, we will specify the conditions under which the propagation of a message stops.

7.2 Computation of the Arousal of an Emotion Determined According to the Intergroup Emotion Theory

Let Sc be the Sociocultural Group of an ndividual who must assess a situation.

As we saw in section 5.2, a Sociocultural Group in SICOMORES is defined, among other characteristics, by social goals, values, norms, artifacts, rituals, institutions and symbols and each value for these characteristics is weighted by its importance/typicality for the group.

Let FlValues, FlNorms, FlArtifacts, FlRituals, FlInstitutions and FlSymbols be the respective sets of the values of the attributes Values, Norms, Artifacts, Rituals, Institutions, Symbols for the group Sc, that represent cultural elements **flouted** in the situation. Let $\text{imp}(\text{fv}_1)$, …, $\text{imp}(\text{fv}_{\text{Card(FlValues)}})$ be the respective importance of the values $\text{fv}_1,\ldots, \text{fv}_{\text{card(FlValues)}}$ of FlValues.

We define analogous notations for FlNorms, …, FlSymbols.

Let FrGoals be the values of the attribute Social Goals, that represent goals **favored** in the situation. Let $\text{imp}(\text{fsg}_1),\ldots, \text{imp}(\text{fsg}_{\text{card(FrGoals)}})$ be the respective importance of the values $\text{fsg}_1,\ldots, \text{fsg}_{\text{card(FrGoals)}}$ of FrGoals.

Let ObGoals be the values of the attribute Social Goals, that represent goals **obstructed** in the situation. Let $\text{imp}(\text{osg}_1),\ldots, \text{imp}(\text{osg}_{\text{card(ObGoals)}})$ be the respective importance of the values $\text{osg}_1, \ldots, \text{osg}_{\text{card(ObGoals)}}$ of ObGoals.

Let AcValues be the values of the attribute Values, that represent cultural values **accentuated** in the situation. Let $\text{imp}(\text{av}_1),\ldots, \text{imp}(\text{av}_{\text{card(AcValues)}})$ be the respective importance of the values $\text{av}_1,\ldots, \text{av}_{\text{card(AcValues)}}$ de AcValues.

- ***If the valence of the emotion is negative***, the factors influencing its arousal are the importance of the breakings, if the emotion is mainly caused by a lack of respect towards some sociocultural elements, and the social goals that are obstructed. In case of incompatibility of the situation with sociocultural characteristics and/or the obstruction of social goals of the info-targets' Sociocultural Group, the arousal of the emotion is more or less intense depending on the importance of the concerned characteristics. For instance, if the situation goes against an important moral value, the emotion will be more intense than if another characteristic is involved.

We define the emotion Arousal Increase Factor AIF (with the previous notations, AIFNorms, …, AIFObGoals being defined in an analogous way as AIFValues):

$$\text{AIF} = (\text{AIFValues} + \text{AIFNorms} + \text{AIFArtif.} + \text{AIFRituals} + \text{AIFInstit.} + \text{AIFSymb.} + \text{AIFObGoals})/70 \quad (1)$$

$$\text{AIFValues} = \sum_{i=1}^{\text{card(FlValues)}} \text{imp}(\text{fv}_i) \Big/ \text{card(FlValues)} \quad (2)$$

- ***If the valence of the emotion is positive***, the factors influencing its arousal are the respective importance of the social goals that are satisfied and the respective importance of the cultural values that are accentuated in the situation.

$$\text{AIF} = (\text{AIFFrGoals} + \text{AIFAcValues})/20 \quad (3)$$

$$\text{AIFFrGoals} = \sum_{i=1}^{\text{card(FrGoals)}} \text{imp}(\text{fsg}_i) \Big/ \text{card(FrGoals)} \quad (4)$$

$$\text{AIFAcValues} = \sum_{i=1}^{\text{card(AcValues)}} \text{imp}(\text{av}_i) \Big/ \text{card(AcValues)} \quad (5)$$

In both cases (negative or positive emotion), the arousal of the emotion is then defined as follows (it varies between 0 and 1):

$$\text{A} = (\text{AIF} + 1) \times 0.5 \quad (6)$$

7.3 Adjustment of the Arousal of an Emotion Due to Prior Experiences

7.3.1 Emotional Memory Databases

To every Sociocultural Group is associated a database of emotional memories. Each emotional memory is defined by a tuple (emotion, arousal, target of the emotion). Let's underline the fact that an emotional memory is the trace of a

dated emotion towards a Social Agent, like Frijda's emotional event. The situation deriving from a PSYOP message, that has caused the occurrence of the corresponding emotion is not stored. Every time a new PSYOP triggers a new emotion, the corresponding emotional memory is stored in the Sociocultural Group's memory database.

7.3.2 Taking into Account of Frijda's Laws

After the determination of an emotion triggered by a psychological message, its arousal is computed as shown in section 7.2 and then adjuted by taking into account Frijda's Laws.

• **Law of Habituation and Law of Hedonic Asymmetry:** If a positive emotion or a negative emotion the arousal of which is higher than a given threshold occurs repeatedly towards a Social Agent, the absolute value of the arousal of this emotion decreases each time, which is not the case for very negative emotions the arousal of which does not change. The decreasing factor is set to a value α (to be adjusted during experimentation).

• **Law of Comparative Feeling:** If several (at least 2) consecutive emotional memories of the same valence have occurred towards a Social Agent and a new emotion towards the same Social Agent appears with the opposite valence, then the absolute value of the arousal of the latter is increased. The increasing factor is set to a value β (to be adjusted during experimentation).

An emotional event which triggers an emotion with an absolute value of its arousal lower than a certain threshold, will not be stored in the concerned Social Agent's memory database.

7.4 Adjustment of the Arousal of an Emotion Due to the Strength of a Psychological Message

The arousal of an emotion computed in both previous steps, is then adjusted again by taking into account the strength of the message. The strength of the message to be propagated (SMP) depends on the previous strength of the message (SM), the credibility of the emitter and a Boolean, EQTL, equal to 1, if the theme of the message is identical to the type of link that connects the emitter of the message and the receiver, to 0 otherwise. The credibility of any Social Agent to the eyes of each Sociocultural Group or leader is predefined (between 0 and 1). It is equal to 1, if the sender and the receiver belong to the same Cultural Group. Initially, the strength of the message propagated by the direct info-targets is equal to 1, otherwise it ranges between 0 and 1.

$$SMP = (EQTL + (1 - EQTL) \times 0.8) \times Credibility \times SM \quad (7)$$

Then the final value of the arousal of the emotion is:

$$\begin{aligned} A_f &= A \times \alpha \times SMP \quad \text{or} \\ A_f &= A \times \beta \times SMP \quad \text{or} \\ A_f &= A \times SMP \end{aligned} \quad (8)$$

7.5 Interest of an Information

Following the Simplicity Theory [1], the interest of an individual in an information is the interest in the event/situation that the information describes or implies and is quantified as the sum of its *unexpectedness* and the arousal of the emotion it causes in this individual:

$$I = U + A_f \quad (9)$$

We define a situation caused by a PSYOP as *unexpected*, if some elements of the situation do not correspond to the sociocultural characteristics of the people involved in it. These elements are:

- the norms and rituals characterizing the people's Sociocultural Group(s),
- the fact that the situation does not respect the hierarchy of power between the Sociocultural Groups in the concerned society.

Let PowerHierarchy be equal to 10, if the hierarchy of power is respected and 0, otherwise. The Unexpectedness is defined between 0 and 10 as follows:

$$U = (AIFNorms + AIFRituals + 10 - PowerHierarchy)/30 \quad (10)$$

7.6 Degree of Well-being Generated by a Message

With the same notations as in the previous sections, let $imp_{N1},\ldots,$ imp_{Nn} be the respective importance of the different types of needs (in [0,1]) defined for a Sociocultural Group Sc and $d_{s1},\ldots,$ d_{sn} their respective satisfaction degrees for the group Sc in the concerned situation, the global degree of well-being of Sc's members in the situation is computed as follows (in [-10, 10]):

$$\sum_{i=1}^{n} imp_{Ni} \times d_{si} \Big/ 10 \times n \quad (11)$$

7.7 End of the Propagation of a Message

Three conditions can cause the partial end of the propagation process:
- the individual who just received the message does not have enough interest about it to transmit it (the interest falls below a certain threshold),
- the strength of the message to be propagated falls below a given threshold,
- if an ndividual is connected to another one in a temporary network and the link is not activated during the propagation, then the propagation stops along this branch.

The complete end occurs if it has been a long time (higher than a given threshold) since the message was spread by the Armed Forces.

The different thresholds are to be defined during the experimentation.

8 CONCLUSION AND FUTUR WORK

We have presented some aspects of SICOMORES, a decision support system intended to simulate the effects of influence operations on the population structured in sociocultural networks, in the framework of asymmetric conflicts. We have focused of the description of a method meant to determine the effects of an emotion-triggering Psychological Operation on the population, based on theoretical works stemming from the Psychology of Emotion and from Social Psychology.

The next step of our work will be to validate our model. A realistic population will be generated using an algorithm that takes into account the sociocultural characteristics of the concerned country [2] and the sociocultural data will be extracted from [18]. The future interface will allow to display "maps of emotion" and well-being indicators for each Sociocultural Group.

ACKNOWLEDGEMENTS

This work is funded by the French Ministry of Defense (DGA – Direction Générale de l'Armement) in the framework of DGA RAPID Project SICOMORES.

REFERENCES

[1] N. Chater and P. Vitanyi. Simplicity: a Unifying Principle in Cognitive Science ? *Trends in Cognitive Sciences,* 7(1):19-22 (2003).

[2] C. Faucher. *An Algorithm for Generating a Realistic Population Based on Sociocultural Characteristics*, unpublished, March (2014).

[3] Headquarters, Department of the US Army, FM 3-05.30, *Psychological Operations* (2005).

[4] N. H. Frijda. The Laws of Emotion, *American Psychologist,* 43(5):349-358 (1988).

[5] P. Garcia-Prieto and K.R. Scherer. Connecting Social Identity Theory and Cognitive Appraisal Theory of Brown and D. Capozza, Eds., *Social Identities: Motivational, Emotional and Cultural Influences*, Psychology Press (2006).

[6] A. Kale. *Modeling Trust and Influence on Blogosphere using Link Polarity*, Master Thesis, University of Maryland, USA, Department of Computer Science and Electrical Engineering (2007).

[7] W. Lee, P. Sungrae, and I.-C. Moon. Modeling Multiple Fields of Collective Emotions with Brownian Agent-Based Model, In *Proceedings of the Fourteenth International Conference on Autonomous Agents and Multi-Agent Systems,* 797-804 (2014).

[8] B. Liu. *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publisher (2012).

[9] D.M. Mackie, T. Devos, and E.R. Smith. Intergroup Emotions:
Explaining Offensive Action Tendencies in an Intergroup Context. *Journal of Personality and Social Psychology*, 79(4):602-616 (2000).

[10] M. Miller, C. Sathi, D. Wiesenthal, J. Leskovec, and C. Potts. Sentiment Flow Through Hyperlink Networks. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (2011).

[11] K. Peters, Y. Kashima, and A. Clark. Talking About Others: Emotionality and the Dissemination of Social Information. *European Journal of Social Psychology*, 39:207-222 (2009).

[12] B. Rimé. *Le partage Social des Emotions*. Presses Universitaires de France (2005).

[13] K. R. Scherer, A. Schorr, and T. Johnstone, Eds. *Appraisal Processes in Emotion.* Oxford University Press (2001).

[14] F. Schweitzer, D. Garcia. An Agent-Based Model of Collective Emotions in Online Communities, *European Physical Journal B*, 77(4): 553-545 (2010).

[15] G. Senez. *Maslow's Pyramid of Needs and the Asian Equivalent*, http://garettsenez.blogspot.fr/2011/07/maslows-hierarchy-of-needs-and-asian.html (2011).

[16] H. Tajfel. Social Identity and Intergroup Relations, *European Studies in Social Psychology*, 7, Cambridge University Press (2010).

[17] J.C. Turner, M.A. Hogg, P.J. Oakes, S.D. Reicher, and M.S. Wetherell. *Rediscovering the Social Group: A Self-Categorization Theory*, Basil Blackwell, New York (1987).

[18] W.D. Wunderle. *Through the Lens of Cultural Awareness: A Primer for US Armed Forces Deploying to Arab and Middle Eastern Countries* (KS 66027, UA26.A2W37), Combat Studies Institute Press, Fort Leavenworth (2006).

[19] R. Zafarani, W.D. Cole, and H. Liu. Sentiment Propagation in Social Networks: A Case Study in LiveJournal. In s.-k. Chai, J.J. Salerno, and P.L. Mabry, Eds., LNCS, vol. 6007, *Advances in Social Computing*. Bethesda, MD: Springer (2010).

[20] Wikipedia, Social Norm, http://en.wikipedia.org/wiki/Norm_social

[21] Education Portal: Cultural Artifact, http:// education portal.com/academy/lesson/cultural-artifact-definition-examples-quiz.html#lesson.

[22] Wikipedia: Rituals, http://en.wikipedia.org/wiki/Ritual

[23] Wikipedia: Cultural Institution, http://en.wikipedia.org/wiki/Cultural_institutions

[24] Cliffnotes: Cultural Values, http://cliffnotes.com/sciences/sociology/culture-and-societies/cultural-values.

Collective Cognition and Distributed Information Processing from Bacteria to Humans

Alexander Almér[1], Gordana Dodig-Crnkovic[1] and Rickard von Haugwitz[1]

Abstract. The aim of this paper is to propose a general info-computational model of cognition that can be applied to living organisms from the level of a single cell´s cognition to the level of groups of increasingly complex organisms with social, distributed cognition. We defend the project of new cognitivism, which unlike the old one acknowledges the central role of embodiment for cognition. Information processing going on in a cognising agent range from transduction of chemical signals and "quorum sensing" in bacteria, via simple local rules of behaviour that insects follow and that manifest themselves as "swarm intelligence", to human level cognition with full richness of human languages and other systems of communication.

1 INTRODUCTION

The smallest living organism is a single cell. It is upholding its existence through interchanges with the environment, by means of energy- and information processing. The central insight in cognitive sciences that we build our framework upon, was made by Maturana and Varela (1980) who recognised that cognition and process of life are synonymous:

"Living systems are cognitive systems, and living as a process is a process of cognition. This statement is valid for all organisms, with or without a nervous system." (Maturana & Varela, 1980: 13)

If we want to study processes and structures of cognition, it is necessary to start by studying organisation of life. The fundamental empirically established property of living systems is that their structures and processes are hierarchically organised. Those structures and dynamics can be modelled computationally as agency-based hierarchies of levels (Dodig-Crnkovic 2013).

The capability of living cells to receive signals from the environment and act adequately upon them is fundamental to life. Information is communicated in a biological system both bottom-up (from input signals up) and top-down (from decision making down) in a circular motion. The lower/basic levels of cognition sort and propagate incoming perceptual information and forward the transduced information to higher levels for a more complex processing.

Here is the detailed description how the process of biological information transduction (transformation) goes on in a cell as fundamental living/cognising unit:

"Bacterial cells receive constant input from membrane proteins that act as information receptors, sampling the surrounding medium for pH, osmotic strength, the availability of food, oxygen, and light, and the presence of noxious chemicals, predators, or competitors for food. These signals elicit appropriate responses, such as motion toward food or away from toxic substances or the formation of dormant spores in a nutrient-depleted medium." (Nelson and Cox 2008:419)

So information for an organism comes in different forms (such as hormones, pheromones, photons (sunlight), changes in some state like acidity or concentrations of glucose and ions such as K+, or Ca2+ in the environment, heat, cold, osmotic pressure, etc.), while receptors of information transduce information for further processing in the cell, transforming input signals into intracellular signals. This involves the same type of molecular processes as metabolism: production and degradation of substances, stimulation or inhibition of chemical reactions, etc.

"In all these cases, the signal represents information that is detected by specific receptors and converted to a cellular response, which always involves a chemical process. This conversion of information into a chemical change, signal transduction, is a universal property of living cells.

The number of different biological signals is large, as is the variety of biological responses to these signals, but organisms use just a few evolutionarily conserved mechanisms to detect extracellular signals and transduce them into intracellular changes." (ibid)

Even though there are many different kinds of signals, basic mechanisms for their transduction are preserved in different signalling pathways. The process of signals transduction (information processing) that provides information transfer in the cell goes on in parallel with cell metabolism that is handling mass/energy transfer. The two processes constrain each other.

2 OLD DISEMBODIED AND NEW EMBODIED COGNITIVISM

The cognitive process presupposes *attention* that enables *information input*, *sensory memory* (allowing an agent to retain impressions of sensory information after the stimulus has gone), *working memory* for actively manipulating information, and *long-term memory* for preserving information so that it can be reused. The process results in decision-making that will affect actuators. An active loop is sustained between inputs from the environment, internal information processing and actuators, which enable organism's response to the environmental inputs.

This view of cognitive processes is different from classical cognitivism in the first place because for old cognitivists,

[1] Department of Communication and Cognition,
Chalmers University of Technology and Gothenburg University, Sweden
Email: alexander.almer@ait.gu.se, dodig@chalmers.se, rickard.von.haugwitz@gu.se

cognition was taken to be a purely intellectual activity of humans. (Scheutz 2002)

The first attempts in 1950s to recreate mind "in silico" as an "electronic brain" without a body, by simply filling an existing digital programmable computer with data failed, as computers at that time had very limited resources – both speed of information processing and memory, apart from the basic fact that they were isolated from the environment and without any adaptive or learning capacities.

The lesson learned from early computationalism was that the brain, in order to function intelligently, cannot be isolated as a "brain in a vat", but must have a body to provide a connection to the environment and thus a source of novel input and learning. After the experience with IBM's Watson machine it may seem that bodily experiences from the interaction with the world could be replaced with the data input provided by the Internet with its open and learning structures. If intelligence is defined as a capacity to successfully process different kinds of information and adequately act upon it, no isolated computers can be expected to be cognitive or intelligent. Instead, robots are being developed as adaptive and learning systems with an ambition to reach in the future the level of general intelligence through a process of adaptation and learning.

In spite of the current impressive progress of computing machinery in performing cognitive and intelligent tasks such as different kinds of machine learning, automatic image and speech recognition, language processing, audio recognition and speech generation, etc., there is still a strong resistance among philosophers of mind to acknowledge that more advanced models of the info-computational nature of cognition do not suffer from the same limitations and problems as the old cognitivism as they embrace both embodiment and embeddedness of info-computation as *conditio sine qua non* of cognition (Scheutz 2002).

The resistance to natural info-computational cognitivism persists although life sciences as well as human, social and behavioural sciences could potentially gain immensely from a general comprehensive definition of cognition that would capture their pre-theoretic overlap at a basic theoretical level, distinguishing it from pure physical information processes in general. Such basic theory integration would eventually have to meet scientific needs of facilitating e.g., explanations of unexplained phenomena in the relevant domains, as well as more comprehensive interpretations. Also, it could be the basis for research and modelling of relations between domains of e.g., biology, psychology, behavioural- and social sciences. The model here proposed must in the end be tested against its capacity to contribute to such goals.

We see cognition as a natural phenomenon, an entirety of information processes in a living organism, organized in hierarchical levels, that meets given evolutionary constraints (Dodig-Crnkovic 2008, 2012, 2013, 2014). Our basic definition of cognitive information processing refers to evolutionary selected mechanisms for information-based production of an organism's activities.

Unpacking the notion of activities being guided by information, we employ a naturalised framework of representation (cf. Almér 2007, Millikan 2004, Dodig-Crnkovic 2008); where representation is defined as something (such as a symbol, or a structure) that stands in place of something else.

3 COLLECTIVE BEHAVIOUR IN LIVING ORGANISMS

Adopting the social ontology proposed by Almér and Allwood (2013), we characterise types of organism-collective activity based on type, complexity, and awareness of represented information. We build the naturalist framework for cognition with the elements from a naturalised perspective on representation (e.g., Almér 2007; Millikan 1984, 2004, Neander 2006, Dodig-Crnkovic 2008) based on the discourse of natural computation within the info-computational approach of Dodig-Crnkovic (2014).

Before moving on, some core notions will be briefly introduced. First, we make use of the notion of living organism in our definition of cognitive information processing. By living organism we refer to:

a) Selected for, co-adapted and co-reproduced system of mechanisms globally selected for; function of which is the survival and reproduction of its genetic type

b) Instance of the above in a normal environment with sufficiently normal processes for survival and reproduction up and running.

This characterisation of living organism relies on the notion of biological function and normal conditions. There are two main approaches to functions in biology. One is the causal-role or causal disposition perspective, originating from Robert Cummins' (1998) work, ascribing functions to components in larger systems based on the components' actual dispositions to causally contribute to some set of capacities ascribed to the whole system. The global capacity of the system is identified with a set of actually produced effects or with a set of actual dispositions of the system to produce such effects under specific conditions. We call functions as conceived of in terms of systems' capacities 'systemic functions'.

A second notion of function is backward-looking, identifying a systems' function with some set of historical effects of its predecessors. Millikan (1984) stands for the most developed version of this type of functional theory. Davies (2000) gives a definition of selected function through conditions that describe the mechanisms of natural selection, the evolutionary outcome of the operation of those mechanisms, the purported normative aspect of functional properties by imposing a role of performance on items previous conditions. For a discussion of various attempts to understand function in biology, see e.g. (Almér 2003).

With a selected-effects characterisation of function we can distinguish between proximate function and distal function – the former being what a mechanism is selected for: A human heart, e.g., contributes to the blood being oxygenated, but its proximate function is to pump blood, while the lungs are directly involved in effecting oxygenation.

Thus we also define the notion of proximate effect. It is the effect of a mechanism directly realising a proximate function, described without reference to the function. An example would be the chameleon's skin, which can change colour – the proximate effect – and thereby function as a social signal, camouflage, or thermal regulation.

4 REPRESENTATION IN HUMANS AND OTHER LIVING ORGANISMS

We indicated above that cognitive information processing is an activity-guiding process in living organisms. One way of framing such claim would be in terms of representation (as in mental and linguistic representation in humans and some animals, or in exchange of physical objects such as molecules or ions in simplest organisms like bacteria, where "language" consists of chemical exchanges governed by much simpler rules than human languages).

Briefly, by a representation we refer to signs co-developed with sign users, which might carry information but could also misrepresent facts, that is, they can be false. (Millikan 2004, Neander 2006). By framing cognitive information processing in this kind of evolutionary framework tied to a corresponding notion of representation, a subset of information processes is selected as bearing particular significance, namely those also giving rise to representation representing something to someone. Note that the notion of falsity does not apply to information *sui generis* that is by (Dodig-Crnkovic 2010) defined as proto-information or intrinsic information as the fabric of reality for an agent.

We must distinguish between what we could call complete correctness conditions for a representation and the part of those conditions which are explicitly codified by the structure of the representation in a way which the system using that representation is adapted to interpret. This pertains what information is accessible to such a user and in what manners it could be used for processing. Almér and Allwood (2013) expressed similar ideas in terms of "complexity of information" distinguishing between representational capacities in terms of degrees of awareness and explicitness of representation. Notice that a false representation carry natural information about the world in the very same manner as a true representation, whereas merely the latter is such that a normal interpreter gets access to the explicitly represented information (corresponding to the sign's correctness conditions) by way of the normal interpretation procedure. It is important to keep apart the notion of correctness condition from the notion of information, although there is a conceptual link in our view as just indicated.

Talking about human-level cognition, much discussed in the fields of pragmatics and philosophy in general is the interplay between contextual parameters and syntactically encoded semantic information in the interpretation of natural language expressions. For an overview of such issues, see (Almér 2007). Take the sentence "it's raining". An instance of an utterance of that sentence type typically "refers" to a particular rain event with a reasonably well-defined location in time and space, whereas the surface structure of the sentence does not seem to encode for location. The million-dollar question, perhaps somewhat surprisingly, is considered to be whether the deep structure of this sentence type contains a hidden variable or parameter for location. Let's assume it doesn't. Then we would have an instance of a representation where the location would be part of the complete correctness conditions while not being explicitly encoded in the sign.

What about awareness? On the human level, organisms are obviously capable of being aware where it rains and normally apply the mentioned sentence type with an intended place in mind. As such, awareness does not automatically connect to the structure of a representation. But we could imagine a cognitive system employing a signal type for rain here and now without being capable of explicitly representing time and location at all. Such an organism could not use their cognitive system to store information about where or when anything happens, like a rain event there and then. Still the time and location of rain would be part of the correctness conditions of a sign, and part of the natural information a true sign of that type carried for a typical user.

We, on the other hand, could use the signs of that organism as natural sign for time and location of rain. Millikan (2004) makes similar points about signs and their conditions of truth in terms of the signs' "articulation". She refers to simple warning signals in the animal kingdom as possible examples of signs not articulating time and location while obviously standing for reasonably well-defined time and location values.

Miłkowski and Talmont-Kaminski (2015) refer to the work of Gładziejewski, who distinguishes between *action-oriented accounts* of representations, characteristic of interactivism (Bickhard, 2008), and the *structural account of representation*, such as proposed by (Ramsey, 2007). They also present results of Clowes and Mendonça regarding the role of representation in embodied, situated, dynamicist cognition, claiming that in several contexts the notion of representation is useful, such as in re-use, fusion and elaboration of information; virtualist perception as well as operations over representations – extension, restructuring and substitution. The role of representation is found in informational economy (more compact manipulation of information) and better understanding of the coupling between the organism and the world. This would mean that the idea of representation in explanation has not become obsolete in enactive and radical embodied theories of cognition.

Traditional approaches to social cognition in humans are well researched compared to animal cognition and to even more scarce sources on the social behaviour and languaging (the cognitive process of developing meaningful output as part of language learning) of unicellular organisms or plants. In spite of the abundant literature and dominant position of the studies of human social cognition, it is important to understand the limitations of approaches to collective intentionality based exclusively on human language and rationality. They are expressed mainly in descriptive, external terms while we need to expand the notion of social cognition to include an embodied, evolutionary, generative approach in all living organisms.

Thus, returning to the question of roots of human representation, we are studying simple organisms interacting with their environment. For understanding them it is important to learn about what type of information (symbolic or sub-symbolic e.g.) as well as what kind of agent (its cognitive info-computational architecture) it is. Of special interest is as well how information is stored. For example, in the case of unicellular organisms it could be stored in the DNA or other cell structures, while in the case of more complex organisms specialised structures such as nervous systems or brains are used for information storage together with other bodily structures, as the body frames the way of agency and thus cognition.

It is important to understand how retrieval of information is enabled, as well as transduction and processing; whether the organism acts completely automatically upon getting information or it can make decisions, reason or plan activities related to that information; whether that information can be

implicitly or explicitly synthesised with other information, and so on.

5 SOCIAL COGNITION, FROM BACTERIA TO HUMANS

With respect to signalling, in the simplest type of collective activity no social signalling (based on type of information processed) is taking part, nor are the organisms conscious of the purpose (evolutionary framed) of their own activities. However, the criterion in this model for an activity to be collective is defined in terms of the function of information-guided actions such that collective activities require contributions from more than one organism for the function to be performed. The collective function is performed without any social signalling, solely depending on mechanisms such as stigmergy, that is indirect, mediated coordination. An example of such coordinated behaviour is that in deep snow people would follow the common path, as it is easier, so collective behaviour will emerge without direct communication, constrained by the interaction with the environment affected by other people.

Thinking about signalling in the case of community of living agents exchanging "messages" we start with the cognitive level of bacteria that are both the simplest kind of organisms and their signalling is simple exchange of chemical molecules. Actually, a single bacterium itself is not so simple when it comes to internal signalling as it may seem. A bacterium is a complex network of functional cooperating parts that orchestrate their mutual interactions, led by chemical and physical exchanges and interactions with the environment. It has been shown (Ben-Jacob, Becker, & Shapira, 2004; Ben-Jacob, Shapira, & Tauber, 2006; Ben-Jacob, 2008) (Ng & Bassler, 2009) that bacterial collectives such as colonies, films and swarms exhibit advanced social cognitive behaviours like "quorum sensing" based on communication between individual bacteria using chemical "language". Bacteria have shown surprising ability to find good strategies to survive under different pressures and to develop defence mechanisms such as anti-biotic resistance.

As an example of the next level of distributed cognition we consider insects such as ants. While an ant colony as a whole is able to efficiently find the shortest path to a food source, individual ants, although capable of learning (Dukas, 2008), do not display the same level of optimisation. Simple behaviour on an individual level gives rise to a more efficient form of learning on a higher level of societal organisation.

Likewise, a slime mould consisting of a colony of unicellular amoebae can "learn" the shortest path to food and exhibit remarkably efficient collective behaviour, despite every single member of the colony lacking any necessary faculty for planning (Nakagaki, Yamada, & Tóth, 2000).

In more complex organisms, however, planning and learning become increasingly evident on an individual level, while in a social setting coordination similarly takes a more long-term form. The behaviour of the organism, then, must be regulated in order to optimise future payoff according to some utility function. Importantly, as the complexity of the organism increases, so does its perceived environment. While an amoeba may be aware of little more than intensity of light and the concentration of sugars around it, and indeed may not need be aware of much more than that, a hare relies on scent, hearing and vision, among other senses, coupled with previous experience to find food and detect predators, which in turn need to employ non-trivial planning based on some learning process in order to catch it. The central mechanism underlying this behaviour is generative – from simple local rules, a global collective pattern emerges (Marsh and Onof 2007).

Social interaction is arguably the largest contributing factor in adding complexity to an environment. Game theory tells us that in an adversarial multi-player game, in most cases an optimal strategy is random (or mixed), and depends on the strategy of the opponents, who may also change their strategies at any time. In such an environment, the dynamics of which are likely to change over time, but where courses of actions nevertheless are dependent on the situation that may need to be analysed in terms of their long-term effects, not only learning becomes crucial, but also a mechanism for modulating learning and behaviour.

Since not all events are equally important in the learning process – one may not get a second chance to learn to escape a lion, for example – the learning rate should be lowered or raised accordingly to reflect this. Likewise, while escaping said lion the long-term implications of one's actions, such as whether running to the left increases or decreases one's chances of finding dinner for the evening, is rather less important than minimising the short-term prospects of ending up as a dinner oneself. The trade-off between exploration and exploitation needs to be struck differently depending on the current environment in much the same way.

It has been suggested by Doya (2000, 2002), following the work of Montague et al. (1996) and Schultz et al. (1997), among others, that the neurotransmitters dopamine, serotonin, noradrenaline and acetylcholine are responsible for the modulation of learning parameters in the brain. Specifically, within the framework of reinforcement learning, the reward system, mainly dopamine, has been shown to correspond to the temporal-difference error, which tells the learning agent how the received reward differs from the expected reward. Serotonin controls the discounting factor, which sets the time horizon of optimisation; noradrenaline determines the level of exploration versus exploitation via the inverse temperature parameter; and acetylcholine regulates the learning rate, that is, how much weight to assign to observed events.

As the signal substances controlling learning have also been shown to cause the physiological and psychological effects associated with emotion in humans, it may be posited that emotion evolved precisely in order to facilitate adaptive learning and behaviour in a complex, non-stationary environment (von Haugwitz et al., 2012). Fear, for example, would serve to lower the discounting factor, making the organism focus on escaping immediate danger, while comfort on the other hand allows for long-term planning.

Humans are on the highest level of hierarchy of social signalling systems. The social ontology framework by Almér and Allwood (2013) has been developed largely as a response to certain philosophical suggestions that social ontology should be understood in terms of what has been called collective intentionality and collective agency (Cf. Gilbert 1989, 2000; Searle 1995, 2010; Tuomela 2007; and Bratman, 1992, 1993). Much of these discussions have been circling around whether there is such a thing as a genuine "we" in some thoughts and actions. Also, the theories tend to put much emphasis on deliberate conscious states of mind, such as "me" consciously thinking and acting together with someone else. Hudin (2008)

adds to this a proposed explanation of selfless "we" mode of social cognition that requires a combination of collective intentionality and social commitment resulting in an emotional bond with the group and presenting a basis for moral sense:

"practical reasons [that] function differently from other types of practical reasons because they do not require rational deliberation in order to motivate, therefore dispensing with any need for satisfaction of members in the motivational set, or any appeal to desire (passion) in any form. " p. 237.

Experimental work of Tomasello and collaborators (2005) supports Hudin's thesis showing that humans naturally possess inclination to act for a common goal, with unique forms of sociality that distinguish humans from other animals such as great apes. That helps to understand position of humans among living organisms with respect to complex forms of cognition and morality. According to Tomasello humans social behaviour is based on the capacity of understanding of each other's intentions, sharing attention, and the capacity to imitate each other (Tomasello 2009, 2014).

The gap between cognition based on molecular languages of unicellular organisms to the human cognition is huge, and possible indications how it could be bridged can be found in the approach proposed by Feldman (2006), in his book *From Molecule to Metaphor: A Neural Theory of Language.* There are still many missing links in his explanations, but they pave the way towards more fundamental understanding of evolutionary mechanisms of cognition.

6 CONCLUSIONS AND FUTURE WORK

Social behavior has its cognitive aspects that are known as distributed cognition. The idea of distributed cognition has been developed in a number of influential works such as Lucy Suchman's *Plans and Situated Action* (1987), Varela, Thompson, and Rosch's *The Embodied Mind* (1993), Edwin Hutchins' *Cognition in the Wild* (1995) as well as Andy Clark's *Being There: Putting Brain, Body, and World Together Again* (1997) and *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* (2008).

However, it should be noticed that mentioned research addresses human social ontology. Work of Searle, Miller, Tuomela, Hutchins, Tomasello, Hudin and others focus on human-level cognition that should be understood as a complex high-level type of cognition.

The model presented in current work starts in another end, with collective activities among cognising agents ranging from the simplest ones like bacteria, via semi-automatic information processing organisms like insects to the highest level cognising agents such as humans, trying to find as general principles as possible to cover all forms of cognition at the individual and at the collective level.

In order to understand the basic mechanisms of social cognition, it is instructive to analyse rudimentary forms of cognitive behaviours such as those in bacteria and insects. Based on the information-processing model of embodied cognition, our hope is to be able to contribute to the common view of cognition as natural, embodied distributed information processing.

Further progress will require building a broadly based, unified cognitive science, capable of multi-level computational modelling of cognitive phenomena, from molecules to (human) language, as emphasized by Feldman (2006). Damasio (2003) aptly notes, that there is a common basis for this unified approach:

"*All living organisms from the humble amoeba to the human are born with devices designed to solve automatically, no proper reasoning required, the basic problems of life. Those problems are: finding sources of energy; incorporating and transforming energy; maintaining a chemical balance of the interior compatible with the life process; maintaining the organism's structure by repairing its wear and tear; and fending off external agents of disease and physical injury." p. 30.*

The process of theory construction for bridging the gap between unicellular cognition and the distributed human cognition is just in the beginning, but we have better than ever models and computational (simulation) tools to explore this uncharted territory.

7 ACKNOWLEDGMENTS

The authors want to acknowledge the constructive and useful comments of the anonymous reviewers.

REFERENCES

Almér, A. & Allwood, J. (2013) "Social facts: collective intentionality and other types of social organization" Conference presentation at ENSO-III Helsinki 23-25.10.2013

Almér, A. (2007) Naturalising Intentionality: Inquiries into Realism & Relativism, Acta Universitatis Gothoburgensis.

Ben-Jacob, E. (2008) "Social behavior of bacteria: from physics to complex organization." The European Physical Journal B, 65(3), 315–322.

Ben-Jacob, E., Becker, I., & Shapira, Y. (2004) "Bacteria Linguistic Communication and Social Intelligence." Trends in Microbiology, 12(8), 366–372.

Ben-Jacob, E., Shapira, Y., & Tauber, A. I. (2006) "Seeking the Foundations of Cognition in Bacteria". Physica A, 359, 495–524.

Bratman, M. 1992. "Shared cooperate activity", The Philosophical Review, 101(2):327-341.

Bratman, M. (1993) "Shared Intention", Ethics, 104:97-113.

Cummins, R. ([1975] 1998), "Functional Analysis", in Colin Allen, Marc Bekoff, and George V. Lauder, Eds. Nature's Purposes: Analyses of Function and Design in Biology. Cambridge, MA: MIT Press, 169-196.

Damasio, A. (2003) Looking for Spinoza: Joy, Sorrow, and the Feeling Brain. William Heinemann. London

Davies, P. S. (2000) "Malfunctions", Biology and Philosophy, 15, pp. 19-38.

Dodig-Crnkovic, G. (2010) "Constructive Research and Info-Computational Knowledge Generation", In: Magnani, L.; Carnielli, W.; Pizzi, C. (Eds.) Model-Based Reasoning In Science And Technology Abduction, Logic, and Computational Discovery Series: Studies in Computational Intelligence, Vol. 314 X, Springer, Heidelberg Berlin, pp. 359-380.

Dodig-Crnkovic G. (2008) "Knowledge Generation as Natural Computation", Journal of Systemics, Cybernetics and Informatics, Vol 6, No 2.

Dodig-Crnkovic G. (2012) "Physical Computation as Dynamics of Form that Glues Everything Together", Information 3(2), pp. 204-218.

Dodig-Crnkovic, G. (2014) "Info-computational Constructivism and Cognition", Constructivist Foundations 9(2) pp. 223-231.

Dodig-Crnkovic, G. (2013) "Information, Computation, Cognition. Agency-based Hierarchies of Levels" PT-AI St Antony's College, Oxford, 20.09.2013 http://arxiv.org/abs/1311.0413

Dodig-Crnkovic, G (2014) "Modeling Life as Cognitive Info-Computation", In: Beckmann A., Csuhaj-Varjú E. and Meer K. (Eds.)

Proc. 10th Computability in Europe 2014, Budapest, Hungary, LNCS, Springer.
Doya K. (2002) "Metalearning and neuromodulation". Neural Networks, 15:495–506.
Doya K. (2000) "Metalearning, neuromodulation, and emotion". In The 13th Toyota Conference on Affective Minds, pp. 101–104.
Dukas, R. (2008). Evolutionary biology of insect learning. *Annual Review of Entomology*, *53*, 145–160. doi:10.1146/annurev.ento.53.103106.093343
Feldman J. A. (2006) From Molecule to Metaphor: A Neural Theory of Language, MIT Press. Bradford book, Cambridge, MA.
Gilbert M. (2000) Sociality and Responsibility: New Essays in Plural Subject Theory. Lanham, MD: Rowman and Littlefield.
Gilbert, M. (1989) On Social Facts, London: Routledge.
Hudin, J. (2008) "The Logic of External Reasons and Collective Intentionality" In: Hans Bernhard Schmid, Katinka Schulte-Ostermann, and Nikos Psarros (eds.), Concepts of Sharedness: Essays on Collective Intentionality, Ontos.
Hutchins, E. (1995) Cognition in the Wild. Cambridge, MA: MIT Press.
Koch C. and Tononi G. (2008) "Can Machines Be Conscious?" IEEE Spectrum, Vol. 45, No. 6, pp 54–59.
Marsh, L. and Onof, C. (2008) "Stigmergic epistemology, stigmergic cognition" Cognitive Systems Research, Vol. 9, No. 1-2, pp. 136-149.
Maturana, H., & Varela, F. (1980) Autopoiesis and cognition: the realization of the living. Dordrecht Holland: D. Reidel Pub. Co.
Millikan, R. (1984) Language, Thought, and Other Biological Categories. MIT Press, Cambridge.
Millikan, R. (2004) Varieties of Meaning, Cambridge, Mass: MIT Press.
Miłkowski, M., & Talmont-Kaminski, K. (2015) Explaining representation, naturally, New Ideas in Psychology http://dx.doi.org/10.1016/j.newideapsych.2015.01.002
Montague P. R., Dayan P. and Sejnowski J. T. (1996) "A framework for mesencephalic dopamine systems based on predictive Hebbian learning". Journal of Neuroscience, 16:1936–1947.
Nakagaki, T., Yamada, H., & Tóth, a. (2000). Maze-solving by an amoeboid organism. *Nature*, *407* (September), 470. doi:10.1038/35035159
Neander, K. (2006) "Content for Cognitive Science", In G. McDonald and D. Papineau (eds.), Teleosemantics, Oxford: Oxford University Press, 167–194.
Nelson D. L. and Cox M. M. Lehninger (2008) Principles of Biochemistry, V Edition: Chapter 12 Biosignalingp.419. Palgrave Macmillan.
Ng, W.-L., & Bassler, B. L. (2009) "Bacterial quorum-sensing network architectures." Annual Review of Genetics, 43, 197–222.
Scheutz, M. (2002) Computationalism New Directions. Cambridge Mass.: MIT Press.
Searle, J. (1995) The Construction of Social Reality, New York: The Free Press.
Searle, J. (2010) Making the Social World, Oxford: Oxford University Press.
Tomasello, M. (2014) "The ultra-social animal." Eur J Soc Psychol. 44(3): 187–194.
Tomasello, M. (2009) Why we cooperate. Cambridge, MA: The MIT Press.
Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). "Understanding and sharing intentions: The origins of cultural cognition." Behavioral and Brain Sciences, 28, 675 691.
Tononi G. (2008) "Consciousness as Integrated Information: A Provisional Manifesto." Biological Bulletin, Vol. 215, No. 3, pp. 216–242.
Tuomela, R. (2007) The Philosophy of Sociality - The Shared Point of View, Oxford University Press.
von Haugwitz, R., Kitamura, Y., & Takashima, K. (2012). Modulating reinforcement-learning parameters using agent emotions. *6th International Conference on Soft Computing and Intelligent Systems, and 13th International Symposium on Advanced Intelligence Systems, SCIS/ISIS 2012*, 1281–1285. doi:10.1109/SCIS-ISIS.2012.6505340

Social Computing Privacy and Online Relationships

Gaurav Misra[1] and **Jose M. Such**[2]

Abstract. Social computing has revolutionized interpersonal communication. It has introduced the aspect of social relationships which people can utilize to communicate with the vast spectrum of their contacts. However, the major Online Social Networks (OSNs) have been found to be falling short of appropriately accommodating these relationships in their privacy controls which leads to undesirable consequences for the users. This paper highlights some of the shortcomings of the OSNs with respect to their handling of social relationships and enumerates numerous challenges which need to be conquered in order to provide users with a truly social experience.

1 Introduction

The emergence of Online Social Networks (OSNs) in recent years has introduced a new paradigm of interpersonal communication. It has provided people with the ability of communicating with a large number of people instantaneously. The nature of communication largely depends on the particular function of the OSN. There are many general purpose OSNs such as Facebook, Google+ and Twitter which are used by millions of users everyday. These sites try to implement all facets of social communication and users are largely free to use the medium according to their convenience and preferences. There are other specialized OSNs which focus on one particular goal (for eg: LinkedIn is an OSN for professionals). The various functions that these sites perform ensure that most people have a presence on one or more of these sites. Facebook, the largest OSN in the World, has about 1.3 billion monthly active users (those users who use the site at least once a month)[3]. A large majority of them (75%) are from outside the United States which exhibits the global reach of Facebook. China has its own social networking giant called Qzone which has more than 600 million users[4]. These figures portray the global reach of these sites which results in a remarkably huge amount of information being exchanged on these networks.

The users of these OSNs share a lot of content on these platforms. They often share information which is personal and related to the activities in their everyday life. Most OSNs require the fulfillment of a "profile page" which contains personally identifiable information (PII) of the user. Details like age, current location, workplace, relationship status, etc., can be enumerated on these pages. However, many of the modern OSNs allow the user to abstain from enumerating these personal details or even regulate access to such information by employing the privacy controls afforded to them by the OSN infrastructure. In such a scenario, it becomes imperative for the users to understand and interpret the risks that information disclosure can have on their privacy. It is also important for them to fully understand the nature and workings of the privacy controls afforded to them in order to fully utilize the potential of these platforms.

Social media users interact with people representing various facets of their life such as work, family, education, etc. In such a scenario, it is essential for them to be able to distinguish between these different types of contacts and form various "virtual relationships" on the network. Moreover, it is important for the users to understand and acknowledge these different relationships and take them into account while disclosing information on the network [17, 40]. This is important in order to preserve the "contextual integrity" of the information which is being disclosed. If some information reaches unintended audiences and they process it without the appropriate context, this can be defined as a privacy breach according to Nissenbaum's theory of contextual integrity [29]. For example, embarrassing photographs of a person enjoying a night-out with his friends being revealed to his boss can lead to undesirable consequences for his professional career. He might think it is acceptable for him to disclose these images to his friends but may not find it desirable or appropriate to find them being disclosed to his boss. Such nuanced disclosure decisions are often required to maintain a favorable image of the user to all his contacts on the OSN. Social media users often use these platforms to project an "online persona" to their audience. This persona is created by the choice of information (such as posts, profile pictures, etc.) disclosed on the network. This careful management of one's presentation is an integral part of interpersonal communication in the offline world as well [14]. With the advent of social media, the opportunities of projecting one's identity to a large and dynamic audience have increased. However, as explained earlier, this also brings a few pitfalls with it if the user is not aware of who the audience really is. It is extremely important for social media users to form and maintain meaningful relationships on OSNs and leverage them while disclosing information in a way which preserves the contextual integrity of the information and also helps them to project a positive image to their audience.

In the subsequent sections of this paper, we focus on how user privacy on OSNs depend on relationships and the ability of the OSN infrastructures to enable and assist the users in accommodating these relationships in the information disclosure process. In section 2, we discuss the types of social relationships in OSNs and how they influence online behavior. Section 3 focuses on the handling of social relationships in OSNs and section 4 outlines some open challenges regarding how this can be improved.

[1] Lancaster University, United Kingdom, email: g.misra@lancaster.ac.uk
[2] Lancaster University, United Kingdom, email: j.such@lancaster.ac.uk
[3] http://www.statisticbrain.com/facebook-statistics/
[4] http://en.wikipedia.org/wiki/Qzone

2 Social Relationships on OSNs

Social media users typically have hundreds of connections on these platforms. In such a scenario, it is important for them to differentiate between different types of relationships to maintain meaningful and relevant communication with all of them. It has been found that different users treat social media communication differently [25, 28]. This diverse range of requirements mandate provisions of relationship management on OSNs. Users should be able to form and maintain relationships on these platforms and utilize them for information exchange. In this section, we look at the various types of relationships supported by OSNs of today. We also focus on how relationships can influence the users' privacy on the network.

2.1 Types of Relationships

There are different types of relationships users may share on OSNs. These typically depend on the nature and functionality of the particular OSN in question. Some OSNs allow the users to simulate offline relationships such as family, friends, co-workers, etc., while others may not offer such granularity. We categorize relationships into two main categories based on directionality:-

1. *Bidirectional* - These are relationships where both participants explicitly approve of and recognize the relationship. An illustrative example is the generic "friend" relationship in many modern OSNs. A user can send a "friend request" to another user who will get notified by the OSN infrastructure about this request. If that user accepts the request, a connection is made between the two users and their "friendship" is established on the network. Thus, both users (the initiator as well as the receiver) have to explicitly agree and accept that they want to be "friends" with each other. Popular OSNs such as Facebook and Google+ also allow the users to enumerate family members, colleagues, classmates, etc., in a similar way. These relationships typically mirror those found in real-life and help the users in acknowledging these relationships on the OSN as well.

2. *Unidirectional* - Some OSNs allow different types of relationships which can be formed unilaterally by a user. For example, the OSN Twitter allows users to become "followers" of other users and subscribe to all their unprotected "tweets". When a user wants to follow someone on Twitter, the followee often doesn't need to accept a request. The follower can start following the followee and can get access to the public content posted by them. Other examples of such relationships are "fans" on the OSN Hi5 and "subscribers" on Facebook (typically used for celebrity or brand pages).

It is evident that the nature of relationships supported by a particular OSN will depend heavily on the nature of its information flow. Moreover, the type of relationship (unidirectional or bidirectional) will determine the nature of access controls afforded to the users.

2.2 Social Relationships and Privacy

Having looked at the different types of relationships users can form on OSNs, we now take a look as to how these relationships can affect information disclosure decisions. Research findings in the past have suggested that the decision of whether or not to disclose a certain piece of information is often dependent on the "identity of the inquirer" [22]. In case of social media, the identity is further defined by the relationship the inquirer shares with the user. In other words, a decision of whether or not a user wants another user to access their information often depends on the relationship they share with them. There are various ways in which the different OSNs provide mechanisms for relationship management to the users. Popular general purpose OSNs like Facebook and Google+ provide the user with the opportunity of enumerating a rich set of relationships including friends, acquaintances, family, co-workers, etc. At the other end of the spectrum, some OSNs such as MySpace and Friendster only allow a binary distinction between "friends" (or contacts) and all other users of the network (often referred to as "public").

Social media users often utilize relationship information to make disclosure decisions. This information can either be explicit (the various relationship types mentioned earlier) or implicit (perceived by the user in the absence of such granularity). It has been observed that disclosure decisions should be made by keeping the balance between intimacy and privacy in mind [37]. The "intimacy-privacy" trade-off is negotiated differently by different users. Some users are more "pragmatic" when it comes to information disclosure as compared to others. Thus, they evaluate this trade-off less liberally than some other users. Nevertheless, irrespective of a particular user's attitude towards privacy, the intimacy-privacy trade-off has to be negotiated by all users. This suggests that the user should have a clear idea of the quality and strength of his relationship with other users in order to make informed decisions regarding information disclosure.

A user's social circle contains ties (or relationships) with a variety of strengths [15]. People utilize these differences in their connections for a number of objectives during interactions [39]. There have been many efforts to try and create a mechanism for determining the strength of social relationships on OSNs (commonly referred to as "tie-strength") in order to assist users in making information disclosure decisions. These approaches try to calculate a value for tie-strength using the information obtained from the amount and nature of interactions between users [13, 34]. Calculation of tie-strength can consider variables like the amount of messages exchanged between users, recency of communication, amount of shared content (such photos in which both the users are tagged), social distance and many others [13]. Some privacy management approaches have proposed using the tie-strength information to assist the user in making access control policies [10, 1, 38, 20]. The user gets access to the tie-strength information while making an information disclosure and can make a decision based on this. Tie-strength is also important as it is one of the factors considered by the algorithms employed by OSNs in order to present information to a user. For example, Facebook used the "EdgeRank" algorithm to prepare a user's newsfeed until recently. This algorithm used to consider "affinity" of one user with another which used many of the variables which are used for tie-strength calculation [4]. Facebook has modified their ranking algorithm in the recent past but it is not implausible to expect that they utilize some calculation to ascertain closeness of individuals on the network. Moreover, since many of the tie-strength calculations depend on the amount of interaction between users, the ranking algorithm also directly influences this value. If a user is not seeing another user's posts on their newsfeed, they do not have the opportunity to interact with it and hence the value for that particular variable is decreased leading to a negative change in their tie-strength.

Relationships on OSNs evolve, much like in real life. As users interact more with each other, their relationships start to change with respect to strength and/or type. It is also possible that people from one facet of someone's life, such as work, can be included and accommodated into another facet such as friends. Thus, it is plausible to imagine that user relationships are dynamic in nature. This dynamism in relationships also makes the task of safeguarding user's privacy a challenging task. It is possible that a change in relationship leads to loss of contextual integrity of some information disclosed by a user. For example, if a colleague from work joins an inner social circle of a user, he may get access to information which he previously didn't have. This may affect the colleague's perception of the individual and also impact their relationship. Such dynamism will also impact the intimacy-privacy trade-off. If a person's level of intimacy evolves with respect to a particular user, their privacy policy with respect to that particular individual should also be re-evaluated.

Recent research mentions that the strength of user relationships on Facebook change with time [5]. This means that users grow closer with other users who interact with them the most on these sites. User interactions can be in the form of visible cues such as comments, likes, etc. They can also be passive especially when receiving content in the form of an update or post made by another user. It is impossible for users to anticipate who has viewed the content posted by them unless any member of the audience interacts with it (with likes, comments, retweets, etc.) [2]. This is significant as it has been found that even such passive interaction results in an increase in strength of a relationship [5]. This means that if a friend simply views the news feed and activity about a friend shows up, the user is likely to feel closer to the friend as he now has some information (even if possibly trivial) about the friend's life. In the present scenario, the OSNs do not enable the users to identify such passive consumption of their content. The user should assume that every member of the audience of the content can and probably will (depending on the algorithm for information presentation to users for a particular OSN) be able to view the information.

This discussion shows the complexity of managing and maintaining social relationships on OSNs. The modern OSNs do allow the users to identify and enumerate individuals having different types of relationships with them. However, they fail to assist the user in maintaining and managing these relationships over time. The user is burdened with the task of interpreting the nature and evolution of their relationships with other users of the OSN and manage their interactions while keeping their privacy preferences in mind.

3 Social Shortcomings of Privacy Controls

It is evident that relationship management is both an important and challenging task for users of social media. Effective relationship management is necessary to maintain contextual integrity of user data and hence safeguard their privacy. In this section, we focus on the problems users face while trying to manage their relationships using existing privacy controls afforded to them by the OSNs.

The lack of granularity in privacy controls afforded to users of social media prevents them from selectively sharing their content to their audience. We have previously discussed the vast spectrum of relationships a user might have on an OSN. Ideally, the user should be able to selectively share content based on factors like relationship type and strength. However, it has been found that users struggle to achieve this objective using the privacy controls afforded to them by the OSN providers [17, 25]. Most OSNs fail to enable the user to differentiate between various relationship types while selecting an audience for their content. More recently, popular OSNs such as Facebook and Google+ have made an effort to assist users in contact management by creating Lists and Circles [19] respectively. These mechanisms help the user in partitioning their contacts and then use these partitions to selectively share their content with an appropriate audience according to their preference. However, it has been observed that users fail to employ these features during audience selection and end up sharing their content with unintended audiences [41]. Many users create these partitions when prompted by the OSN interface but fail to utilize them for selective sharing. Moreover, as discussed earlier, relationships evolve with time and these features do not offer any mechanism to the user to deal with this evolution. The responsibility of maintaining the appropriateness of these groupings lies solely on the user. This puts a cognitive burden on the user and hence most users end up not using these mechanisms for selective sharing. As a result, they end up "over-sharing" with unintended audiences [41, 18, 16]. It has also been shown that users often misinterpret privacy controls afforded to them. There can be a difference in what they expect from the privacy controls and what actually happens [24]. This cognitive gap is a significant one and it is important to attempt to try and bridge this gap as research has shown that users who are unaware of the full potential of the privacy controls afforded to them by OSNs are found to be more concerned about their privacy [36]. Thus, a failure to bridge this gap will result in a lot of cynicism among users about the privacy mechanisms being offered to them which can adversely affect the information flow on the network itself.

In the absence of suitable sharing mechanisms for users, they employ various "coping mechanisms" to try and safeguard their privacy [42]. Some of these coping mechanisms include "self-censorship" (not sharing something due ot the fear of a privacy breach) and "un-friending" contacts [32, 42]. Such mechanisms are often counter-productive for the user and diminish the utility of having a profile on these platforms. The users feel the need to resort to such coping mechanisms due to the effects of possible privacy breaches which can range from mild embarrassment to truly dire consequences [16].

The persistence of privacy problems on OSNs and the self-reported concerns of the users suggest that the OSNs fall short of delivering a truly social experience in which they can suitably share and disclose information according to their preferences. It is evident that the development of more usable and intelligent privacy controls are needed which will effectively reduce the cognitive burden on the user and enable them to selectively share their content within their social network depending on the various types of relationships they have with other users.

4 Mitigations and Open Challenges

In this paper, so far, we have highlighted the importance of relationship management on OSNs in order to safeguard the privacy of user data. We have also enumerated the aspects where the present OSN infrastructures fall short in supporting the user in this regard. In the remaining sections of this paper, we highlight some of the

mitigations which have been either adopted by the OSNs or have been suggested in literature but are yet to be adopted. We conclude the paper by outlining some unmitigated issues which can lead to further research in this domain.

4.1 Contact Management and Friend Grouping

Given the vast and varied nature of contacts any user interacts with on OSNs, it is important for them to be assisted with contact grouping. There is evidence to suggest that users conceptualize their social networks as constituting social groups and not a collection of individuals [21, 19]. We have already discussed the steps taken by OSNs such as Facebook and Google+ in providing their users with Lists and Circles in order to maintain their contacts. However, the responsibility for populating these partitions lies with the user. The user decides how to group their contacts and this can put a cognitive burden on them.

An alternative method of implementing contact grouping is by implementing community detection algorithms. Most traditional community detection algorithms leveraged network information and aimed to optimize modularity of the network [30]. However, communities formed using such techniques do not necessarily reflect the user's conception of their social network. Therefore, some recent techniques aim to mine "social circles" within a user's social network based on profile features (such as location, age range, education, etc.) of the contacts [27, 33]. Facebook has also introduced "smart lists" which automatically creates groups based on different life facets such as current location, school, workplace, etc., and populates them with the relevant contacts. However, their minimal use for audience selection suggests that their utility should be explained more clearly to the user to enable them to selectively share their content.

"ReGroup" suggests an alternative approach based on an interactive machine learning system which enables users to create on-demand contextual groups of their contacts [1]. Its machine learning component uses 18 features (such as gender, age range, hometown, recency of correspondence, friendship duration, etc.) to create profile vectors of all the friends of the user. The user can start the process of group creation by selecting some of the contacts for a particular group. The system suggests other contacts to be included in the group after learning the implicit context of the group creation and the similarity of the contacts with those that have already been selected by the user. These dynamically created groups can then be used by the user for audience selection to enable him to selectively share the content and preserve its contextual integrity.

4.2 Relationship-Based Access Controls (ReBAC)

The discussions in the preceding sections of this paper highlight the important role social relationships have in influencing information disclosure decisions made by users of social media. However, traditional access control models such as Role-Based Access Control (RBAC) fail to capture social relationships among the users [11]. In this section, we discuss some of the proposed Relationship-Based Access Control (ReBAC) models.

A major requirement of a suitable ReBAC model is that it should be able to support multiple types of relationships that users may have on the OSNs. Many approaches leverage tie-strength information to provide the users with usable access control mechanisms based on their social relationships [6, 7]. As we have discussed previously in this paper, tie-strength plays a key role in influencing disclosure decisions on OSNs. Thus, ReBAC models leveraging this information are likely to produce user-friendly mechanisms for access control and assist the users in information disclosure to appropriate audiences. Another important factor to be considered while designing ReBAC systems are the directional nature of relationships [12, 3]. The direction of the relationship determines the pattern of information flow in the network between the connections and hence it is important to consider this information while designing access control systems. It is also important to consider the users' relationship with the content that is being shared for a ReBAC system to be effective [6].

4.3 Improving Usability of Privacy Controls

Evidence from research suggests that there is a clear lack of understanding among users regarding the various privacy controls afforded to them by the OSNs [24]. This is also manifested in the lack of usage of contact grouping mechanisms for selective sharing [41]. Thus, there is a need for providing users with more usable privacy controls and also ensuring greater comprehension of the utility of these controls. There have been extensive efforts by researchers to try and suggest mechanisms to improve the visualization of privacy controls. Lipford, et al. suggested the use of an "audience view" which would enable the user to view their profile as it would appear to audiences having varying levels of access [23]. This mechanism has subsequently been adopted by Facebook which now allows its users to view their profiles as "friends" or "public". This ensures that the user is aware as to what information is accessible to what kind of an audience. Armed with this information, the user can then tweak the access control settings according to their preferences. An alternative visualization is the use of color-coding to signify the visibility controls of profile information [31]. The color code depends on whether the information is shared with no one (red), only selected friends (blue), all friends (yellow) and everyone (green).

The above mentioned approaches are useful in understanding the visibility controls with respect to a user's profile. However, the granularity of the different classes of audience (friends, network and public) is not precise enough. They do not account for the different social groups that the user may have created to organize the contacts. *PViz* is a privacy comprehension based on a graphical display which shows all the sub-groups which a user has in his friend network [26]. It can be seen as an extension of the "audience view" model which accommodates the option of viewing visibility controls for sub-groups of the user's contacts.

The different approaches mentioned here would help the user in comprehending the effects of their chosen access control policy. However, the usability of audience selection techniques also needs improvement to be geared towards assisting the user in selecting an appropriate audience for their content. A particular way of assisting the user to select the appropriate audience for their content is by providing them with information such as their "tie-strength" with different members of their social network [10, 20]. If the user is provided with this information while selecting an audience, they can consider the sensitivity of the content and evaluate the intimacy-privacy trade-off and select an appropriate audience. Other assisting information can be community membership of the contacts. This can be espe-

cially helpful if the communities are a true reflection of the user's conception of their social groups or if they represent different life facets. This information can be presented in the form of interventions during the information disclosure process. There is evidence to suggest that such interventions can lower the risk of unintended dissemination of information on the network [38]. However, it is important to acknowledge the fact that such interventions should not disturb the dynamic nature of information exchange on these platforms and should preserve the seamless user experience. Thus, any intervention or user assistance mechanism should be computationally light-weight.

4.4 Privacy Protection Models

This paper has highlighted many areas where current OSNs fall short in addressing the privacy concerns of the users. In this section of the paper, we look at some of the proposed approaches in literature which aim to mitigate these privacy problems.

There have been some proposed approaches which look to mine privacy policies from a user's peer network. This can potentially guide the user in setting the privacy controls based on what other users in their network have done. A similarity metric for identifying similar users of the network is required to provide meaningful linkages between relevant privacy policies. When a user sets a privacy policy for a particular piece of content, the algorithm checks for privacy policies listed by similar users for similar content and comes up with a predicted policy to suggest to the user [35]. Such models are required to leverage metadata of the content as well in order to understand similar content to provide relevant suggestions. Such an approach can significantly reduce the cognitive burden on the user by providing meaningful policy suggestions from which they can choose a desired policy. A similar approach is to leverage network connections and extract contexts for information disclosure from high density sub-graphs [8]. The underlying assumption here is that if a network connection exists between two users, they are likely to exchange information independent of the network as well. This assumption helps to identify shared contexts between users which can assist in framing access control policies which will preserve the contextual integrity of the information which is exchanged.

There have been a lot of efforts whcih are geared towards trying to provide OSN users with usable content dissemination systems. We have already discussed the relative rigidity and lack of granularity of some of the controls provided by OSNs to the users. Many approaches aim to address this problem by employing machine learning techniques in order to provide dynamic suggestions to users. Fang, et al. [9] propose a model for designing "Privacy Wizards" which use active learning techniques aimed at providing the user of a social network with a concise representation of their privacy choices (typically allow or deny type) for their personal data with respect to their friends in the social network. The user is required to assign access control labels to each contact with respect to the data item. The algorithm learns from the choices made by the user who can choose to abandon the labeling at any point. The algorithm aims to understand the implicit rules employed by the user in assigning access controls to different contacts. It then interprets these rules and comes up with suggested access controls for the unlabeled contacts of the user. This can potentially reduce a lot of effort as the task of exhaustively creating access control lists for each and every contact is a prohibitively complex task for most social media users. This approach can further be enhanced by leveraging features like community membership and tie-strength to provide more meaningful suggestions with minimum number of labeled contacts. "PriMa" is a semi-automated privacy protection mechanism which considers the intimacy-privacy trade-off for information disclosure decisions [34]. It considers a "risk factor" associated with the sensitivity of the content. It balances this risk factor with the "relationship score" which simulates tie-strength calculation. These two factors are weighed and a user-access score is created which suggests whether the user should allow or deny access to a particular user for a data item. The user has the ability to make the final decision and can fix the threshold of user-access score to automate the process.

As we have observed in this section, there have been some proposed privacy protection models which leverage some of the important aspects of social relationships (such as intimacy-privacy tradeoff) that have been discussed in this paper. Adoption of similar mechanisms in the OSN functionality will enhance the social aspect of audience selection and information disclosure.

5 Conclusions

Users of social media are required to form and maintain relationships with their contacts on these platforms to enable effective and manageable communication. These relationships are an important factor in helping the user to conceptualize and organize their vast social network. In this paper, we have discussed the important role these social relationships have with respect to privacy of user data. The various features of these relationships such as directionality and strength are considered to be important deciding factors by the users while making information disclosure decisions on OSNs. This suggests that privacy controls offered by OSNs should adequately accommodate and account for the various facets of these relationships in order to provide usable audience selection controls to its users.

We have observed, however, that most OSNs fall short of accommodating these social relationships in the access control mechanisms provided to their users. Due to this gap, users often encounter privacy breaches and have to face the unpleasant consequences which follow. Recently, major OSNs like Facebook and Goolge+ have made various attempts to rectify the situation by introducing contact management tools such as Lists and Circles but even these provisions have been found to fall short of solving users' privacy problems. We have highlighted some important challenges that need to be addressed for development of usable privacy controls and also enumerated some of important research efforts in this domain. Based on the analysis presented in this paper, we conclude that there is still a fair way for the OSNs to go before they can be deemed to be truly social and cater to the dynamic and multifarious needs of the OSN users.

REFERENCES

[1] Saleema Amershi, James Fogarty, and Daniel Weld, 'Regroup: Interactive machine learning for on-demand group creation in social networks', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 21–30. ACM, (2012).

[2] Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer, 'Quantifying the invisible audience in social networks', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 21–30. ACM, (2013).

[3] Glenn Bruns, Philip WL Fong, Ida Siahaan, and Michael Huth, 'Relationship-based access control: its expression and enforcement through hybrid logic', in *Proceedings of the second ACM conference on Data and Application Security and Privacy*, pp. 117–124. ACM, (2012).

[4] Taina Bucher, 'Want to be on the top? algorithmic power and the threat of invisibility on facebook', *new media & society*, **14**(7), 1164–1180, (2012).

[5] Moira Burke and Robert E Kraut, 'Growing closer on facebook: changes in tie strength through social network site use', in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 4187–4196. ACM, (2014).

[6] Barbara Carminati, Elena Ferrari, Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham, 'Semantic web-based social network access control', *computers & security*, **30**(2), 108–115, (2011).

[7] Barbara Carminati, Elena Ferrari, and Andrea Perego, 'Enforcing access control in web-based social networks', *ACM Transactions on Information and System Security (TISSEC)*, **13**(1), 6, (2009).

[8] George Danezis, 'Inferring privacy policies for social networking services', in *Proceedings of the 2nd ACM workshop on Security and artificial intelligence*, pp. 5–10. ACM, (2009).

[9] Lujun Fang and Kristen LeFevre, 'Privacy wizards for social networking sites', in *Proceedings of the 19th international conference on World wide web*, pp. 351–360. ACM, (2010).

[10] Ricard L Fogués, Jose M Such, Agustin Espinosa, and Ana Garcia-Fornes, 'Bff: A tool for eliciting tie strength and user communities in social networking services', *Information Systems Frontiers*, 1–13, (2013).

[11] Ricard L Fogués, Jose M Such, Agustin Espinosa, and Ana Garcia-Fornes, 'Open challenges in relationship-based privacy mechanisms for social network services', *International Journal of Human-Computer Interaction, In press*, (2014).

[12] Philip WL Fong, 'Relationship-based access control: protection model and policy language', in *Proceedings of the first ACM conference on Data and application security and privacy*, pp. 191–202. ACM, (2011).

[13] Eric Gilbert and Karrie Karahalios, 'Predicting tie strength with social media', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 211–220. ACM, (2009).

[14] Erving Goffman, 'The presentation of self in everyday life', (1959).

[15] Mark S Granovetter, 'The strength of weak ties', *American journal of sociology*, 1360–1380, (1973).

[16] David J Houghton and Adam N Joinson, 'Privacy, social network sites, and social relations', *Journal of Technology in Human Services*, **28**(1-2), 74–94, (2010).

[17] Gordon Hull, Heather Richter Lipford, and Celine Latulipe, 'Contextual gaps: Privacy issues on facebook', *Ethics and information technology*, **13**(4), 289–302, (2011).

[18] Maritza Johnson, Serge Egelman, and Steven M Bellovin, 'Facebook and privacy: it's complicated', in *Proceedings of the Eighth Symposium on Usable Privacy and Security*, p. 9. ACM, (2012).

[19] Sanjay Kairam, Mike Brzozowski, David Huffaker, and Ed Chi, 'Talking in circles: selective sharing in google+', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1065–1074. ACM, (2012).

[20] Michaela Kauer, Benjamin Franz, Thomas Pfeiffer, Martin Heine, and Delphine Christin, 'Improving privacy settings for facebook by using interpersonal distance as criterion', in *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pp. 793–798. ACM, (2013).

[21] Patrick Gage Kelley, Robin Brewer, Yael Mayer, Lorrie Faith Cranor, and Norman Sadeh, 'An investigation into facebook friend grouping', in *Human-Computer Interaction–INTERACT 2011*, 216–233, Springer, (2011).

[22] Scott Lederer, Jennifer Mankoff, and Anind K Dey, 'Who wants to know what when? privacy preference determinants in ubiquitous computing', in *CHI'03 extended abstracts on Human factors in computing systems*, pp. 724–725. ACM, (2003).

[23] H.R. Lipford, A. Besmer, and J. Watson, 'Understanding privacy settings in facebook with an audience view', in *Proceedings of the 1st Conference on Usability, Psychology, and Security*, pp. 1–8. USENIX Association Berkeley, CA, USA, (2008).

[24] Yabing Liu, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove, 'Analyzing facebook privacy settings: user expectations vs. reality', in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 61–70. ACM, (2011).

[25] Alice E Marwick et al., 'I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience', *New Media & Society*, **13**(1), 114–133, (2011).

[26] Alessandra Mazzia, Kristen LeFevre, and Eytan Adar, 'The pviz comprehension tool for social network privacy settings', in *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pp. 13:1–13:12, New York, NY, USA, (2012). ACM.

[27] Julian McAuley and Jure Leskovec, 'Discovering social circles in ego networks', *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **8**(1), 4, (2014).

[28] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai, 'Is it really about me?: message content in social awareness streams', in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 189–192. ACM, (2010).

[29] Helen Nissenbaum, 'Privacy as contextual integrity', *Washington Law Review*, **79**, 119, (2004).

[30] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos, 'Community detection in social media', *Data Mining and Knowledge Discovery*, **24**(3), 515–554, (2012).

[31] Thomas Paul, Daniel Puscher, and Thorsten Strufe, 'Improving the usability of privacy settings in facebook', *arXiv preprint arXiv:1109.6046*, (2011).

[32] Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber Lynn McConahy, Jason Wiese, and Lorrie Faith Cranor, 'The post that wasn't: exploring self-censorship on facebook', in *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 793–802. ACM, (2013).

[33] Anna Squicciarini, Sushama Karumanchi, Dan Lin, and Nicole DeSisto, 'Identifying hidden social circles for advanced privacy configuration', *Computers & Security*, (2013).

[34] Anna Squicciarini, Federica Paci, and Smitha Sundareswaran, 'Prima: an effective privacy protection mechanism for social networks', in *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*, pp. 320–323. ACM, (2010).

[35] Anna Cinzia Squicciarini, Smitha Sundareswaran, Dan Lin, and Josh Wede, 'A3p: adaptive policy prediction for shared images over popular content sharing sites', in *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pp. 261–270. ACM, (2011).

[36] Jessica Staddon, David Huffaker, Larkin Brown, and Aaron Sedley, 'Are privacy concerns a turn-off?: engagement and privacy in social networks', in *Proceedings of the Eighth Symposium on Usable Privacy and Security*, p. 10. ACM, (2012).

[37] Jose M Such, AgustíN Espinosa, Ana GarcíA-Fornes, and Carles Sierra, 'Self-disclosure decision making based on intimacy and privacy', *Information Sciences*, **211**, 93–111, (2012).

[38] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh, 'A field trial of privacy nudges for facebook', in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 2367–2376. ACM, (2014).

[39] Barry Wellman and Scot Wortley, 'Different strokes from different folks: Community ties and social support', *American journal of Sociology*, 558–588, (1990).

[40] Jason Wiese, Patrick Gage Kelley, Lorrie Faith Cranor, Laura Dabbish, Jason I Hong, and John Zimmerman, 'Are you close with me? are you nearby?: investigating social groups, closeness, and willingness to share.', in *UbiComp*, pp. 197–206, (2011).

[41] Pamela Wisniewski, Bart P Knijnenburg, and H Richter Lipford, 'Profiling facebook users privacy behaviors', in *SOUPS2014 Workshop on Privacy Personas and Segmentation*, (2014).

[42] Pamela Wisniewski, Heather Lipford, and David Wilson, 'Fighting for my space: Coping mechanisms for sns boundary regulation', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 609–618. ACM, (2012).

Computational Aspects of Autonomous Discursive Practices

Raffaela Giovagnoli[1]

Abstract. A "pragmatic conception" of computation can help to isolate (1) what capacities and abilities are common to human and non-human animals, and machines and (2) what capacities and abilities are typical of human beings. I'll show the motivation for a pragmatic philosophical approach and, in particular, the original application of "Analytic Pragmatism" to AI. The results of this analysis is a form of weak AI, which admits some important differences between animal and non-animal reasoning[1].

1. INTRODUCTION

To choose a pragmatic strategy is to presuppose that we understand pragmatism in a distinctive way. So, it is useful to distinguish between a "narrow" interpretation and a "wide" one [1]. Why should we adopt this distinction?

Classical pragmatism of Charles Peirce, Williams James and John Dewey is a form of narrow pragmatism that rests on Peirce's famous maxim in "How to make our Ideas Clear": Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of those effects is the whole of our conception of the object. It has a verificationist character "our idea of anything is our idea of its sensible effects". So we mean by wine something that has certain distinctive effects upon the senses. This idea introduces the difference between reality and truth. The first is what has some effects on our senses, whether the second depends on the agreement in the scientific community; the final opinion is the truth and the object represented in it is the real. James has a different conception of truth, which rests on the idea that beliefs are made true by the fact that they enable us to make accurate predictions of the future run of experience. James seems to show other similar interpretations of the "goodness of belief". For instance, the truth of a theological proposition is due to the fact that it has "a value for concrete life". The idea of God possesses a majesty, which can "yield religious comfort to a most respectable class of minds". A theoretical important consequence is that pragmatism is the role of practice to contribute to the constitution of objects. Dewey conception is more radical about the problem of "fixing" a situation, which is indeterminate at the beginning of the research. He uses "logical forms" as ideal instruments that help us to transform things and to resolve our problem. So, we can underscore a peculiar conception of experience that overcomes classical empiricism, namely the fact that experience is "full of inferences". This is because what we experience is shaped by our habits and expectation. So are shaped also our representations of reality, namely the content of our thoughts. The content of a belief is determined by its role in our action, namely what we should do in the light of our desires and our background knowledge. According to James and Dewey all our concepts and theories are instruments to be judged by how they achieve theory's intended purpose. Peirce develops the famous theory of signs, which rests on the triadic sign-relation: a sign or thought is about some object because it is understood, in subsequent thought, as a sign of that object. Because of the role of the subsequent thought as interpretant we can observe that the content of a thought is determined by the ways in which we can use it in inference and the planning of action.

Tradition apart, we can consider important pragmatic issues from C. I. Lewis, Murray Murphy and G. Herbert Mead. I would like to embrace Bob Brandom's suggestion for including some perspectives in the "wide" interpretation of pragmatism [2]. The reason for enlarging the notion is the search for a role of practices, which is not restricted to an instrumental nature. If we think to the use of language we think that it constitutes the content or meaning of linguistic expressions. We can distinguish between:

1. Methodological pragmatism: the content of linguistic expression must be explained in terms of some distinctive characteristic of their use (Dummett, Tarsky, Quine);
2. Semantic Pragmatism: the speakers constitute the meaning or content by using expression in a manner that determines the association between expression and content;
3. Fundamental Pragmatism: the capacity to know-that or believe-that is parasitic of a more primitive know-how, namely the capacity to adapt to environment (early Heidegger, Dreyfus and Haugeland);
4. Linguistic Pragmatism: to take part to linguistic practices is a necessary condition to have thoughts and beliefs in a strict sense (Sellars, Davidson and Dummett).

[1] Faculty of Philosophy, Pontifical Lateran University, Rome. Email: raffa.giovagnoli@tiscali.it; giovagnoli@pul.it.
[1] I wish to thank the referees for very fruitful comments to this early version of my paper.

This distinctions help to introduce Brandom's analytic pragmatism that focuses on the normative regulation of our practices; in particular, practices involved in reasoning and cognitive activities. He follows Sellars according to which rationality means the ability to recognize the force of reasons and this very capacity is a kind of activity that allows us to take responsibility for how well we reason and act.

2. A SOCIAL MODEL FOR THE GAME OF "GIVING AND ASKING FOR REASONS"

Brandom's enterprise in his most relevant book *Making It Explicit* is devoted to develop a new social model for describing the Sellarsian "game of giving and asking for reasons" [3]. Beyond the classical conception of representation, the notion of content or meaning of linguistic expressions is intended in inferential and social terms. Social practices are discursive practices (inferentially articulated), which confer content to expressions and actions according to a precise normative vocabulary. The idea of learning the inferential use of a concept is bound to social attitudes that imply "responsibility"" and " authority". The game of giving and asking for reasons becomes, therefore, dependent on the social practices by which we recognize commitments and entitlements. The "scorekeeper" takes the place of the Sellarsian knower and becomes a "social role". The scorekeeper is the one who is able to reliably recognize inferentially articulated commitments that constitute the content of beliefs. He possesses an "expressive" rationality as the capacity to perform inferences in the game of giving and asking for reasons.

According to Hegel, the very nature of negation is incompatibility, which is not only formal but also material, i.e., entails material properties as, for example, "triangular". In this sense, we can say that *non-p* is the consequence of anything materially incompatible with *p*. From an idealistic point of view we cannot objectively acknowledge relations of material incompatibility unless we take part in processes and practices by which we subjectively acknowledge the incompatibility among commitments. This is the reason why to apply a concept is to occupy a social position, i.e., to undertake a commitment (to take responsibility of justifying it or to be entitled to it). Thus, judgments, as the minimum unit of experience, possess two sides: the subjective side which indicates who is responsible for the validity of his claims, and the objective one, which indicates whatever the speaker considers as responsible for the validity of his/her claims. Through specific attitudes we can specify the social dimension of knowledge. The *de dicto* ascription such as "he believes that…", determines the content of a commitment from a subjective point of view, i.e., from the point of view of the one who performs a certain claim. The *de re* ascription such as "he believes **of** this thing that…", determines the content of a commitment from an objective point of view, i.e., the inferential commitments the scorekeeper must acknowledge [4]. How does this acknowledgment happen? We can use the above mentioned ascriptions. If, for example, I am a scorekeeper who performs the *de dicto* ascription «Vincenzo says that this golden agaric must be cooked in butter» and contemporarily I acknowledge that the mushroom is totally similar to an *amanita caesarea* (a good golden agaric) yet it is dangerous because it is an *amanita muscaria* (an evil golden agaric), I can isolate the content of Vincenzo's assertion through the *de re* ascription «Vincenzo says **of** this golden agaric that it must be cooked in butter» and make explicit the commitments I undertake and the ones I refuse from an objective point of view [5].

2. AUTONOMOUS DISCURSIVE PRACTICES AND AI

Making It Explicit aims at describing the social structure of the game of giving and asking for reasons, which is typical of human beings. *Between Saying and Doing* has a different task: it pursue the pragmatic end to describe the functioning of autonomous discursive practices (ADPs) and the use of vocabularies [6]. ADPs start from basic practices that give rise to different vocabularies and the analysis is extended to nonhuman intelligence.

The so-called "analytic pragmatism" (AP) represents a view that clarifies what abilities can be computationally implemented and what are typical of human reasoning. First, Brandom criticizes the interpretation of the Turing's Test given by strong artificial intelligence or GOFAI, but he accepts the challenge to show what abilities can be artificially elaborated to give rise to an autonomous discursive practice (ADP). What is interesting to me is that AI-functionalism or "pragmatic AI" simply maintains that there exist primitive abilities that can be algorithmically elaborated and that are not themselves already "discursive" abilities. There are basic abilities that can be elaborated into the ability to engage and ADP. But these abilities need not to be discovered only if something engage in any ADP, namely there are sufficient to engage in any ADP but not necessary. Brandom's view could be seen as a philosophical contribution to the discussion about how to revisit some classical questions: the role of symbols in thought, the question of whether thinking just is a manipulation of symbols and the problem of isomorphism as sufficient to establish genuine semantic contentfulness. It becomes interesting to continue the Wittgensteinean trend in the theory of action, which brings light on the differences between proper action and bodily movement, which are mechanical as in the case of machines, and the problem of rule following that is related to the question of the peculiarity of non-human and human learning. I just would like to remember Habermas early essay *Handlungen, Operationen, körperlichen Bewegungen* [7], in which several fruitful distinctions are introduced. To summarize:

- humans have a kind of consciousness of the rule-following as in suitable circumstances they can make explicit the propositional content of the rule they are following,
- non-humans have a kind of derived consciousness according to which we make sense of their rule

following and we give an interpretation of their behaviour,
- we speak of mere behaviour in case of absence of implicit consciousness of rule following so that there is only a minimal capacity of action.

Very interesting ideas come from the book *The Shape of Actions: What Humans and Machines Can Do*, in which Harry Collins and Martin Kusch propose a thoughtful theory of action that sets the boundaries between humans and machines [8]. Humans can do three things: polymorphic actions (actions that draw on an understanding derived from a sociological structure); mimeomorphic actions (actions that are performed like machines and do not require an understanding derived from a sociological structure) and they cam merely behave.
The strategy of AP is based on a "substantive" decomposition that is represented in algorithms. Any practice-or-ability P can be decomposed (pragmatically analyzed) into a set of primitive practices-or-abilities such that:

1. they are PP-sufficient for P, in the sense that P can be algorithmically elaborated from them (that is, that a*ll* you need in principle to be able to engage in or exercise P is to be able to engage in those abilities plus the algorithmic elaborative abilities, when these are all integrated as specified by some algorithm); and
2. one could have the capacity to engage or exercise *each* of those primitive practices-or-abilities without having the capacity to engage in or exercise the target practice-or-ability P.

For instance, the capacity to do long division is "substantively" algorithmically decomposable into the primitive capacities to do multiplication and subtraction. Namely, we can learn how to do multiplication and subtraction without yet having learning division. On the contrary, the capacities to differentially respond to colours or to wiggle the index finger "probably" are not algorithmically decomposable into more basic capacities because these are not things we do *by* doing something else. Starting from Sellars, we can call them *reliable differential capacities to respond to environmental stimuli* but these capacities are common to humans, parrots and thermostats [9]. Along the line introduced by Sellars, Brandom intends ADP typical of human practices in an "inferential" sense and strictly correlated with capacities to deploy an autonomous vocabulary (namely a vocabulary typical of human social practices). They are grounded on the notion of "counterfactual robustness" that is bound to the so-called "frame problem". It is a cognitive skill namely the capacity to "ignore" factors that are not relevant for fruitful inferences. The problem for AI is not *how* to ignore but *what* to ignore. This is a way to overcome the analogical notion of intentionality that connotes Sellars' thought, by introducing a "relational" one. Basic practices that provide the very possibility to talk involve the capacity of attending to complex relational properties lying within the range of counterfactual robustness of various inferences.

CONCLUSION

I sketched the classical ideas from Pragmatism and introduced new conceptions, which enlarge the classical notion to overcome an instrumental sense of the philosophical research. Analytic Pragmatism has the advantage to introduce the logical structure of discursive practices that are typical of human beings while retaining a fruitful relation with basic practices characterizing machine learning. I would point on Brandom's thesis that only creatures that can talk can do that, because they have access to the combinatorial productive resources of a *language,* which allows humans to attend to many complex relational properties. But, I do not intend this thesis as a way of stating a primacy for human practices, rather the weaker descriptive end to analyze different practices we can observe in natural, artificial and social reality.

REFERENCES

[1] C. Hookway. *Pragmatism. The Stanford Enciclopedia of Philosophy*. http://plato.stanford.edu/entries/pragmatism, (2013).
[2] R. Brandom. Pragmatics and Pragmatism. In: *Hilary Putnam:Pragmatism and Realism*. J. E. Conant, U. M. Zeglen (Eds.). Routledge, London. UK. (2002).
[3] R. Brandom. Making It Explicit. Harvard University Press. Cambridge. USA. (1994).
[4] R. Brandom. Making It Explicit. Cambridge University Press. Cambridge. USA. Chap. 8. (1994).
[5] R. Giovagnoli. Razionalità espressiva. Scorekeeping: inferenzialismo, pratiche sociali e autonomia. Mimesis. Milano. Italy. (2004); R. Giovagnoli. On Normative Pragmatics. A Comparison between Brandom and Habermas. *Teorema*, XXIII; 51-68, (2003).
[6] R. Brandom. Between saying and Doing. Oxford University Press. Oxford. UK. (2008); R. Giovagnoli, Representation, Analytic pragmatism and AI. In: *Computing Nature*. G. Dodig-Crnkovic, R. Giovagnoli (Eds.), Springer, Germany, 2013; R. Kibble, Discourse as Practice: from Bordieau to Brandom. In: *Proceedings of the 50th Anniversary Convention of the AISB*, Goldsmith, UK, 2014.
[7] J. Habermas. Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns. Suhrkamp. Frankfurt am Main. Germany. (1984).
[8] H. Collins, M. Kusch. The Shape of Actions: What Humans and Machines Can Do. The MIT Press. USA. (1999).
[9] W. Sellars. Empiricism and The Philosophy of Mind. Harvard University Press. Cambridge. USA. (1997); D. MacBeth, Inference, Meaning and Truth in Brandom, Sellars and Frege, www.haverford-academia.edu

Digital identity: finding me

Yasemin J. Erden [1]

Abstract. Identity is neither simple nor static, and in many ways the multiplicity of identity that this paper will consider is not in itself either novel or controversial. Who I am as a writer, academic, sister, teacher, learner is as complex as who you are as a reader and everything else that you may be. Our everyday roles and experiences contribute to the complex nature of our identity, and we are both defined by (and define ourselves according to) the actions, choices, beliefs and emotions that we either choose or deny. In these respects it seems likely that what we might call a *digital identity* would merely add to the multiplicity of our existing complex picture of ourselves. What this paper will consider is whether this is indeed just another facet of what it is to be me, you, or anybody else, or whether our digital identity affects identity in differently, and (either way) in which direction of travel that relation follows. Am I me because of Facebook, or is my Facebook me?[2] Or are these relations reciprocal, or something else entirely?

1 THE DIGITAL GENERATION?

The concept of a digital identity (or footprint, tattoo, etc.) picks out the idea that a terrestrial human identity can stretch into the digital web. The term can point to a life lived online (through games or avatars), or one that is portrayed after the fact (such as on social networks, message-boards, or blogs). It can reference a digital network of friends, as well as work associates and colleagues. A digital identity can in principle be singular. Whether this is through one output only, which is increasingly rare, or through the persistence of one single identity through all digital output, which is still possible. The connection can be drawn by an individual alone, and can include a single representation of a perceived identity by the person, or can be identified or created by an observer who can access and associate photos or personal information to a single user. Indeed, certain data mining software can already achieve this with relative ease.[3] It may also consist of multiple yet discrete individual strands of identity manipulated by a single user who yet (purposefully or otherwise) does not draw attention to, or does not perceive there to be, links between them. As Palfrey and Gasser [1] show, there is generally a lack of agreement about whether there are one or multiple identities amongst the generation of digital users born after the so-called *digital-explosion.*

On the one hand a digital representation of identity can seem fleeting, or open to change, for instance where information is easily amended, deleted, constructed, reconstructed. On the other hand, information persists. An online identity can remain tethered to inaccessible and/or persistent threads of information that remains on the web long past a person's own mortal existence in the world. Yet the concept of permanence on the web—the limitless persistence of uploaded information—is in fact one that is uncertain. For instance, Case C-131/12 was heard at the Court of Justice in May 2014, on the topic of Personal data and the "Protection of individuals with regard to the processing of such data". In this case the court ruled that a data subject "may oppose the indexing by a search engine of personal data relating to him where their dissemination through the search engine is prejudicial to him and his fundamental rights to the protection of those data and to privacy — which encompass the 'right to be forgotten' — override the legitimate interests of the operator of the search engine and the general interest in freedom of information." [2] The *right to be forgotten* as it's come to be known has yet to be fully tested, and it seems unlikely to be the end of the matter. Yet it is clear that some data, where such data is considered valuable in one form or another, is either carefully or haphazardly, and not always anonymously, catalogued and stored. This is not always with either the explicit or informed consent of the user, and where consent is sought, for instance in the ticking of agreements for services, users may not always be considered *informed*. But who is this user, and whose identity is at stake? It is this that this paper will explore.

Before this there are some distinctions to bear in mind and some to dispel. Palfrey and Gasser [1, p. 4] draw a distinction for instance between those to whom digital media is second nature, and those for whom it is learned behaviour: digital native for the former, digital immigrant for the latter. But who should we say occupies the former category, who the latter? Is it simply a matter of *age*? In fact, this terminological shorthand rather polarises between two groups, when many people may flit between one group and another (native to certain technologies, immigrant to some, alien to others). Buckingham [3] offers an alternative reading of the term "digital generation",[4] and this account may prove more fruitful. He cites a need to account for the fact that the impact of these technologies is not restricted to just the emerging identities of the young, but to the developing identities of all ages. He further notes [3, p. 2] that "generations are defined both historically and culturally", such that while the time frame may be important, it is not restricted to those who are *born* within that particular time frame. Indeed, and at the other end of the scale, there is little reason to suppose the generational distinction to be the most important distinction. This may be for a number of reasons. First because older generations within particular cultures may have more economic advantages, thus enabling better access to the digital world than many young people. This may be true across cultures. It is also the case that during the latter half of last century, and even in this century, the

[1] Philosophy, St Mary's University, London
Email: yj.erden@stmarys.ac.uk

[2] Other Social Networks are available.

[3] Cf. http://business.time.com/2012/07/31/big-data-knows-what-youre-doing-right-now/ [accessed 20/03/15]

[4] Buckingham [3, p. 11] also suggests caution with respect to the term digital generation since he claims it "runs the risk of attributing an all-powerful role to technology". This seems a reasonable comment, especially as it is fair to say that the *technology* in and of itself does not contain within it the potential for power. Rather it is how it is used, manipulated, and used as manipulation that should be of concern (by others, corporations, etc.), where risks and benefits are not equally considered.

majority of young people across the world still have little or limited access to such technologies.[5]

2 THE PHILOSOPHICAL *I*

The distinction between *society* and the *individual*, including where, what, and even the possibility of such distinction, has been hotly debated. The answer you give about where that distinction might lie will give an indication of your cultural upbringing, political affiliations and/or beliefs. Perhaps all three. Those philosophies which hold identity to be an *individual* matter, whereby a person is born with an essence, or develops this on their own account no longer hold much sway. Rorty [4, p. xiii] provides an easy account of why this might be the case, noting that those who deny "there is such a thing as 'human nature' or the 'deepest level of the self'" have, as their strategy "to insist that socialisation, and thus historical circumstance, goes all the way down…" This is the approach I will adopt in this paper, and in what follows I will present what I believe are convincing arguments regarding the necessarily *social* nature of identity formation. Along the way it should become clear that individualistic views, on this account, are untenable.

To do this, we can begin by examining the work by Taylor [5] who argues that a general feature of human life is "its fundamentally dialogical character." To which he adds that "The genesis of the human mind is…not 'monological,' not something each accomplishes on his or her own, but dialogical." For these reasons he suggests that our *identity* is thus defined "in dialogue with, sometimes in struggle against, the identities our significant others want to recognise in us" [5, p. 33]. This sense of struggle is encapsulated by this need for *recognition*. Taylor states that our identity "needs and is vulnerable to the recognition given or withheld by significant others" [5, p. 49]. Here we need to understand recognition of a person and/or their identity as pointing to more than just the action of *seeing*. A *willing* to recognise someone *as* is also important. As explained elsewhere, recognition and acceptance are key elements in both personhood and identity [6].

Along the same line, Markell [7, p. 41] concludes that the politics of recognition "actively constitutes the identities of those to whom it is addressed." The influence of Hegel's discussion of recognition is particularly relevant here:

> we are the sorts of beings we are with our characteristic "self-consciousness" only on account of the fact that we exist "for" each other or, more specifically, are *recognized* or *acknowledged* (*anerkannt*) *by* each other, an idea we might refer to as the "acknowledgment condition" for self-consciousness [8, p. 1]

Gilbert and Lennon [9, p. 140] discuss the "embodied nature of subjectivity," on which they describe "The constitution of subjectivity by other subjects," whether these are *general* or *particular* others. To this they add that the "Experiences of *sameness* with others serve to constitute the self." This includes where the construction of the *I* involves the self as engaged in the process of *differentiating itself*. Even here, the self requires and involves others (in simple terms the possibility of comparison requires that something must stand in comparison *to*).

If these philosophical accounts—supported by accounts offered in both social theory and psychology—of identity formation are taken seriously, we see that it is not only the other who forms our notion of self but the interaction through which this dialogical formation occurs. Following this line, we can see that questions need to be asked about the manner of interaction. If on this account our identity forms in relation to the other (including the myriad of social, cultural, political, and religious contexts), what then is the effect when that primary interaction or engagement with the other is *virtual*?

3 THE DIGITAL *WE*

With the expansion of online communication and more recently social networking, there has been the potential for closer and more immediate cross-linguistic and cross-cultural interactions. Given the infancies of these technologies and societal participation in them, the implications for broader notions of society and culture, as well as for notions of individual identity and personhood remain somewhat uncertain. On this, Palfrey and Gasser [1, p. 32] offer the claim that "what it means to be a young person hasn't changed; what has changed is the manner in which young people choose to express themselves." In one sense this may be true. In and of itself what it means to be *young* (as in *to not be old*) may not have changed, but it seems that now more than ever newer generations can engage with the world around them in new and distinct ways. Added to which the boundary for young-ness itself has shifted (it is less common to presume that adulthood necessarily and always begins at 18).

Multimedia interaction—gaming, social networks, online message-boards, instant messaging, blogging—impacts on the way we engage with others and the ways in which we make our voices heard, hear the voices of others, and how much time we give to each. By this stage however we only have speculative ideas about the sort of impact these subtle or major shifts in interaction may have on identity, or on our brains. What the effects of a continuous and complex multi-tasking may have on brain processing, for example, remains to be seen, and while there are claims that that such activity has already affected the manner in which our brains process information, and the relation between short and long term memory storage, these are certainly not conclusive (cf. [11] for further discussion on this topic, including conflicting accounts, research and evidence). Yet beliefs about the impact of such changes already impacts on the provision of education, such that the expectation in UK Higher Education is that teaching should and often must include digital platforms and content. Modern learning, educational methods, and even students are seen as somehow different to their predecessors, and students are as likely to be described in terms of their online, interactive, and collaborative learning identities (*digital clients*, is one such example) as by their analogue experience. Arguments are offered about whether and how such changes affect students, and much is assumed, but here as with much that is digital, there is little consensus, and even less certainty.

Prensky's seminal paper from 2001 'Digital Natives, Digital Immigrants' argues that students born into the digital world "think and process information fundamentally differently from their predecessors" [12]. This claim and the arguments that follow lead him to conclude that those who teach such students "speak an outdated language (that of the pre-digital age), are

[5] The reasons for this are both vast and important, but there is not the space to consider them here. Nevertheless it should be noted that where the consideration of a digitial identity is considered, access to such digitial media is necessarily assumed. This is neither a politically nor ethically neutral position, and the use of the "we" throughout this paper should be considered alongside the recognition that I offer here.

struggling to teach a population that speaks an entirely new language." A call to changes in education followed these and similar claims, but the evidence for this is largely anecdotal and (as I note above) is certainly not definitive. As Bennett, Maton and Kervin note, calls for major change in education, though "widely propounded", have in fact "been subjected to little critical scrutiny, are under-theorised and lack a sound empirical basis" [13]. In their exploration of the field, they instead found that while "a proportion of young people are highly adept with technology and rely on it for a range of information gathering and communication activities", this cannot be taken for granted since there is also "a significant proportion of young people who do not have the levels of access or technology skills predicted by proponents of the digital native idea." In conclusion they offer the following sober conclusions:

> While technology is embedded in their lives, young people's use and skills are not uniform. There is no evidence of widespread and universal disaffection, or of a distinctly different learning style the like of which has never been seen before. We may live in a highly technologised world, but it is conceivable that it has become so through evolution, rather than revolution. Young people may do things differently, but there are no grounds to consider them alien to us. Education may be under challenge to change, but it is not clear that it is being rejected.

Changes in general communication are perhaps less controversial and are more immediately apparent. It's indubitable, for instance, that there are differences in the ways that we communicate now as a result of technology, as well as the expectations that these changes bring. We send emails rather than letters, text messages rather than make phone calls, but how it is changing *us* is likely to prove a more difficult analysis. A subtle shift from thinking in one way to thinking in another is not always easy to track (we're not even sure about the way in which we currently think). Nevertheless, it is possible that our thinking *is* changing, and it is equally likely that the digital age has a hand in this. As noted above and in [11] current research into the way digital interaction may be changing our very brain processing, such that on foundational levels our very nature (as persons) is altered is still in its infancy.

In terms of expectation, the assumption that there could or should be immediate responses to messages (email, SMS) is striking, as well as the idea that we can and may even be expected to engage quickly and with less effort to large audiences of friends or acquaintances (Facebook, Reddit). There is even now a belief that our voices can or should be heard by the public or by those who we would not otherwise have access to (Twitter). These are just a few of the more common examples. The perception of the nature of information and information-exchange seems also to be changing, though again with caveats as to the extent. For instance information is no longer static, evolutionary but slow moving (encyclopaedias, books, libraries), and is instead malleable or even fleeting (wikis, forums, semantic web searches). Mono- or one-way consumption has been replaced by immediately dialogical, information-manipulating (editing, creating) interaction. Information is not an endgame, and though the process of information gathering may be dynamic (the idea of being wed to one newspaper, for instance, is no longer as common as it was), but there is reason to doubt that there have been substantial changes in our perceptions of information as something that is accurate or definitive. The proliferation of false celebrity-death stories is only one such reason for caution,[6] which sits uneasily alongside the scepticism of the unreliability of what is read on the web.

Of most interest for this paper are the changes in relationship formation and development. Online relationships mirror analogue engagement in some ways, and can be fleeting, long distance, or entirely non-physical [3, p. 6]. If we accept that identity is formed dialogically however, we must question the impact of whatever changes there are. Discussion about the so-called *filter bubble* is one such example. As Pariser [14] explains, the algorithms employed by internet search engines narrow searches according to user history. Thus ensuring you are likely to see more of the same each time you search. Filter bubbles are also self-perpetuating. In our choices of Twitter followers, Facebook friends, Reddit sub-groups, we share and follow those who we perceive to share affinity for our interests, beliefs, and ideas. This is not always true of course, and some may actively seek out antagonistic or opposing parties or opinions, but this is certainly not a given. At this stage it also seems increasingly less likely. With the rise of the *safe space* in UK university campuses (and even with the backlash against these, whether in the name of liberalism or free speech)[7] the mechanism for deciding whose voices are heard and by whom seems to be following a trend of narrowing rather than expanding, and it's perhaps not surprising. Arguments can be fun of course, but in friendships people seek common ground (even if the common ground is a love of argument). That such tendency would be mirrored online is unsurprising.[8]

This is important when we think about dialogical identity formation. If identity is indeed formed *in response to, because of* or even *in spite of* the way in which others perceive us, the fact that we can manipulate what others perceive on the one hand (selfies are an excellent example of this), or delete those who do not view us as we might wish to be seen, on the other, means that the formation of identity may also be open to our own manipulation. This may not in itself be unusual or controversial. Groups of analogue friends are also self-selecting to some extent. But it is precisely the question of *extent* that matters here. Simply put, if I didn't like the views of those around me in a pre-digital age my choices were limited: physically remove myself from those people, or choose to ignore, adapt, respond, or confront the views that I faced. In digital dialogue the confrontation need not be so obvious (I can simply delete, block or otherwise silence such views), nor do I ever need to hear them at all, since I can unfriend, block or otherwise remove the access that those people have to me, or me to them. This can be long before they have the chance to offer the views that I might wish to avoid. Examples of people who unfriend or unfollow those with whom they disagree are not difficult to find. Thus an opportunity to define oneself in dialogue with, including in contrast with, those people antithetical to ourselves may be lost. If there is an impact of this, and even if this develops as a trend, remains to be seen.

In a broader sense how we *use* digital resources already affects the way in which an online identity is perceived by others. In the same way that we define an artist according to their

---

[6] Cf. http://www.nytimes.com/2012/09/20/fashion/celebrity-hoax-death-reports.html?_r=0
Also see attempts by some sites like Facebook to mitigate the impact of false information and news stories on their pages: http://www.reuters.com/article/2015/01/20/us-facebook-hoaxes-idUSKBN0KT2C820150120

[7] Cf. http://www.theguardian.com/education/2015/feb/06/safe-space-or-free-speech-crisis-debate-uk-universities

[8] There is of course more to be said about these ideas, and it is a topic to which I hope to return in the next incarnation of this paper.

engagement with, and usually production of art, someone who has a blog is a blogger. In this way the person becomes associated with a sub-culture of internet uploaders (or contributors). If, on the other hand you surf the internet without leaving more of a mark than the occasional status update or cookie trail, then you might be considered a downloader or lurker (or less flatteringly a *consumer*). Rather like the person who visits and consumes art but does not actively create art. The fluidity of such identities online is particularly noteworthy since each unique or individual interaction, with more or less anonymity can define an individual quickly and with more or less permanence. While overnight stardom in historically analogue terms was relatively infrequent, and normally included a lot of behind-the-scenes work and participation in a field—whether willing or otherwise—an overnight internet star or sensation can happen *overnight* in rather more of a literal way. This has been found to some cost by unwitting users, such as Justine Sacco, Lindsey Stone, and Adria Richards, all of whom used internet media to share their ideas and experiences, and all of whom faced quite serious backlash, bullying and smearing as a direct result.[9] Add to this *trolling* that includes sustained campaigns, or even identity appropriation or theft, and it becomes more and more apparent that in simple terms your identity online is up for grabs, for good or for bad. The possibility of anonymity is part of these trends, though it would be difficult to cite this as the only reason. While a person may be less likely to insult someone in the analogue world as online, this does not mean that they wouldn't do so. As an interesging aside, *anonymity* itself has lately been cemented as a grammatical person, sometimes even with proper noun capitalisation ("posted by Anonymous").

There are of course advantages to anonymity. Holloway and Valentine's research into the way in which young people engage with the internet [10, p. 133] found that anonymity allows "users to construct 'alternative' identities, positioning themselves differently in online space than off-line space." Identities, they further note, that are both *played with* and at times *abandoned.* This anonymity offers control, flexibility, as well as "time to think about what they want to say and how they want to represent themselves" [10, p. 134]. Despite this, they also found that the off- and online worlds of children are not utterly disconnected, but rather "mutually constituted" [10, p. 140]. It is easy to see the benefits this can bring, especially where such identities may be otherwise isolated, but the question of narrowing dialogical engagement once again remains unanswered. A positive example of where this support may be helpful in identity formation is for transgender identities that are otherwise less common in an analogue community. Yet there are other identities that can be perpetuated by online communities in ways that may be harmful, such as pro-ana sites, which promote eating disorders, and propagate myths about weight and health.

Palfrey and Gasser [1, p. 36] claim that "increasingly, what matters most is one's social identity, which is shaped not just by what one says about oneself and what one does in real space but also by what one's friends say and do." While the immediate impact of one's social identity may be more apparent, more permanent, or perhaps just more accessible, it is a misnomer to distinguish identity in this manner. Identity (according to the dialogical account) is at once always and necessarily social (cf. [17] for further discussion on the social aspect], at least in its formation, and perhaps the clearest differences are likely to be the overt and immediacy of the perception of such formation.

5 CONCLUSION

This paper has sought to engage in the conversation on digital identity, and in so doing has attempted to offer a picture of online identity that reflects the complexity and uncertainty that is not antithetical to pre-digital discussion of identity. To some extent the online identities that we construct (or are constructed for us) are, on the one hand, just another strand of what it is to be me or what it is to be you. On the other hand, the paper has tried to show ways in which the dialogical formation of identity may face challenges in the narrowing selection process of those dialogues, and from silencing the voices that are *other* in some way. The paper has sought to broaden the scope of the discussion on this topic. The hope is that it attracts the attention of many different voices (including dissenting or unconvinced), and that from this dialogue the identity of the paper can be expanded.

REFERENCES

[1] J. Palfrey and J. Gasser, *Born Digital: Understanding the first generation of digital natives*. New York: Basic Books, 2003.

[2] Case C-131/12, Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González. URL: http://curia.europa.eu/juris/document/document.jsf;jsessionid=9ea7d0f130d5b6c0ef0cfcc34664af65824af1275c09.e34KaxiLc3eQc40LaxqMbN4OaNmNe0?text=&docid=152065&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=433471 [accessed 20/03/15]

[3] D. Buckingham, Is there a Digital Generation? in Buckingham, D. and Willett, R. (eds.), *Digital generations: Children, Young People, and new Media*. Mahwah, NJ: Lawrence, Erlbaum Associates, 2006.

[4] R. Rorty, *Contingency, Irony, and Solidarity,* 1989.

[5] C. Taylor, *The Ethics of Authenticity,* 1992.

[6] Y. J. Erden and S. Rainey, Turing and the real girl: thinking, agency and recognition, in *The New Bioethics: A Multidisciplinary Journal of Biotechnology and the Body*, 18: 2, pp.133-144, September 2013.

[7] P. Markell, *Bound by Recognition,* Princeton: Princeton University Press, 2003.

[8] P. Redding, The Independence and Dependence of Self-Consciousness: The Dialectic of Lord and Bondsman in Hegel's Phenomenology of Spirit in *The Cambridge Companion to Hegel and Nineteenth-Century Philosophy*, CUP, 2008 (pp. 94 – 110).

[9] P. Gilbert and K. Lennon, *The world, the flesh and the subject: Continental themes in philosophy of mind and body*, Edinburgh: Edinburgh University Press 2005.

[10] S. L. Holloway and G. Valentine, *Cyberkids: Children in the Information Age*, London: RoutledgeFalmer, 2008.

[11] S. Greenfield, *Mind Change: How Digital Technologies Are Leaving Their Mark on Our Brains*. London: Rider Books, 2014.

[12] M. Prensky, Digital Natives, Digital Immigrants in *On the Horizon,* NCB University Press, 9: 5, October 2001. URL: http://www.nnstoy.org/download/technology/Digital%20Natives%20-%20Digital%20Immigrants.pdf (Accessed 20/03/15).

[13] Bennett, S. J., Maton, K. A. & Kervin, L. K. The 'digital natives' debate: a critical review of the evidence. *British Journal of Educational Technology*, 39: 5, pp. 775-786, 2008. URL: http://ro.uow.edu.au/cgi/viewcontent.cgi?article=2465&context=edupapers (Accessed 20/03/15)

[14] E. Pariser *The filter bubble: What the Internet is hiding from you.* London: Penguin, 2011.

[9] Cf. http://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html
The fact that all these examples are of women is not unintentional. While it would be untrue to say that *only* women experience online shaming, bullying or harassment, it *is* true to say that women face a disproportionate volume of such abuse. This reference purposefully does not include comment on whether such criticism as each received was deserved or not, since that is beyond the scope of this paper. For the purposes of my argument, what is of interest is the identity they forged, and that which was forged for them online.

[15] A. Clark and D. Chalmers, The Extended Mind in *Analysis* 58. pp. 10-23, 1998. Reproduced here: http://consc.net/papers/extended.html (Accessed 02/03/15).
[16] L. Wittgenstein, L. *Last Writings on the Philosophy of Psychology: Volume II: The Inner and the Outer* (G. H. von Wright & H. Nyman, Trans.). Oxford: Blackwell, 1992.
[17] I. Burkitt, *Social Selves: Theories of self and society*. London: Sage, 2008.

Projective Simulation and the Taxonomy of Agency

Léon Homeyer[1] and **Giacomo Lini**[2] [3]

Abstract. In this paper we focus on behaviourism and materialism as theory-driven approaches to the classification of AI and agency in general. We present them and we analyse a specific utility-based agent, the PS model presented first in [2], which has as its key feature the capability to perform projections. We then show that this feature is not accounted for solely by materialistic or behaviouristic stance but represents rather a functional link between the two approaches. This is at the same time central for agency. This analysis allows us to present a feature-driven (or reversed) taxonomy of the concept of agency: we sketch its main characteristics and we show that it allows a comparison of different agents which is richer than the solely behaviouristic and materialistic approaches. The reason for that lies in the fact that we have reversed the approach to agency from a theory-driven stance to a process-driven one.

1 Introduction

The notion of "agent" has a very broad spectrum of uses both in everyday life and in academic debates, such as in computer science, economics, or in the philosophical discussion on free will – to mention a few. In this paper we are concerned with the following question: *How can one distinguish and categorise different agents?*. In order to answer this question we need a taxonomy, and since we are addressing agency in general this taxonomy must not be bound by the origins of the specific agents – artificial or natural. In the following article we provide the outlines of a taxonomy of agency which supports such a holistic perspective. The philosophical interest of this topic is on the one side related to the fact that suggesting a holistic view often, if not always, has multiple applications, while on the other side the taxonomy we describe merges advantages and avoids pitfalls of behaviourism and materialism.

The paper is structured as follows. In section 2 we introduce two main theory-driven approaches to the classification of agency, namely behaviourism and materialism, and we highlight their distinctive features. In section 3 we consider a specific form of utility-based agent, the PS model, which has the capability to perform projections of itself into future situations. We argue that this feature cannot be accounted for solely by the presented proposals, but it can rather be considered as a functional link between those two perspectives. This characteristic allows us – in section 4 – to build a taxonomy for categorising different agents. By reversing the methodology of taxonomy building and concentrating on the feature of projection as a functional link, we suggest a perspective turnaround from "category $\longrightarrow$ features" to "features $\longrightarrow$ category ". We then close with some concluding remarks.

[1] Universtiy of Stuttgart, Germany, email: leon.homeyer@philo.uni-stuttgart.de

[2] Universtiy of Stuttgart, Germany, email: giacomo.lini@philo.uni-stuttgart.de

[3] This paper is fully collaborative, authors are listed in alphabetical order.

2 Theory-Driven Approaches to AI

Two theory-driven approaches contribute to the research of artificial intelligence in significant ways:

i Behaviourism as a connection to the role model of human intelligence and as a basis for assessing successful AI.
ii Materialism as the general proposal of founding higher order mental functions in physical structures.

In the following section we want to work out this meaning of behaviourism and materialism for AI and why they do not succeed on their own in giving a full-blown account of (artificial) intelligence.

2.1 Behaviourism

Behaviourism is an approach to psychology which does not refer to introspection and its mental phenomena directly in order to explain and predict human actions. By analysing the behaviour of an agent, a behaviourist reduces "mindfulness" to its consequences in behaviour. Behaviourism aims then at avoiding the metaphysics of mental entities while still explaining and predicting human actions.

The origins of the research endeavour of AI are intertwined with the theory of behaviourism. In his influential paper [10] Alan Turing stresses this connection by substituting his imitation game for the provoking philosophical question "Can machines think?". Turing's motivation was to reduce the phenomena of thinking to the behaviour of an agent in its environment. The imitation game itself is a behaviouristic test arrangement to the core. The system consists of an interrogator and two agents one of which is a machine. The task for the interrogator is to find out by questioning, through written communication, which of the two is the machine. The question "Can machines think?" becomes in this setting "Are there imaginable digital computers which would do well in the imitation game?" [10, p. 442].

It is important to note here that this central behaviouristic approach of AI construes intelligence as the successful interaction of an agent with its environment, while its physical realisation is considered irrelevant. Behaviourism considering AI enables us to map a vast variety of agents based on their stimulus-response patterns onto one scale. This approach promotes a continuum idea of intelligence, where different degrees of it can be derived from the agent's behaviour, without the burden of considering how intelligence is physically implemented.

Agents that seem to be ontologically heterogenic in terms of mindfulness become comparable from the behaviouristic stance. This leads to an evolving account of intelligence in AI research.[4]

[4] By concentrating on the interaction of agent and environment one can determine different degrees of success and the notion of intelligence becomes a gradual idea independent of its (meta)physical realisation.

2.2 Classification of AI

It is difficult to provide a unitary view on AI, since the term covers various research fields and questions, such as in computing, philosophy and psychology.[5] In [8], a definition of agency is provided by the authors, which we find to be very simple and at the same time not committed to any specific school of thought with respect to agency and artificial intelligence:

> An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors. [8, p. 31]

The extension of this idea in terms of agent-performances leads to the notion of an ideal rational agent. Given a performance measure for the actions of an agent, an ideal rational agent is able to perform such that its action maximize its performance, according to perceptions and built-in knowledge.

It is evident from that definition that rationality according to AI, although well defined, is a general concept: the reference to built-in knowledge implies the impossibility of defining a unified rationality criterion. A close look at the agent and the methods to describe its built-in knowledge are necessary elements in order to define the restricted criterion for rationality. According to the behaviour of the agent with respect to percepts, actions and goals [8], it is possible to identify four different instances of AI: simple-reflex agents, "keeping-track-of-the-world" agents, goal-based agents and utility-based agents.[6]

Simple-reflex agents get activated by stimuli in such a way that input and action are directly linked. These agents can perform well in a specific environment but are hard to program, because the more complex the environment gets the more effort one has to put into the hardwired behaviour in order to perform successfully in the environment. The success of its action is not a relevant part of the agents perception and unforeseen input tends to produce unsuccessful interaction, or no interaction at all.

Agents that keep track of the world introduce an intermediate step, where their environment (and past states of their environment) are represented as a state of the agent. Changes in the environment become relevant when analysing the input and the agent can react to more complex stimuli in sufficient ways.

Besides these past states of the environment, a goal-based agent also considers a (programmed) goal as part of his internal state. This goal describes a future state of the system that is desirable. Future states and the anticipated influence of the agent's actions now define the right activator. A behavioural description makes actions of the agent seem purposeful in a more abstract way. Complex actions, which involve a chain of actions and anticipated states of the environment, become possible.

From an outside perspective, the differentiation between a very detailed simple-reflex agent and a goal-based agent gets possible only when unforeseen environmental states are present. While a simple-reflex agent probably fails due to his missing hard-wired behaviour, the goal-based agent profits from the decoupling of desired behaviour and specific input. He can learn from the changes in the environment and pursue his goals on the collected information and anticipated future states.

By decoupling desired behaviour from specific output the abstraction-level of goals gets introduced and with it a variability of possible actions to achieve them. Goal-based agents might pursue their goals in weird and complicated ways and might therefore seem less efficient than a complex designed reflex agent from a behavioural perspective.

Utility-based agents encounter the problem of choice by considering side goals that determine the efficiency of an action. Utility matters when an agent has to choose between different actions to achieve his goal, when conflicting goals are present or the likeliness of anticipated future states has to be evaluated. In a changing environment, the process of evaluating possible outcomes of actions gets more complex and the effort of abstraction becomes crucial for success.

An essential feature in realising utility-based agents is that the internal states of the agent "can be of its own subject matter"[10, p.449]. In evaluating possible outcomes of actions, an agent has to consider the future state of the whole system. A self-representation in this sense is a central feature to create rational behaviour. Turing anticipated this quality and stated that "it may be used to help in making up its own programmes, or to predict the effect of alterations in its own structure. By observing the results of its own behaviour it can modify its own programmes so as to achieve some purpose more effectively"[10, p. 449]. The Projective Simulation Model developed in [2] we are going to discuss later is a proposal for realising a utility-based agent by embedding a self-representation through projection. A taxonomy that describes these different realisations of AI by degree can be partly realised by considering the performances of agents in their environment.

2.3 Materialism

The behavioural stance lacks the capability to assess how rational behaviour is produced, and it becomes difficult to compare different agents due to the limitation in observations. Besides AI research being an endeavour to produce an agent that *behaves* rationally in its environment, it has an inevitable *materialistic* component. In order to explain rationality, one has to ground intelligent behaviour in physical structures, hence one can interpret the materialistic understanding of AI as the simple fact, that when implementing AI, rational behaviour gets reduced to physical structures. An engineering process naturally begins (and ends) with a physical structure, in order to create rational behaviour in an artificial agent. Nevertheless, AI is undeniably guided by a higher-order notion of intelligence and rationality. It therefore joins materialism in reducing these notions to its physical basis. Human intellectual capacities are a role model for AI research and the insights into physical realisations of AI can guide our understanding of human rationality. It is important to note a distinction between mechanism and materialism, as Shanker highlighted in [9, p. 56]. While in a mechanistic sense the physical realisation of AI serves as an analogy for a psychological theory of the human mind, a materialistic AI approach would assume that human intelligence is actually computed in the same manner.

Although this distinction might be clear in theory, practice in neuroscience and AI provides us with another picture. It is equally hard to apply a strictly materialistic approach as well as a rigid behaviouristic stance. Both positions need to be informed by the other in order to gain significance in the domains of cognitive neuroscience or AI research. One might argue that the connecting elements of the two are mental entities, to begin with. Because that is what both theories wanted to avoid – behaviourism – or neglect – materialism – in the first place, bridging them via mental entities would corrupt their original intent.

[5] We thank anonymous reviewers for pinpointing this specific topic.
[6] See, again [8, pp. 40–45].

Nevertheless, what drives the research in this area is, at least partly, wondering about psychological features, e.g. intelligence. The bridging element that refers to these qualities is a functional understanding of mental phenomena. By reducing psychological phenomena to their functional role, functionalism establishes functional links between physical realisation and observed behaviour. In this sense functionalism is a materialistic informed behaviourism, or a phenomena-enriched materialism.

Let us consider learning as an example of this involvement and summarise its different levels:

- From a behaviouristic stance, learning is recognised via observing alterations in the behaviour of agents.
- A materialistic approach may consider neural networks in the brain as the deciding structures for mental phenomena. The challenge is then to connect changes in this structures with different kinds of behaviour.

The process of learning needs to be redefined by means of a function that enhances successful behaviour through strengthening the structure that led to it. This approach allows for a functional link, which is evident for example in Hebb's theory of learning [3]. Learning is defined by strengthening of cellular connections that have casual interdependencies. The more they fire together, the more likely their application gets in the future.

- AI research takes the functional link of learning and Hebbian theory as models, and employs mathematical tools when implementing the feature of learning into an agent.

3 Projective Simulation

In the following section we present a model which shows interesting features with respect to the characterization of agency offered in the previous section. The PS (Projective Simulation) model, is a simple formal description of a learning agent introduced in [2] which provides a new step into the characterization of intelligence in the field of "embodied cognitive science".

3.1 PS Model

A PS model is a formal automata-description able to perform some specific tasks. Its key feature is that the agent, in which the PS model is embedded, is able to project itself into future possible – even not occurred – situations, and to evaluate possible feedback received from the environment. Note that the evaluation is done before a real action is performed.

The procedure that allows the agent to perform the projective simulation can be described as follows. The environment sends an input – percept – to the agent, which elaborates it in order to produce an answer – action, output. After this exchange the environment provides feedback – which might be either positive or negative – and the agent updates its internal structure [2].

The analysis of the internal structure of the agent is necessary in order to understand its interactions with the environment. This will allow us to comprehend what projective simulation is, how it is implemented, and what its consequences are for the present study.

3.2 Agent Description

Given the above description of the overall system, we must clarify two points in order to furnish a suitable description of the agent:

- How does the elaboration of the percept allow the agent to perform an action?
- How does the incoming feedback allow the agent to update its internal structure?

The answer is given by describing the so-called ECM (Episodic and Compositional Memory). The ECM is defined as a stochastic network of clips, with lines connecting them. Every clip constitutes a node in the network and it is individuated by the couple $c = (s, a)$ where s refers to a percept and a to an actuator. Every clip is a "remembered percept-action". The lines connecting different clips are to be interpreted as the probabilities of passing from one to another; hence $p(c_1, c_2)$ individuates the probability that the agent in the state c_1 will switch to c_2. The process of projective simulation is implemented as a random walk through the ECM, which allows the agent to recall past events, and to evaluate fictitious experiences, before performing actions. The procedure of data elaboration is then reducible to the following steps:

- the agent gets a percept from the environment,
- the percept activates a random walk trough the ECM,
- via reaching a clip corresponding to a suitable actuator an action is produced.[7]

Turning our attention to the second question – regarding the updating of the internal structure of the agent – we should focus on the relationship between the feedback and the subsequent modification of the ECM.

Once the agent reaches a suitable actuator and performs an action, the environment sends a reward, either positive or negative, and this constitutes the evaluation of the performed action. The activity of updating the internal structure represents then the learning capacity of the agent. In the case of a specific percept-action sequence which is rewarded with positive feedback all of the transitions between different clips are modified according to some rule – for example Bayesian updating – in such a manner that all the probabilities between clips involved in the procedure that led to the action are enhanced, while others are normalised. To sum up, the evaluation of an action triggers a deterministic process of probability-updating that makes clips associated with positive feedback more "attractive".

3.3 Relevant Features

Initially, every pattern of the PS has the same probability to happen. When the agent gets a feedback from the environment it builds "some experience", and the updating process of probabilities in the ECM consists in a dynamic description that keeps track of experiences (previous or fictitious) as the main relevant element for future decisions. The relevance of the PS model for our research relies mostly in two specific features which are realised within the model.

- Decisions are taken not only according to previous experience, but also allow the agent to project itself into future possible situations.
- The agent shows compositional features – in terms of the creation of new clips – during its learning process.

The general concept underlying these two characteristics is the possibility for the PS model to create new clips; it is in fact the content of the created clip which allows us to make a distinction between

[7] For further characterization of the features we remand to [2] and [6] where performances of the PS model are tested in some applied scenarios. By "suitable actuator" here we refer to the definition given in [2, p. 3].

compositional and fictitious experience. In general, the process of creation is associated with parallel excitation of several clips, an idea which leads to the extension of the presented scheme in a quantum context, see [11] and [7]. This deterministic scheme is nonetheless sufficient to describe the process of clip-creation in the ECM: if two (or more) clips are activated during a projective simulation frequently and with similar probabilities it is possible to define a relative threshold for the involved clips: if the connection between them exceeds this threshold, they are then merged together into a new one.

This procedure – implemented in the PS model in e.g. [2, p. 12], [6] – allows us to understand how compositional features of the PS model emerge: given two clip associated with different actuators a_1, a_2 their merging gives a new clip, associated with an actuator a_3, which is obtained by means of composition.

Composition is also the key feature in order to understand fictitious projection. The creation of new clips can be defined in such a manner that actions of the agent are not only guided by previous experience; the agent can in fact create episodes which have not happened before, testing them according to the eventual reward given by the environment. The selection over all possible fictitious episodes are implemented then according to the confrontation with past rewards.

How does the idea of the creation of new clips constitute a relevant quality for both the behaviouristic and materialistic approach? On the one side it is evident from the previous discussion that the creation of new clips can be translated into new learning and acting behaviours – see, e.g. the composition case. On the other side, from a materialistic stance it is interesting to see that a structure with defined physical elements – the agent in the previously discussed case – "evolves" not only by stating a redefined compositional framework, but by also merging existent elements into new ones.

These two facets allow us to highlight the relevant role of the PS model in the agency/intelligence debate: it seems that the feature of projection constitutes a key element in order to build a taxonomy of agency, which – as we will see in the next section – guarantees several advantages over the solely behaviouristic or materialistic points of view.

4 A Broader View on Agency

In this section we focus on the relevance of the key feature of the PS model, namely its capability to perform projections, in order to comprehend to what extent it guarantees a broader understanding than the solely behaviouristic and materialistic stances. We provide then a feature-driven classification of the concept of agency, which we represent by means of an "empty" graph (fig.1) outlining the general structure of our taxonomy. This picture keeps projection as a central item, since we account for that by merging physical and behavioural aspects. We consider then three different instances of agency namely a standard non-projecting AI device, the PS model, and a human being. We locate them in our hierarchy and we analyse the resulting picture.

4.1 Projection and Behaviourism

If we consider behaviourism and its approach to AI and agency it is clear that the process which allows the agent to perform actions does not have any relevance, since what matters is just the final result.[8]

If we want to offer a broader overview of agency, this approach seems to be unsatisfying: even though it considers behaviour as a central feature, this position completely disregards the producing process of the behaviour itself. Two agents that perform with the same accuracy in a given scenario are indistinguishable according to behaviourism. But it is easy to imagine a situation in which the first agent works in a genuinely random manner without processing environmental inputs, and its accuracy is just determined by "luck", while the second agent processes the input in some specific manner in order to produce behaviour.[9] Alteration of behaviour has to be manifest in order to be considered according to behaviourism.

Projection, considered as a creative internal process [1], does not fit the constraint of being manifest, while it may modify final behaviour, and hence it can be regarded as an additional feature.

4.2 Projection and Materialism

Materialism constitutes the "other side of the moon" in the interpretation of AI, so to say. According to this position, we are solely concerned with the internal processes of the device that result in actions. The idea of projection is nevertheless not comprehensible, since according to this stance what is disregarded is the environment in which the agent is situated. The examination of physical realisation ends with the boundaries of the agent, while projection does not only involve internal states, since it considers possible environmental rewards. As we have seen while analysing the PS model and its description, the capability to perform projections constitutes a distinctive portrait of the agent and accounts for the produced action as an internal process; hence, again, it cannot be simply disregarded.

According to these two characterisation of the missing connections between behaviourism/materialism on the one side, and the capability to perform projections on the other, it is then evident that neither of the two research approaches to AI can account for agency and cope with projection as a key feature. The description of the PS model suggests that projection takes on a central role with respect to the categorisation of different agents; hence we provide a merged account which is concentrated on projection as a functional link – i.e. as a distinct feature which we cannot account for according to the separate views, but which is necessary in order to build a link between them – in order to sketch a taxonomy for AI.

4.3 Merging through a Functional Link

By merging both research stances together one gains the possibility to grasp the functional link between them and, therefore, also a broader view on intelligent agents. We want to promote a visualisation of the resulting taxonomy for intelligent agents as shown in the graph (fig.1).

Why should we be concerned with an empty graph?

- It provides us with the general outline and structure of the taxonomy we would like to promote: this graph allows us to show how projection as a functional link is dependent on both physical and behavioural features, as we will see in the example in sec. 4.4.
- By reversing the methodology of taxonomy building,[10] we take the need of explanation away from the categories of physical realisation and behavioural interaction, and we concentrate on the feature that defines the content of the taxonomy – i.e. the empty space of the graph, which is to be filled.

[8] The imitation game sketched in a previous section is a good instance of this concept.

[9] Although unlikely, this situation can be imagined and is hence possible.

[10] The reverse procedure goes from a "category → features" characterisation to a "feature → category" one.

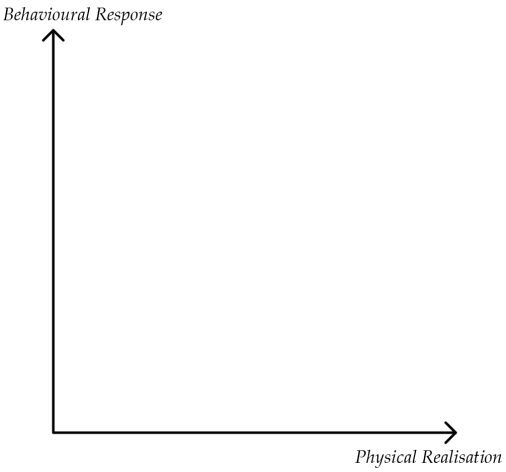


Figure 1. This graph represents a naïve visualisation of the idea of merging the behavioural response towards the environment and the physical realisation of the agent. Note that this visualisation is not meant to represent a mathematical function, but it is rather a supporting element for comprehending the taxonomy.

Different agents can be distinguished according to their capability to perform projections. This function links behavioural interactions and physical realisations of the agents and defines the content of fig.1. While it is difficult to define qualities and quantities according to a theory-driven approach, the suggested feature- and process-driven taxonomy allows us to assign relevant scopes to both sides. With regard to the behavioural inquiry, this quality consists in the flexibility to cope with a changing environment or a rising complexity. The implementation of the capacity of projecting allows an agent to consider different actions and to anticipate future changes in the environment, both whether those changes are induced by the agent itself or by external sources. On the materialistic side, structures that represent the internal state of the agent become important. Feedback loops and other recursive structures are necessary to perform projections and enable self induced state-changes and -creation [5, p. 22 ff.].

By concentrating on the functional link of projection-performing, we are concerned with a second order quality, i.e. a quality which gets its ontological status not independently, but rather through the combination of behavioural interactions and physical realisations.

Even though a distinction based on these rather vague categories is difficult,[11] the benefit of our reversed taxonomy is twofold. It enables us to compare different intelligent agents originating from nature and AI, while at the same time it points to the direction of research in order to clarify the categories that amount to the functional link of projection. Instead of adopting a bottom-up approach which starts from well-defined aspects of agency (such as behavioural interaction and physical realisation) with the scope to categorize individual agents and the functions they perform, our reverse taxonomy takes a top-down view by identifying the functional link first, and then map different agents into a hierarchy, trying to connect the functional link to the "classical" categories.

[11] One can think at the following question as an example: "How could one give a unified measurement of the physical realisation of various agents?".

4.4 An Example

Let us consider three different sorts of agents. A standard non-projecting AI, a PS model and a human being. Our projection-based taxonomy offers a straightforward strategy to compare them. The PS model constitutes a step forward with respect to the non-projecting AI since it takes into account possible not-yet occurred events, which might be the objects of a projection. Still, the PS model does of course not realise human intelligence. According to our approach one of the reasons for this is that the PS model lacks the capability to simulate other agents. One of the distinctive traits of human intelligence is that they not only project themselves but also other agents into many different situations. Consider two different human agents Alice and Bob, such that Alice has some experience of how Bob behaves in a certain situation x. One of the distinctive traits of Alice as a human agent is that, facing the situation x, she has the possibility to ask herself the question "What would Bob do?" before acting and she can take a decision influenced by the evaluation of previous Bob's experience. The PS model lacks this "theory of mind" as a level of abstraction. This is one aspect that distinguishes humans from the other elements in our taxonomy.[12]

The possibility to distinguish those three different sorts of agents according to the functional link of projection allows us to display them into different levels as shown in fig.2. The resulting picture raises the question of how to connect elements represented on different levels. One can either think of the overall evolvement of agency as a set of discrete steps or as a continuous evolving "machinery". Fig.2 shows – among many others – two possible connection patterns for the three individuated levels.

Our argument for projective simulation as an essential functional link between behaviourism and materialism implicitly supports the idea that there is at least one discrete step in the evolvement of AI.[13] Nevertheless, we want to stress the fact that one of the main advantages of this approach is that it does not require any sort of commitment to specific schools in philosophy of science or ontology. In the first case, one can address both a discontinuous perspective in the evolution of science, see e.g. [4], as well as a continuous one. The two lines represent those two approaches. Ontologically, discontinuous steps in fig.2 may as well be read as qualitative gaps between AI and humans, while the continuous picture provides the possibility to think of them as being in the same ontological category.

5 Conclusion

In this paper we have shown why two main theory-driven approaches of AI, i.e. behaviourism and materialism, do not succeed on their own in giving a full-blown account of (artificial) intelligence. This was also done by presenting the PS model, a form of utility-based agent which has the capability to perform projections. We have argued that this key element constitutes a functional link between the two theory-driven approaches.

The overall analysis allowed us to introduce a feature-driven (or reversed) taxonomy of the concept of agency, which gives a broader and richer view on intelligent agents. We provided a general scheme for the distinction of different agents according to their capability to perform projections. This perspective considers both behavioural interactions and physical realisations, via the identification of flexibil-

[12] We are of course aware that there are many other missing items in order to simulate human intelligence with a PS model. It is the present scope that requires us to individuate projection as the key feature.

[13] This argument supports the overall discrete picture in an inconclusive manner. This topic is the subject of further research.

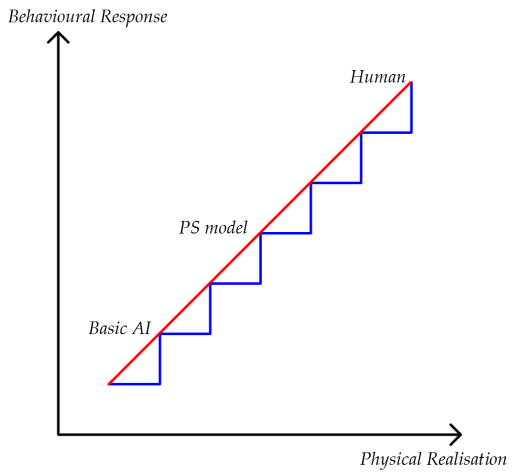


Figure 2. A representation of the comparison of non-projecting AI, PS model and human agent. Note that many patterns allow to connect those three distinct points, leaving open the question whether this should be a continuous or discrete "evolution".

ity in interactions on the one side and the possible physical structures and their complexity on the other. This conclusion is supported by giving an example and comparing different agents according to the individuated functional link. The emerging question of how the evolution between different realisations of AI should be understood is briefly sketched and constitutes a possible follow-up research question, but we have argued in this paper that our approach seems to not require any ontological or epistemological commitment.

ACKNOWLEDGEMENTS

The authors wish to thank Ulrike Pompe-Alama, Thomas Müller and Tim Räz for comments and discussion of a previous draft of this paper. We also wish to acknowledge the two anonymous reviewers for their helpful comments. Authors take full responsibility for every mistake in the paper.

REFERENCES

[1] Hans Briegel, 'On Creative Machines and the Physical Origins of Freedom', *Scientific Reports*, (522), 1–6, (2012).
[2] Hans Briegel and Gemma De Las Cuevas, 'Projective Simulation fo Artificial Intelligence', *Scientific Reports*, (400), 1–16, (2012).
[3] D.O. Hebb, *The Organization of Behaviour. A Neuropsychological Theory*, John Wiley & Sons, New York, 1949.
[4] T. Kuhn, *The Structure of Scientific Revolutions*, Chicago University Press, Chicago, 1962.
[5] H. Maturana and F. Varela, *Autopoiesis and Cognition: The Realization of the Living*, D. Reidel Publishing Co., Dodrecht, 1980.
[6] J. Mautner, A Makmal, D. Manzano, M. Tiersch, and H. Briegel. Projective Simulation for Classical Learning Agents: a Comprehensive Investigation, 2013. Online at http://arxiv.org/abs/1305.1578.
[7] G. D. Paparo, V. Dunjko, A. Makmal, M. A. Martin-Delgado, and H. J. Briegel, 'Quantum Speed Up for Active Learning Agents', *Physical Review X*, **4**, 1–14, (2014).
[8] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice-Hall, 1995.
[9] S. Shanker, 'Turing and the Origins of AI', *Philosophia Mathematica*, **3**, 52–85, (1995).
[10] A.M. Turing, 'Computing Machinery and Intelligence', *Mind*, **59**, 433–460, (1950). doi:10.2307/2251299.
[11] Seokwon Yoo, Jeongho Bang, Changhyoup Lee, and Jinhyoung Lee, 'A Quantum Speedup in Machine Learning: Finding an N-bit Boolean Function for a Classification', *New Journal ofPhysics*, **16**, 1–15, (2014). doi:10.1088/1367-2630/16/10/103014.

Rationality in the Behaviour of Slime Moulds and the Individual-Collective Duality

Andrew Schumann[1]

Abstract. We introduce the notion of the so-called context-based games to describe rationality of the slime mould. In these games we assume that, first, strategies can change permanently, second, players cannot be defined as individuals performing just one action at each time step. They can perform many actions simultaneously. In other words, each player can behave as an individual or as a collective of individuals. This significant feature of context-based games is called individual-collective duality.

1 INTRODUCTION

In *Physarum Chip Project: Growing Computers From Slime Mould* [1] supported by FP7 we are going to design an unconventional computer on programmable behaviour of *Physarum polycephalum*, a one-cell organism that behaves by its plasmodium that is sensible to different stimuli called attractants, it looks for them and in case it finds them, it propagates protoplasmic tubes toward those attractants. These motions can be regarded as the basic medium of simple actions that are intelligent [1], [3], [4].

Notice that the *Physarum* motions are a kind of *natural transition systems*, $\langle \text{States}, \text{Edg} \rangle$, where States is a set of states presented by attractants and $\text{Edg} \subseteq \text{States} \times \text{States}$ is a transition of plasmodium from one attractant to another. The point is that the plasmodium looks for attractants, propagates protoplasmic tubes towards them, feeds on them and goes on. As a result, a transition system is built up. Now, labelled transition systems have been used for defining the so-called *concurrent games*, a new semantics for games proposed by Samson Abramsky. Traditionally, a play of the game is formalized as a sequence of moves. This way assumes the polarization of two-person games, when in each position there is only one player's turn to move. In concurrent games, players can move concurrently.

On the medium of *Physarum polycephalum* we can, first, define concurrent games and, second, extend the notion of concurrent games strongly and introduce the so-called *context-based games*. In these games we assume that strategies can change permanently. Another feature of context-based games is that players cannot be defined as individuals who perform just one action at each time step. They can perform many actions simultaneously. So, each player can behave as an individual or as a collective of individuals. This significant feature of context-based games is called *individual-collective duality*.

[1] Dept. of Social Science, Univ. of Information Technology and Management in Rzeszow, Sucharskiego 2, 35-225 Rzeszow, Poland. Email: andrew.schumann@gmail.com.

In this paper we will talk about the notion of rationality within context-based games.

2 ACTIONS OF PLASMODIA

Physarum polycephalum verifies the following three basic operations which transform one states to others in $\langle \text{States}, \text{Edg} \rangle$: fusion, multiplication, and direction. (i) The *fusion* means that two active zones (attractants occupied by the plasmodium) either produce new active zone (i.e. there is a collision of the active zones) or just a protoplasmic tube. (ii) The *multiplication* means that the active zone splits into two independent active zones propagating along their own trajectories. (iii) The *direction* means that the active zone is not translated to a source of nutrients but to a domain of an active space with certain initial velocity vector. These three operations can be examined as the most basic forms of intelligent behaviour of living organisms. For example, in the paper [4] we showed that the behaviour of collectives of the genus Trichobilharzia Skrjabin & Zakharov, 1920 (Schistosomatidae Stiles & Hassall, 1898) can be simulated in the *Physarum* spatial logic. This means that, first, a local group of Schistosomatidae can behave as a programmable biological computer, second, a biologized kind of process calculus such as *Physarum* transition system can describe concurrent biological processes at all.

The main result of our research is that, on the one hand, the Physarum motions are intelligent, but, on the other hand, they do not verify the *induction principle* (when the minimal set satisfying appropriate properties is given). This means that they can implement Kolmogorov-Uspensky machines or other spatial algorithms only in a form of approximation, because Physarum performs much more, than just conventional calculations (the set realised is not minimal), i.e. it achieves goals (attractants) not only by "Caesarian" straight paths.

Let us consider the following thought experiment as counterexample showing that the set of actions for the plasmodium is infinite in principle, therefore we cannot implement Kolmogorov-Uspensky machines. Assume that the transition system for the plasmodium consists just of one action presented by one neighbour attractant. The plasmodium is expected to propagate a protoplasmic tube towards this attractant. Now, let us place a barrier with one slit in front of the plasmodium. Because of this slit, the plasmodium can be propagated according to the shortest distance between two points and in this case the plasmodium does not pay attention on the barrier. However, sometimes the plasmodium can evaluate the same barrier as a repellent for any case and it gets round the barrier to reach the attractant according to the longest distance. So, even if the environment conditions change a little bit, the

behaviour changes, too. The plasmodium is very sensible to the environment.

Thus, simple actions of Physarum plasmodia cannot be regarded as atomic so that composite actions can be obtained over them inductively. In other words, it is ever possible to face a hybrid action which is singular, but it is not one of the basic simple actions. It is a hybrid of them.

In the transition system with only one stimulus presented by one attractant, a passable barrier can be evaluated as a repellent 'for any case'. Therefore the transition system with only one stimulus and one passable barrier may have the following three simple actions: (i) pass trough, (ii) avoid from left, (iii) avoid from right. But in essence, we deal only with one stimulus and, therefore, with one action, although this action has the three modifications defined above.

Simple actions which have modifications depending on the environment are called *hybrid*. The problem is that the set of actions in any labelled transition systems must consist of the so-called atomic actions – simple actions that have no modifications.

3 INDIVIDUAL-COLLECTIVE DUALITY AND NON-ADDITIVITY

In context-based games, we cannot use conventional probability theory. The matter is that if we assume the existence of hybrid actions, then the entities of games are certain and, therefore, cannot be additive.

The double slit experiment with the plasmodium of *Physarum polycephalum* is the best example of that conventional probability theory is unapplied for *Physarum* acts. Let us take the first screen with two slits which are covered or opened and the second screen behind the first at which attractants are distributed evenly. Before the first screen there is an active zone of plasmodium. Then let us perform the following three experiments: (i) slit 1 is opened, slit 2 is covered; (ii) slit 1 is covered, slit 2 is opened; (iii) both slit 1 and 2 are opened. In the first (second) experiment protoplasmic tubes arrive at the screen at random in a region somewhere opposite the position of slit 1 (slit 2). Let us denote all tubes landing at the second screen by A, thereby all tubes that pass through slit 1 by A_1 and all tubes that pass through slit 2 by A_2. Now we can check in case of *Physarum* if there is a partition of set A into sets A_1 and A_2. We open both slits. Then we see that the plasmodium behaves like electrons, namely it can propagates just one tube passing through either slit 1 or slit 2 or it can propagates two tubes passing through both slits simultaneously. In the second case, these tubes split before the second screen and appear to occur randomly across the whole screen. Thus, the total probability P(A), corresponding to the intensity of plasmodium reaching the screen, is not just the sum of the probabilities $P(A_1)$ and $P(A_2)$. This means that the plasmodium has the fundamental property of electrons, discovered in the double-slit experiment. It is the proof of non-additivity of probabilities.

Economics and conventional business intelligence tries to continue the empiricist tradition, where reality is measurable and additive, and in statistical and econometric tools they deal only with the measurable additive aspects of reality. They try to obtain additive measures in economics and studies of real intelligent behaviour, also. Nevertheless, there is always the possibility that there are important variables of economic systems which are unobservable and non-additive in principle. We should understand that statistical and econometric methods can be rigorously applied in economics just after the presupposition that the phenomena of our social world are ruled by stable causal relations between variables. However, let us assume that we have obtained a fixed parameter model with values estimated in specific spatio-temporal contexts. Can it be exportable to totally different contexts? Are real social systems governed by stable causal mechanisms with atomistic and additive features?

Hence, our study of context-based games on the medium of *Physarum polycephalum* can make impacts for many behavioural sciences: game theory, behavioural economics, behavioural finance, etc.

Non-additivity of phenomena does not mean that they cannot be studied mathematically. There are some rigorous approaches such as p-adic probability theory, which allow us to do it. The most significant feature of p-adic probabilities (or more generally, non-Archimedean probabilities or probabilities on infinite streams) is that they do not satisfy additivity. On the one hand, the p-adic analogies of the central limit theorem in real numbers face the problem that the normalized sums of independent and i.d. random variables do not converge to a unique distribution, there are many limit points, therefore there is no connection with the usual bell type curve. In other words, in p-adic distributions we cannot build up the Gauss curve as fundamental notion of statistics and econometrics. On the other hand, the powerset over infinite streams like p-adic numbers is not a Boolean algebra in general case. In particular, there is no additivity (we cannot obtain a partition for any set into disjoint subsets whose sum gives the whole set). Using p-adic (non-Archimedean) probabilities we can disprove Aumann's agreement theorem and develop new mathematical tools for game theory, in particular define context-based games by means of coalgebras or cellular automata. In these context-based games we can appeal just to non-Archimedean probabilities. These games can describe and formalize complex reflexive processes of behavioural finances (such as short selling or long buying).

Notice that the p-adic number system for any prime number p extends the ordinary arithmetic of the rational numbers in a way different from the extension of the rational number system to the real and complex number systems. The extension is achieved by an alternative interpretation of the concept of absolute value.

Let us suppose that the sample space of probability theory is not fixed, but changes continuously. It can grow, be expanded, decrease or just change in itself. In this case we will deal not with atoms as members of sample space, but with streams. The powerset of this growing set cannot be a Boolean algebra and probability measure is not additive.

We can consider *Physarum* behaviours within a certain topology of attractants and repellents as growing sample space. Assume that there are two neighbour attractants a and b. We say that there is a string ab or ba if both attractants a and b are occupied by the plasmodium. As a result, we observe a continuous expansion of the set of strings. It can be regarded as a sample space of probability theory. Its values will be presented by p-adic integers.

Let us show, how we can build up the sample space Ω^{ω} constructively. Suppose that Ω consists of $p - 1$ attractants and A, B, ... are subsets of Ω. Such A, B, ... are conditions (properties) of the experiment we are performing. For instance,

let A := “Attractants accessible for the attractant N_1 by protoplasmic tubes” and B := “Neighbours for the attractant N_1”, etc. Some conditions of the experiment, fixed by subsets of Ω^ω, do not change for different time $t = 0, 1, 2, \ldots$ Some other conditions change for different time $t = 0, 1, 2, \ldots$ So, we can see that the property B is verified on the same number of members of Ω for any time $t = 0, 1, 2, \ldots$ Nevertheless, the property A is verified on a different number of members for different time $t = 0, 1, 2, \ldots$ Thus, describing the experiment, we deal not with properties A, B, etc., but with properties A^ω, B^ω, etc. Let us define the cardinality number of $X^\omega \subseteq \Omega^\omega$ as follows: $|X^\omega| := (|X|$ for $t = 0$; $|X|$ for $t = 1$; $|X|$ for $t = 2$, ...), where $|X|$ means a cardinality number of X. Notice that if $|\Omega| = p - 1$, then $|A^\omega|$, $|B^\omega|$, and $|\Omega^\omega|$ cover p-adic integers.

The simplest way to define p-adic probabilities is as follows:

$$P(A^\omega) = |A^\omega| \text{ or } P(A^\omega) = |A^\omega| / |\Omega^\omega|$$

Notice that in p-adic metric, $|\Omega^\omega| = -1$

Agent *i*'s knowledge structure is a function $\mathbf{P}_i$ which assigns to each $a \in \Omega^\omega$ a non-empty subset of Ω^ω, so that each world a belongs to one or more elements of each $\mathbf{P}_i$, i.e. Ω^ω is contained in a union of $\mathbf{P}_i$, but $\mathbf{P}_i$ are not mutually disjoint. The function $\mathbf{P}_i$ is interpreted on p-adic probabilities.

$$K_i A^\omega = \{a : A^\omega \subseteq P_i(\omega)\}$$

The double-slit experiment with *Physarum polycephalum* shows that, first, we cannot extract atomic actions from all the kinds of the plasmodium behaviour, second, probability measures used in describing this experiment are not additive. We can deal just with hybrid actions.

The informal meaning of hybrid actions (e.g. hybrid terms or hybrid formulas) is that any hybrid action is defined just on streams and we cannot say in accordance with which stream the hybrid action will be embodied in the given environment. It can behave like any stream it contains but there is an uncertainty how exactly.

7 CONCLUSIONS & FUTURE WORK

Thus, context-based games on the medium of *Physarum polycephalum* can have many impacts in the development of unconventional computing: from behavioural sciences to quantum computing and many other fields.

So, if we perform the *double-slit experiment* for *Physarum polycephalum*, we detect self-inconsistencies showing that we cannot approximate atomic individual acts of Physarum as well as it is impossible to approximate single photons. From the standpoint of measure theory, it means that we cannot define additive measures for *Physarum* actions. In our opinion, it is a fundamental result for many behavioural sciences. Non-additivity of actions can be expressed in different ways: (i) *natural transition systems, such as Physarum behaviour, cannot be reduced to Kolmogorov-Uspensky machines*, although their actions are intelligent, (ii) *there is an individual-collective duality, when we cannot approximate atomic individual acts* (an individual, such as plasmodium, can behaves like a collective and a collective, such as collective of plasmodia, can behaves like an individual).

ACKNOWLEDGEMENTS

This research is supported by FP7-ICT-2011-8.

REFERENCES

[1] A. Adamatzky, V. Erokhin, M. Grube, Th. Schubert, A. Schumann, A. Physarum Chip Project: Growing Computers From Slime Mould, *Int. J. of Unconventional Computing*, 8(4): 319-323, (2012).

[2] A. Schumann, Payoff Cellular Automata and Reflexive Games, *J. of Cellular Automata*, 9(4): 287-313, (2015).

[3] A. Schumann, L. Akimova, Simulating of Schistosomatidae (Trematoda: Digenea) Behavior by Physarum Spatial Logic, *Annals of Computer Science and Information Systems, Volume 1. Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*. IEEE Xplore, (2013), 225-230.

[4] A. Schumann, K. Pancerz, Towards an Object-Oriented Programming Language for Physarum Polycephalum Computing, [in:] M. Szczuka, L. Czaja, M. Kacprzak (eds.), *Proceedings of the Workshop on Concurrency, Specification and Programming* (CS&P'2013), Warsaw, Poland, September 25-27, (2013), 389-397.

Reasoning, representation and social practice

(extended abstract)

Rodger Kibble[1]

Abstract. The idea that human cognition essentially involves symbolic reasoning and the manipulation of representations which somehow stand for entities in the real world is central to "cognitivist" approaches to AI and cognitive science, but has been repeatedly challenged within these disciplines; while the very idea of representation has been problematised by philosophers such as Dreyfus, Davidson, McDowell and Rorty. This extended abstract discusses Robert Brandom's thesis that the representational function of language is a derivative outcome of social practices rather than a primary factor in mentation and communication, and raises some questions about the computational implications of his approach.

1 Introduction

"Where do correct ideas come from? Do they fall from the sky? Are they innate? No, they come from social practice".
Mao Zedong, "On Practice".

What Varela et al [13] labelled "cognitivism" (also known as the Computational Theory of Mind or CTM) is an approach to AI and cognitive science that postulates symbolic representations as fundamental to cognition: representations are taken to be some kind of internal constructs that somehow stand for entities in the real world, and function as "arguments" for internal deductive reasoning. On this view, representations involve physical states of the organism, so cognitive processes must be associated with identifiable physical changes of state.

Some early critiques of the representational thesis from the standpoints of cognitive science and AI can be found in Varela et al [op cit] and Brooks [5]. Varela et al argue that the purported representations and operations that manipulate them are inaccessible to conscious (phenomenological) experience. Brooks reports on the development of systems which manifest intelligent behaviour but make no use of central representations; each layer or process in a system has access to relevant pieces of information, but it is only from a third-party observer's standpoint that the data can be interpreted as representing states of the real world. Varela et al class Brooks' work along with their own as belonging to the (then) new *enactivist* paradigm.

Representationalism has also taken a battering within 20th century analytic philosophy (see [8,11] for discussion). In this extended abstract we consider whether the "analytic pragmatism" of Robert Brandom [1,2,3,4] can offer a bridge between enactivist approaches and representational schemes. Brandom argues that while language does have an essentially representational dimension, this should not be considered as its primary function but can be best captured within the context of discursive social practices (see [6,11]). In the course of these practices, language users assume responsibility and authority for their various claimings while attributing and ascribing both doxastic (propositional) and practical commitments and entitlements to themselves and others. Representations and symbolic reasoning are not primary or causal, but are a means of characterising invariants in (material) inferential reasoning. Brandom sets out to show how one can develop accounts of linguistic meaning and purposeful action which are grounded in normative social practice, eschewing semantic or intentional concepts, and in particular how formal logic can be shown to be grounded in everyday linguistic practice

Brandom is classed by Joseph Rouse as a "practice theorist" ([12]; see [7] for discussion), and this aspect of his work seems to offer a good fit with the enactivist stance. Practice theory is a term that has been applied to a variety of approaches (or practices?) in the social sciences and humanities. What these approaches have in common is that theyseek to study the behaviour of individuals in

1. Department of Computing, Goldsmiths University of London. Email: r.kibble@gold.ac.uk

social contexts by focussing on habitual performances classed as practices against a background of other practices, in place of such monolithic categories as culture, class, gender, rules, values, norms and so on. One motivation for this is that analysts can focus on observable events rather than postulating unobservable entities such as beliefs, values or traditions, or speculating about the psychology of the participants' motives. In fact, in the course of Brandom's works it turns out that his discursive practices are assumed to rely on a fair amount of behind-the-scenes cogitation, which we consider in some detail in section 3.

2. Some key themes from Brandom

The essentials of the framework presented in [1] and [2] can be cursorily sketched as follows. Brandom claims to follow Kant and Frege in insisting on the primacy of the propositional, as the smallest linguistic unit for which we can take *responsibility*. To assert a proposition is both to take on a commitment to defend that assertion if challenged, and to claim an authority to which others may defer when making the same assertion. A commitment is understood here not as a state of mind but as a social status, which is constituted by the normative attitudes of one's interlocutors. Participants in a dialogue are taken to maintain "deontic scoreboards" with a record of claims to which each participant has committed themself, consequential commitments which the scorekeeper derives by (material) inference, and commitments to which the scorekeeper judges the speaker to be entitled [1:190ff].

It is important to note that the commitments that a speaker will acknowledge may not match those that will be attributed by scorekeepers: in particular the scorekeepers may calculate *consequential* commitments of which the speaker is unaware. This is claimed to capture a difference between two senses of "belief": what one is aware of or will admit to believing, and what follows (logically or otherwise) from one's avowed beliefs. Levesque [9] sought to capture this distinction with a "logic of implicit and explicit belief", while Olsen [10] argues that Brandom's notion of consequential commitments enables us to handle these phenomena, in particular the problem of "logical omniscience", without resorting to non-standard logics.

"Inference" here is meant as "material" or content-based inference as in: Edinburgh is to the East of Glasgow, so Glasgow is to the West of Edinburgh. According to Brandom these inferences are immediate, and do not rely on an enthymeme or hidden premise or meaning postulate "X is to the East of Y iff Y is to the West of X". Rather, this biconditional *makes explicit* the implicit basis of the inference which acculturated users of a language make unthinkingly. The argument is correct by virtue of the meanings or appropriate uses of the words, not because of some covert formal deduction. This leads up to Brandom's logical expressivism: logical reasoning supervenes on material inference, in that an argument is considered to be logically good just in case it is materially good, and cannot be made materially bad by any substitution of non-logical for non-logical vocabulary in its premises or conclusion [2:55].

Finally (for the purposes of this abstract) material inference has a role to play in analysing the semantic content of subsentential expressions:

> "Two subsentential expressions of the same grammatical category share a semantic content just in case substituting one for the other preserves the pragmatic potential of the sentences in which they occur… a pair of sentences may be said to have the same pragmatic potential if across the whole variety of possible contexts their utterance would be speech acts with the same pragmatic significance…" [2:128-9].

So for example, one might say that two terms have the same denotation ("representation") if replacing one with the other makes no difference to the appropriate circumstances in which a speech act may be uttered and its pragmatic consequences, in terms of the speaker's deontic score (see [8] for extended critical discussion of this approach). Much of the second half of [1] consists of elaborations of this substitutional technique to handle the traditional subject matter of formal semantics such as reference, anaphora, deixis, quantification and propositional attitudes.

3. Processing implications of background practices

Having briefly outlined some key elements of Brandom's inferentialism, we now turn to some of the assumptions that seem to be made about the processing capabalities of communicating agents.

3.1 Scorekeeping

Chapter 4, Section IV of *Making it Explicit* includes detailed instructions for deontic scorekeeping, including the requirement that if speaker *B* claims that *p*, scorekeeper *A* *must* add *p* to the list of commitments attributed to *B* and *should* also add "commitments to any claims *q* that are committive-inferential consequences of *p*…" (my emphases). It appears from this that agents are obligated to be "perfect reasoners" when scorekeeping even if they are not when speaking. This seems to threaten to revive the issue of "omniscience", displaced onto the "scorekeeper" rather than the speaker, and has implications for the computational complexity of scorekeeping. Levesque [9] shows that for his formal system, the time taken to calculate what an agent believes grows linearly with the size of the KB (in the propositional case), while the time taken to calculate the implications of the belief grows exponentially. Of course these results do not necessarily carry over to Brandom's setup, but they are certainly suggestive.

Furthermore, the status of scoreboards themselves and the practice of deontic scorekeeping seem somewhat uncertain. Scorekeeping is clearly not a directly observable practice, but is presumably meant to be manifest in the practical attitudes displayed towards utterances: one may for example **challenge** a speaker's entitlement to a commitment, or **endorse** it either explicitly (by repeating the claim) or implicitly (by remaining silent). The scoreboards themselves are only notional entities, with a troubling resemblance to *representations* within a quasi-formal system.

3.2 Substitution and expressivism

Kremer [8] questions Brandom's reading of Kant and Frege and offers a detailed examination of the decompositional strategy of analysing the content of subsentential expressions, and identifying different subcategories such as terms and predicates according to the contribution they make to the inferential potential of propositional utterances. For example: the fact that one can infer "Thora is a mammal" from "Thora is a dog", but not vice versa, indicates that *mammal* and *dog* are **predicates** which licence asymmetric substitution inferences, rather than **terms** which may license symmetric inferences [2:133ff]. Kremer argues that Brandom's account is plagued with circularity, since it claims to define syntactic categories in terms of substitution inferences but turns out (on Kremer's account) to assume a prior grasp of these very categories. One could add that the substitutional techniques are presented in rather general terms, using simple examples, and would constitute a formidable machine learning problem if applied to corpora of actual discourse. For one thing, it is unlikely that any corpus would provide instances of "all possible contexts" for any given sentence-pair (see above). This suggests some interesting directions for future applied research.

As noted above, the expressivist programme seeks to develop a notion of formal validity based on exhaustive substitution of nonlogical for nonlogical vocabulary. There is a persuasive argument that the ability to endorse material or content-based inferences such as "Brighton is to the east of Worthing, so Worthing is west of Brighton" does not necessarily presuppose a notion of "formally valid inference", as this threatens to set off a "regress of rules" of the kind depicted by Lewis Carroll in "Achilles and the Tortoise". However the substitutional approach also has its problems: no worked examples are presented, and the claimed parallels with other domains such as "theological vocabulary" are unconvincing [2:55]. Logical words like "if", so", "then" do not necessarily behave the same in all possible contexts, and a "fuzzy" or probabilistic approach may turn out to be more appropriate. The assumption that agents are capable of evaluating universal statements involving the entire non-logical vocabulary of a language is surely an idealisation.

4. Conclusion

Brandom's practice-oriented approach to language and purposeful action appears at first to offer theoretical support for non-cognitivist approaches to AI and cognitive science. This extended abstract has highlighted some computational and processing issues which argue against adopting the inferentialist model wholesale. The practices ascribed to individual language users turn out to rely on a complex and sophisticated analytical machinery which appears to require the processing resources of a cognitivist agent and makes idealised, perhaps unrealistic assumptions about agents' processing capabilities. As [7] argues, Brandom [3] essentially offers a "competence" model of an ideal speaker-hearer/scorekeeper rather than an "anthropological" account of actual practice: "Brandom's automata appear to be rather unconstrained both in terms of their internal operations and in the range of entities that can be discriminated as inputs or generated as outputs." Any restrictions are labelled as "psychological" and thus extrinsic to the explanatory model, though it is precisely these psychological restrictions which must be confronted if Brandom's model is to be pressed into the service of AI and cognitive science.

REFERENCES

[1] R. Brandom, *Making it Explicit*, 1994.
[2] R. Brandom, *Articulating Reasons*, 2000.
[3] R. Brandom, *Between Saying and Doing*, 2009
[4] R. Brandom, "Global anti-representationalism?" in *Expressivism, Pragmatism and Representationalism*, Huw Price et al., Cambridge University Press, 2013.
[5] R. Brooks, "Intelligence without representation", *Artificial Intelligence* 47, 1991.
[6] R. Giovagnoli, "The relevance of language for the problem of representation", in *Proceedings of the 50th Anniversary AISB Convention: Symposium on the Representation of Reality: Humans, Animals and Machines*. 2014
[7] R. Kibble, "Discourse as practice: from Bourdieu to Brandom", in *Proceedings of the 50th Anniversary AISB Convention: Symposium on Questions, discourse and dialogue: 20 years after Making it Explicit*. 2014.
[8] M. Kremer, "Representation or inference: must we choose? Should we?" in *Reading Brandom: on Making it Explicit*, eds B. Weiss and J. Wanderer, 2010.
[9] H. Levesque, "A logic of implicit and explicit belief", in *Proceedings of AAAI-84*, 1984.
[10] N.S. Olsen, "Logical omniscience and acknowledged vs consequential commitments." in *Proceedings of the 50th Anniversary AISB Convention: Symposium on Questions, discourse and dialogue: 20 years after Making it Explicit*. 2014.
[11] R. Rorty, "Robert Brandom on social practices and representations", in *Truth and Progress: Philosophical Papers volume 3*, 1998.
[12] J. Rouse. Practice theory. *Division I Faculty Publications. Paper 43,* 2007. http://wesscholar.wesleyan.edu/div1facpubs/43.
[13] F. Varela, E. Thompson and E. Rosch, *The Embodied Mi*nd, 1991.

Digital Footprints: Envisaging and Analysing Online Behaviour

Giles Oatley and **Tom Crick**[1] and **Mohamed Mostafa**

Abstract. Our long-term research goal is the development of complex (and adaptive) behavioural modelling and profiling using a multitude of online datasets; in this paper we look at suitable tools for use in big social data, specifically here on how to 'envisage' this complex information. We present a novel way of representing personality traits (using the Five Factor model) with behavioural features (fantasy and profanity). We also present some preliminary ideas around developing a scalable solution to modelling behaviour using swear words.

1 Introduction

There are large-scale research efforts in developing new and robust techniques for modelling online behaviour and identity. There exists numerous domains in which it is essential to obtain knowledge about user profiles or models of software applications, including intelligent agents, adaptive systems, intelligent tutoring systems, recommender systems, e-commerce applications and knowledge management systems [32]. The rise of Web 2.0 and social networking has facilitated the publishing of user-generated content on an exponential scale; its analysis is becoming increasingly important (and applicable) to the empirical study of society (and thus societal change).

Big datasets from social networking platforms are now being used for a multitude of purposes, alongside the obvious advertising, marketing and revenue generation; increasingly for government monitoring of citizens[2,3,4], along with covert security, intelligence community and military user profiling. However, the publishing of user-generated content on an exponential scale has significantly changed qualitative and quantitative social research, with its analysis becoming increasingly important to the empirical study of society. There are interesting sociological uses of studying or mining big social data, for instance exploring cyber-physical crowds using location-tagged social networks or the study of personality with large-scale benchmark social datasets and corpora.

However, this "big social data" from social media platforms, for instance social networks, blogs, gaming, shopping and review sites, differs significantly from more traditional/formal sources. With the advent of the social web, there are now orders of magnitude more data available relating to uncensored natural language, requiring the development of new techniques that can meaningful analyse it. This uncensored language is rich in 'unnatural' language (as opposed to 'natural' language, used in formal/traditional published media such as books and newspapers), defined as "*informal expressions, variations, spelling errors...irregular proper nouns, emoticons, unknown words*"[5]. We have been interested in profiling complex behaviours [20], particularly for crime informatics [22, 21] and in this paper we include in our models such bad behaviour that is found in big social data, for example so-called unnatural language with its poor language construction but also context dependent acronyms, jargon, "leetspeak" and swear words or profanity. Leet, also known as eleet or leetspeak, is an alternative alphabet for the English language that is used primarily on the Internet and in geek/cyber communities. It uses various combinations of ASCII characters to replace Latin script. For example, leet spellings of the word "leet" include *1337* and *l33t*; eleet may be spelled *31337* or *3l33t*. See Perea et al. [29] for an discussion of leet from a cognitive processing perspective.

2 Modelling Fantasy and Profanity

2.1 Rude Words: The Language of Pornography

A research project investigating opinions on a range of topics related to pornography usage was carried out; a web-based questionnaire received over five thousand respondents (n=5490). Several of the questions were open-ended, for instance how the person became involved with the subject of pornography, their particular interests and so on, eliciting a number of detailed responses (c.2000 words). From the initial findings [33], the data is ill-structured, with frequent usage of bad grammar and contains a large number of jargon (swear) words relating to pornography and sexuality.

An aim of the original study was the investigation of the usage of fantasy. This resonated with our general interest in determining behaviour from data, and so explored the language characteristics of the answers related specifically to fantasy. We analysed the respondents text using the psycholinguistic databases LIWC and MRC. The Dictionary of Affect in Language (DAL) [35] was also used, due to its specific uses for imagery-based language. We used methods derived from LIWC and MRC to determine personality traits and measures such as formality and deception. We wanted to get a general feel for the level of the text, and to see if there were any correlations between literacy and readability.

Initially we focused on the specific questions that might reveal something about the role of fantasy. For instance, among the many options for the question "*What are your reasons for looking at pornography?*", among the list were the following:

[1]All authors: Department of Computing, Cardiff Metropolitan University, UK; {goatley,tcrick,mmostafa}@cardiffmet.ac.uk

[2]Twitter Transparency Report 2014: https://transparency.twitter.com/

[3]Facebook Global Government Requests Report 2014: https://govtrequests.facebook.com/

[4]Google Transparency Report 2014: http://www.google.co.uk/transparencyreport/

[5]2nd Unnatural Language Processing Contest, part of the 17th Annual Meeting of the Association for Natural Language Processing (NLP2011): http://www.anlp.jp/nlp2011/

(A) "*To see things I might do*";
(B) "*To see things I can't do*";
(C) "*To see things I wouldn't do*";
(D) "*To see things I shouldn't do*".

The '*can't*' and '*wouldn't*' choices clearly indicate respondents utilising pornography more strongly as a form of fantasy. For this we explored the Five Factors personality traits, in particular expecting some correlation with the *Openness to Experience* factor (see Figures 1–4.

	A	B	C	D
A	1			
B	-0.72974	1		
C	-0.46635	-0.06469	1	
D	-0.33821	0.08321	0.091183	1

Table 1. Correlation between question items (where: A="*To see things I might do*"; B="*To see things I can't do*"; C= "*To see things I wouldn't do*" D="*To see things I shouldn't do*")

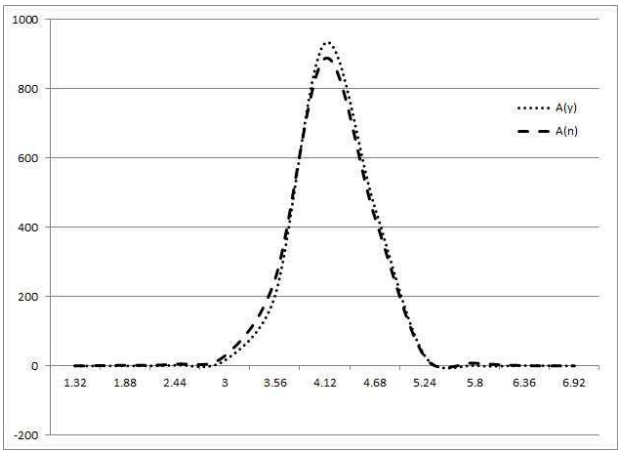


Figure 1. Openness to experience for A(y) (dotted) versus non-A (dashed)

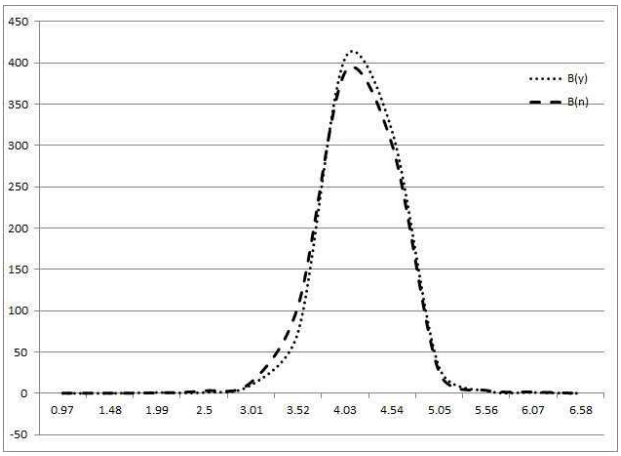


Figure 2. Openness to experience for B(y) (dotted) versus non-A (dashed)

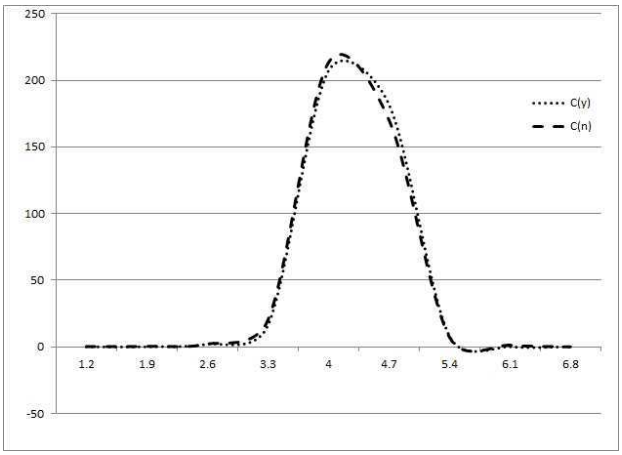


Figure 3. Openness to experience for C(y) (dotted) versus non-A (dashed)

Analysis is ongoing, with the results to be published in the near future; however there appears to be a strong negative correlation between participants who chose "*A. To see things I might do*" versus "*B. To see things I can't do*", as originally hypothesised. What was less convincing was our analysis of the Five Factors, and we put this down to the measures we used from [16] being derived from a very different corpus. We are currently concentrating on the lower level features from LIWC, MRC and DAL.

2.2 Disambiguating Profanity

WordNet[6] is a large lexical database of English; nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept, and each synset is interlinked by means of conceptual-semantic and lexical relations. Words that are found in close proximity to one another in the network are semantically disambiguated. WordNet Affect[7], a hierarchical set of emotional categories, and SentiWordNet[8], synsets are assigned sentiment scores (positivity, negativity, objectivity), are built on top of WordNet.

Millwood-Hargrave's study [17] for Ofcom (formerly, the Broadcasting Standards Commission), the UK's regulatory and competition authority for the broadcasting, telecommunications and postal industries, in 2000 was designed to test people's attitudes to swearing and offensive language, and to examine the degree to which context played a role in their reactions. Included in the report were attitudes towards swearing and offensive language 'in life', including a range of swear words and terms of abuse. Appendix 2's 'list of words' contained positions of the top swear words (categorised as "*very severe*", "*fairly severe*", "*quite mild*" and "*not swearing*") and their ranking from 1998 to 2000.

[6] http://wordnet.princeton.edu/

[7] http://wndomains.fbk.eu/wnaffect.html

[8] http://sentiwordnet.isti.cnr.it/

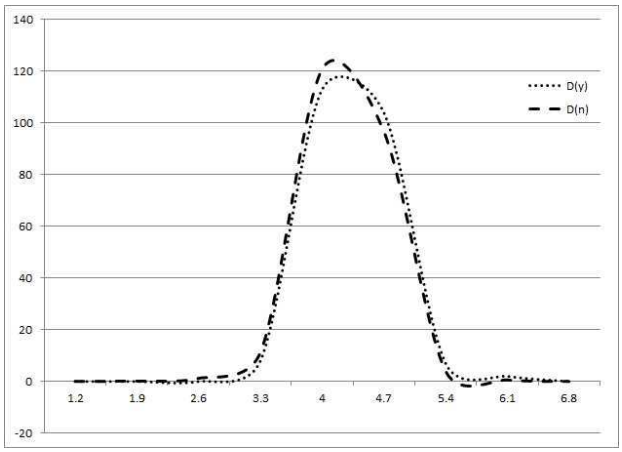


Figure 4. Openness to experience for D(y) (dotted) versus non-A (dashed)

The study of swear words has a longstanding position in linguistics, with the academic journal *Maledicta: The International Journal of Verbal Aggression* running from 1977 until 2005. Maledicta was dedicated to the study of the origin, etymology, meaning, use and influence of vulgar, obscene, aggressive, abusive and blasphemous language. Unfortunately we do not have resources such as databases in the literature; furthermore, WordNet does not contain the range of swear words we encountered in our data and is no use for disambiguating our text. Wikipedia, however, fared much better; but even better than these were Roger's Profanisaurus and Urban Dictionary.

Roger's Profanisaurus[9] is a lexicon of profane words and expressions; the 2005 version (the Profanisaurus Rex), contains over 8,000 words and phrases, with a further-expanded version released in 2007. Unlike a traditional dictionary or thesaurus, the content is enlivened by often pungent or politically incorrect observations and asides intended to provide further comic effect.

Urban Dictionary[10] is a Web-based dictionary that contains nearly eight million definitions as of December 2014. Originally, Urban Dictionary was intended as a peer-reviewed dictionary of slang or cultural words or phrases not typically found in standard dictionaries, with words or phrases on Urban Dictionary having multiple definitions, usage examples and tags.

We created different gazetteers related to rude words; one list was based on Wikipedia entries, and another on lists from Urban Dictionary. The Wikipedia list was created from link text on the Wikipedia porn sub-genre page[11] (link "anchor text" is a typical approach in semantic relatedness studies). This was comprised of 250 words. The Urban Dictionary list was created from the "sex" category[12] (by no means exhaustive – it is a fraction of the pornography-related terms in Urban Dictionary). This was comprised of 156 words. We implemented two metrics for rude words, the key idea of which is to have a simple mathematical model that enables us to estimate the life-history value of a token.

There are numerous other lists of pornographic words, which we compiled from miscellaneous sources; however, we are mainly interested in sources such as Wikipedia and Urban Dictionary as these are maintained by a similar community that uses the words in social networking. In this way we do not have to concern ourselves about this knowledge engineering process, merely concern ourselves about the representation and quality of meaning or definitions. We will in future work make use of the voting scores available on Urban Dictionary, and look to incorporate new resources such as Roger's Profanisaurus.

[9] http://www.viz.co.uk/profanisaurus.html

[10] http://www.urbandictionary.com/

[11] http://en.wikipedia.org/wiki/List_of_pornographic_sub-genres

[12] http://www.urbandictionary.com/category/sex

3 Psycholinguistic Models and Representing Complex Behaviour

Advances in psychology research have suggested it is possible for personality to be determined from digital data [28, 41, 15]. Recent studies [44] have suggested certain keywords and phrases can signal underlying tendencies and that this can form the basis of identifying certain aspects of personality. Extrapolating this suggests that by investigation of an individual's online comments it may be possible to identify individual's personality traits. Initial evidence in support of this hypothesis was demonstrated in 2012 by analysis of Twitter data for indicators of psychotic behaviour [34]. While in the past this has mainly been the textual information contained in blogs, status posts and photo comments [2, 3], there is also a wealth of information in the other ways of interacting with online artefacts. For instance, it is possible to observe the ordering/timings of button clicks of a user. Several researchers have looked at personality prediction (e.g. Five Factor personality traits) based on information in a user's Facebook profile [1, 14] and speech [9, 37], as well as also demonstrating significant correlations with fine affect (emotion) categories such as that of excitement, guilt, yearning, and admiration [18]. There are also several strands of related work based on the benchmark myPersonality Project[13] dataset [7], providing a platform for much-needed comparative studies.

Mairesse et al. [16] highlighted the use of features from the psycholinguistic databases LIWC [27] and MRC [43] to create a range of statistical models for each of the Five Factor personality traits [19, 26].

In previous work [20] we utilised these methods to develop a complex behavioural profile that included 'two faces' to model that we can have several different modes of operation (ego states). We performed our Five Factor analysis, and elaborated two sets of Five Factor results for each user. We chose Chernoff faces [8] for the visual representation. The Five Factors are displayed as five features on a stylised face, where:

- Width of hair represents *Conscientiousness*;
- Width of eyes represents *Agreeableness*;
- Width of nose represents *Openness to experience*;
- Width of mouth represents *Emotional stability*;
- Height of face represents *Extraversion*.

It should be noted that while researchers have continued to work with the Five Factors model, there are well known limitations [13, 25, 4] that are often overlooked by researchers. In particular, it has been criticised for its limited scope, methodology and the absence of an underlying theory. However, attempts to replicate the Big Five in other countries with local dictionaries have succeeded in some countries but not in others [36, 11]. While [10] claim that their Five Factors model "represents basic dimensions of personality", psychologists have identified important trait models, for instance Cattell's 16 Personality Factors [6] and Eysenck's biologically-based theory [12].

[13] http://mypersonality.org/

Figure 5. Two faces of a person. Personality traits from the Five Factors model are mapped on a Chernoff face (see later figure for specific trait mappings). Two different faces are drawn from two different linguistic sources, for the same person.

4 Envisaging Information

By analysing the myriad approaches of representing complex information, it is easy to be inspired by Tufte's clarity, precision, and efficiency [40, 39, 38]. We have integrated the profanity and fantasy behavioural features into our Chernoff face representing the Five Factor traits – see Figure 6 – represented on a Chernoff face are the Five Factors plus the additional behaviours for swearing level (darkness of blue colour on face) and fantasy level (amount of 'thought bubbles').

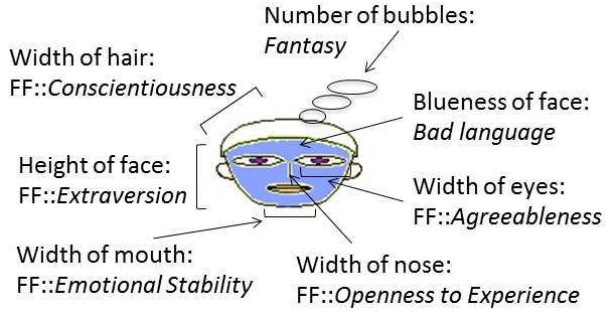


Figure 6. Traits and behaviours. Represented on a Chernoff face are the Five Factors (prepended by FF::) plus the additional behaviours for swearing level (darkness of blue colour on face) and fantasy level (amount of 'thought bubbles').

4.1 Modelling Timelines

Elsewhere we have presented ways to fuse social network (graph) information with geographical information [24, 23], and from spatial statistics there exists methods for space and time such as the Knox and Mantel indices. In this section we look at a method to represent temporal events, something very necessary when developing a behavioural profile.

Our data comes from an online portal for a European Union (EU) international scholarship mobility hosted at a UK university. The case study looked at how people interact with complex online information systems, the online portal for submitting applications. We analysed the document uploading behaviour (also motivation letters, and social media interactions) of the applicants. By examining the upload footprint for the users we determined several classes of behaviour.

There were several thousand applications submitted by over a thousand candidates, applying to 10 EU universities and 10 non-EU universities. Each mobility call has an opening date/time and closing date/time, with occasional extensions given for specific reasons (for instance due to administrative reasons or technical issues with the portal). Applicants are required to submit for their application certain mandatory files, such as motivation letter, passport/identification, curriculum vitae), as well as optional files (supporting documents).

We simplified an applicant's interaction, or timeline, with the portal to include the following milestones: *T0* Registration Time; *T1* First Action; *T2* Last Action; and, *T3* Submission. Additionally we represented an extension to the submission deadline as *T4* Extension. In this way we can represent an applicants interaction as shown in Figure 7, which shows seven example timelines.

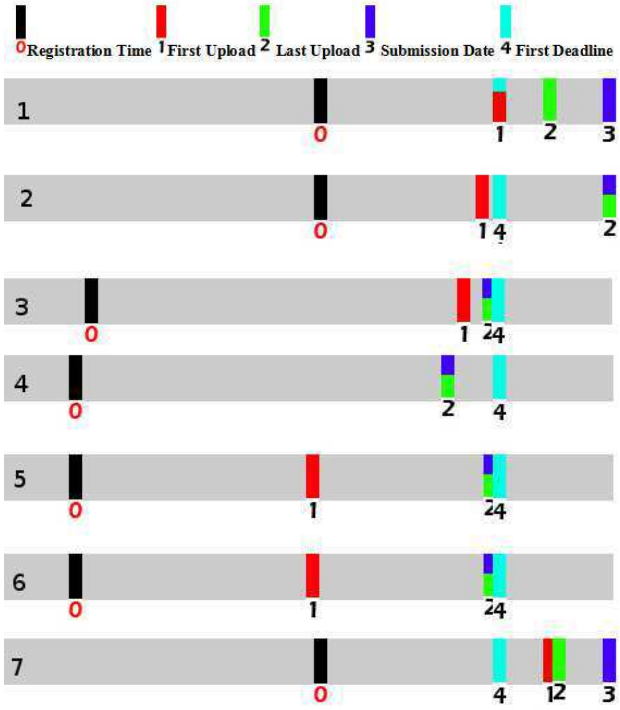


Figure 7. Seven user timelines. *T0* (black bar) is when the applicant first registered with the call. *T1* (red bar) represents when the applicant uploaded their first document, or First Action. *T2* (green bar) represents an applicants' Last Action. *T3* (blue bar) represents the applicants' Submission. *T4* (aquamarine bar) represents the first deadline (certain calls had initial deadlines extended).

Using these milestones we are able to identify interesting behaviours that compare and contract with personality traits and other sources of information. Behaviours such as: how long it was before an applicant became aware of the call, and when they registered; how long after registration did the applicant carry out their first action with the system; how long did they take to complete their application; and, how close to the deadline did they submit their application.

The complete timeline from opening to final close was 125 days. There was an extension from day 112 until day 125. We divided the timeline of the call into five equally spaced segments (S0-S4).

Using these segments we were able to assign the various applicant actions (*T0* Registration, *T1* First Upload, *T2* Last Upload, *T3* Submission) to various time periods. This allowed us to assign appli-

cants to statistically significant categories, and also to add in a few categories from observations. These are shown in the following Table 2; as you can see, a small number of applicants (n=4) registered within the segment S1 (20-40% of timeline), and then uploaded all of their documents and submitted within the segment S3 (60-90% of timeline). This is represented by Class A, the first row. Successive rows can be interpreted in the same manner.

Class	*n*	*T0*	*T1*	*T2*	*T3*
A	4	S1	S3	S3	S3
B	14	S2	S2	S2	S2
C	128	S2	S3	S3	S3
D	29	S2	S3	S4	S4
E	678	S3	S3	S3	S3
F	202	S3	S3	S4	S4
G	9	S3	S4	S4	S4
H	54	S4	S4	S4	S4

Table 2. Applicants' timeline actions assigned to segments

We did not want to ascribe a premature alias to the behaviours, as we recognise that there are several possible interpretations; nevertheless, we have used the 'Potential Alias' column in Table 3 to indicate some initial thoughts.

Combining this information with the earlier trait and behaviour model, it could be possible to present several faces along the timeline, or to represent the temporal aspect as a 'clock-type' metaphor, the straight line curved around, surrounding the face. The latter would perhaps be preferable, as we would expect that traits persist through time, but behaviours change. Likewise we would expect the blueness (rudeness) of the Chernoff face to change, and the amount of bubbles (fantasy) to change, but the facial features to remain constant (personality traits).

5 Conclusions and Future Work

The linguistic methods for determining personality traits are still in their infancy, and we have already noted some of the opposition to the lexical hypothesis [4]. Generally, information conveyed by the use of terms in human dialog studied in linguistics follows precise rules; other important rules are now introduced in the philosophy of language, investigating the meanings of terms and their extra-linguistic reference. We would expect that in time these additional information sources, like how people project identities through personal websites [41], judging people by their music preferences [30], personalisation of workspaces [42], etc, will all help with classification.

Further problems related to using social media for classification are that existing NLP tools are known to struggle with unnatural language: "*demonstrated that existing tools for POS tagging, chunking and Named Entity Recognition perform quite poorly when applied to tweets*" [31] and "*showed that [lengthening words] is a common phenomenon in Twitter*" [5], presenting a problem for lexicon-based approaches. These investigations both employed some form of inexact word matching to overcome the difficulties of unnatural language. We have made no attempt to use inexact string matching or to make use of a leetspeak parser. This will form part of future work.

The Web constitutes a world made of a precise formal-social ontology which hardly reflect the complexity of human personality; it is a difficult enterprise to try to mediate between the personal world of humans and the impersonal one of the Web. To assist with the ongoing knowledge modelling problem in this domain we recognise the need to utilise specific lexicons that keep pace with the language used, for instance the use of Urban Dictionary to resolve swear words. We thus need to study how in what precise manner this resource keeps pace with popular culture.

REFERENCES

[1] Mitja D. Back, Juliane M. Stopfer, Simine Vazire, Sam Gaddis, Stefan C. Schmukle, Boris Egloff, and Samuel D. Gosling, 'Facebook Profiles Reflect Actual Personality, Not Self-Idealization', *Psychological Science*, **21**(3), 372–374, (2010).

[2] Benjamin Blamey, Tom Crick, and Giles Oatley, 'R U :-) or :-(? Character- vs. Word-Gram Feature Selection for Sentiment Classification of OSN Corpora', in *Research and Development in Intelligent Systems XXIX, Proceedings of the 32nd SGAI International Conference on Artificial Intelligence (AI-2012)*, pp. 207–212. Springer, (2012).

[3] Benjamin Blamey, Tom Crick, and Giles Oatley, ''The First Day of Summer': Parsing Temporal Expressions with Distributed Semantics', in *Research and Development in Intelligent Systems XXX, Proceedings of the 33rd SGAI International Conference on Artificial Intelligence (AI-2013)*, pp. 389–402, (2013).

[4] Jack Block, 'The Five-Factor Framing of Personality and Beyond: Some Ruminations', *Psychological Inquiry*, **21**(1), 2–25, (2010).

[5] Samuel Brody and Nicholas Diakopoulos, 'Cooooooooooooooollllllllllllllllllllllllllll: using word lengthening to detect sentiment in microblogs', in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pp. 562–570, (2011).

[6] Raymond B. Cattell, *The description and measurement of personality*, Harcourt, Brace & World, 1946.

[7] Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski, eds. *Workshop on Computational Personality Recognition (Shared Task)*, 2013.

[8] Herman Chernoff, 'The Use of Faces to Represent Points in k-Dimensional Space Graphically', *Journal of the American Statistical Association*, **68**(342), 361–368, (1973).

[9] Cindy Chung and James W. Pennebaker, *Social Communication: Frontiers of Social Psychology*, chapter The psychological function of function words, 343–359, Psychology Press, 2007.

[10] Paul T. Costa and Robert R. McCrae, *Neo PI-R Professional Manual*, Psychological Assessment Resources, 1992.

[11] Filip De Fruyt, Robert R. McCrae, Zsófia Szirmák, and János Nagy, 'The Five-Factor Personality Inventory as a Measure of the Five-Factor Model: Belgian, American, and Hungarian Comparisons with the NEO-PI-R', *Assessment*, **11**(3), 207–215, (2004).

[12] Hans J. Eysenck, *Dimensions of Personality*, Routledge & Kegan Paul, 1947.

[13] Hans J. Eysenck, 'Four ways five factors are not basic', *Personality and Individual Differences*, **13**(6), 667–673, (1992).

[14] Jennifer Golbeck, Cristina Robles, and Karen Turner, 'Predicting personality with social media', in *Proceedings of Human Factors in Computing Systems (CHI'11)*, pp. 253–262. ACM Press, (2011).

[15] Francisco Iacobelli, Alastair J. Gill, Scott Nowson, and Jon Oberlander, 'Large Scale Personality Classification of Bloggers', in *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, volume 6975 of *Lecture Notes in Computer Science*, pp. 568–577. Springer, (2011).

[16] François Mairesse, Marilyn A. Walker, Matthias R. Mehi, and Roger K. Moore, 'Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text', *Journal of Artificial Intelligence Research*, **30**, 457–500, (2007).

[17] Andrea Millwood-Hargrave, 'Delete expletives?', Technical report, Research undertaken jointly by the Advertising Standards Authority, British Broadcasting Corporation, Broadcasting Standards Commission and the Independent Television Commission, (2000).

[18] Saif M. Mohammad and Svetlana Kiritchenko, 'Using Nuances of Emotion to Identify Personality', in *Proceedings of the ICWSM Workshop on Computational Personality Recognition*, (2013).

[19] Warren T. Norman, 'Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rat-

Class	Description	Potential Alias
A	Register early, and take some time to upload documents, but submit with plenty of time before deadline	EverythingEarly
B	Register reasonably early, but then upload documents and submit straight after with plenty of time before deadline, making no amendments	QuiteEarlyAndQuick
C	Similar to Class B, but submitting more slowly	Cautious
D	Registers reasonably early, and then takes time to upload, and only submits at the last days	VeryCautious
E	Latecomer to registration, but then uploads and submits quickly thereafter	Cautious
F	Latecomer to registration, but then uploads and submits slowly	Cautious
G	Latecomer to registration, but delays uploading and submission to last days	Cautious
H	Does everything at the last days, from registration to submission	EverythingLastMinute

Table 3. Description of each class

ings', *Journal of Abnormal and Social Psychology*, **66**(6), 574–583, (1963).

[20] Giles Oatley and Tom Crick, 'Changing Faces: Identifying Complex Behavioural Profiles', in *Proceedings of 2nd International Conference on Human Aspects of Information Security, Privacy and Trust (HAS 2014)*, volume 8533 of *Lecture Notes in Computer Science*, pp. 282–293. Springer, (2014).

[21] Giles Oatley and Tom Crick, 'Exploring UK Crime Networks', in *2014 International Symposium on Foundations of Open Source Intelligence and Security Informatics (FOSINT-SI 2014)*. IEEE Press, (2014).

[22] Giles Oatley and Tom Crick, 'Measuring UK Crime Gangs', in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE Press, (2014).

[23] Giles Oatley, Tom Crick, and Ray Howell, 'Data Exploration with GIS Viewsheds and Social Network Analysis', in *Proceedings of 23rd GIS Research UK Conference (GISRUK 2015)*, (2015). (in press).

[24] Giles Oatley, Kenneth McGarry, and Brian Ewart, 'Offender Network Metrics', *WSEAS Transactions on Information Science & Applications*, **12**(3), 2440–2448, (2006).

[25] Sampo V. Paunonen and Douglas N. Jackson, 'What is beyond the Big Five? Plenty!', *Journal of Personality*, **68**(5), 821–836, (2000).

[26] Dean Peabody and Lewis R. Goldberg, 'Some determinants of factor structures from personality-trait descriptor', *Journal of Personality and Social Psychology*, **57**(3), 552–567, (1989).

[27] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic Inquiry and Word Count. Erlbaum Publishers, 2001.

[28] James W. Pennebaker and Laura A. King, 'Linguistic styles: language use as an individual difference', *Journal of Personality and Social Psychology*, **77**(6), 1296–1312, (1999).

[29] Manuel Perea, Jon A. Dunabeitia, and Manuel Carreiras, 'R34D1NG W0RD5 W1TH NUMB3R5', *Journal of Experimental Psychology: Human Perception and Performance*, **34**(1), 237–241, (2008).

[30] Peter J. Rentfrow and Samuel D. Gosling, 'Message in a Ballad: The Role of Music Preferences in Interpersonal Perception', *Psychological Science*, **17**(3), 236–242, (2006).

[31] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni, 'Named entity recognition in tweets: an experimental study', in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pp. 1524–1534, (2011).

[32] Silvia Schiaffino and Analía Amandi, 'Intelligent User Profiling', in *Artificial Intelligence: An International Perspective*, volume 5640 of *Lecture Notes in Computer Science*, pp. 193–216. Springer, (2009).

[33] Clarissa Smith, Feona Attwood, and Martin Barker. pornresearch.org Prelininary Findings. Available from: `http://www.pornresearch.org/Firstsummaryforwebsite.pdf`, 2013.

[34] Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J. Park, 'Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets', in *Proceedings of the 11th International Conference on Machine Learning and Applications (ICMLA 2012)*. IEEE Press, (2012).

[35] Kevin Sweeney and Cynthia Whissell, 'A dictionary of affect in language: I, establishment and preliminary validation', *Perceptual and Motor Skills*, **59**(3), 695–698, (1984).

[36] Zsófia Szirmák and Boele De Raad, 'Taxonomy and structure of Hungarian personality traits', *European Journal of Personality*, **8**(2), 95–117, (1994).

[37] Yla R. Tausczik and James W. Pennebaker, 'The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods', *Journal of Language and Social Psychology*, **29**(1), 24–54, (2010).

[38] Edward R. Tufte, *Envisioning Information*, Graphics Press USA, 1990.

[39] Edward R. Tufte, *Visual Explanations: Images and Quantities, Evidence and Narrative*, Graphics Press USA, 1997.

[40] Edward R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press USA, 2nd edn., 2001.

[41] Simine Vazire and Samuel D. Gosling, 'e-Perceptions: Personality Impressions Based on Personal Websites', *Journal of Personality and Social Psychology*, **87**(1), 123–132, (2004).

[42] Meredith Wells and Luke Thelen, 'What Does Your Workspace Say about You? The Influence of Personality, Status, and Workspace on Personalization', *Environment and Behavior*, **34**(3), 300–321, (20062).

[43] Michael Wilson, 'The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2.00', *Behavior Research Methods, Instruments & Computers*, **20**(1), 6–10, (1988).

[44] Michael Woodworth, Jeffrey Hancock, Stephen Porter, Robert Hare, Matt Logan, Mary Ellen OToole, and Sharon Smith, 'The Language of Psychopaths: New Findings and Implications for Law Enforcement', *FBI Law Enforcement Bulletin*, (July 2012).

On the rationality of emotion: a dual-system architecture applied to a social game

David C. Moffat
Department of Computing
Glasgow Caledonian University, UK.
D.C.Moffat@gcu.ac.uk

Abstract. The insightful dichotomy between fast and slow thinking, as identified by Kahneman [5], is explored here with a simple model of a rational agent playing the Ultimatum Game.

It is an interesting game to model because it creates a social context between the two players that induces apparently irrational behaviour. One explanation for this is that the players react emotionally to each other in the game; and the emotions are irrational.

Consideration of the model leads to a conclusion that the irrational behaviour patterns can indeed be reproduced by the artificial agent; although the question of whether emotions are truly irrational is not resolved or even addressed here. Another conclusion is that the distinction between fast and slow thinking may not be the most important criterion to distinguish Kahneman's notions of system-1 and system-2. Instead, the related concept of precedence could be prior.

1 Dual systems of cognition — fast and slow

Kahneman has popularised the concept of what he calls system-1 and system-2 thinking, in his excellent book [5]. He is also to be credited for originating many of the key ideas that led the rest of the field in that direction.

To summarise the fundamental dichotomy that Kahneman raises, system-1 thinking is characterised by being *fast*, *intuitive*, *automatic* and *subconscious* or largely *impenetrable* to introspective analysis. On the other hand, system-2 thinking is relatively *slow*, *deliberate*, *logical* and *conscious*, costing *effort* to the thinker.

Into system-2 would go thinking about what to do tomorrow, for example; or solving a puzzle. In AI terms we could associate this kind of thinking with its symbolic, traditional approaches.

System-1 would be for thinking that is more closely following perception, and otherwise more closely coupled to the environment. Kahneman puts emotion into this category as well.

2 The Ultimatum Game

The Ultimatum Game (UG) is an artificial mathematical game that is used in laboratory experiments to probe participants' judgements of fairness in social interactions.

There are two players in the game: the *proposer* and the *responder*, and a sum of money that they have to split between them as follows. The proposer offers a split, which we may express as a percentage of the sum. The responder then chooses to accept the offer, or reject it. Accepting the offer means that both players get their part of the split; but rejecting it means that both get nothing.

For example, if the proposer offers 50% then the responder would surely accept it, and both players would get half the sum. But if the proposer offers much less, say only 4%, then any human responder is likely to reject it. It is easy to see why (if you are also human): the responder is angered by the tiny offer, in which proposer keeps nearly all the money for himself. However, that angry human has behaved irrationally, according to standard economic theory and mathematical game theory. The responder should accept any offer made to him, to maximise his gain in "utility", because even a tiny amount of money is better than nothing.

The fact that people are consistently and robustly "irrational" in this way is what makes the UG such an interesting game for researchers. Is it really true that humans are an inherently irrational species? Is it our emotions that make us irredeemably irrational? Or is there something deeply wrong with standard economic theory?

There are some indications in the literature that it is indeed emotion to be blamed, and probably the emotions of anger or disgust. For example, dosing participants with the oxytocin before they play the UG makes them less likely to reject the offer [9]. As oxytocin is a hormone that fosters affiliative feelings in mammals, (and as we are mammals,) the suggestion is that responders feel more forgiving toward the proposers, and are thus less inclined to punish them.

Let us explore the possibilities of modeling these emotional reactions towards other agents in social situations like the UG. First we construct an abstract architecture of a purely rational agent, in the form of a traditional symbolic-AI planning system. Then we shall add an emotional system to itm and see if it can be made to behave in the "irrational" manner of real humans playing the UG.

3 The Rational Algorithm

1. event perceived
2. maybe replan
   - if no current plan, then
     - maybe construct one from current state and goals
   - else (have current plan), so
     - maybe replan it if the new event was unexpected
   - also generate expectations of any events other than own actions
3. execute next action in the plan
4. repeat from (1)

As an example of a planning algorithm that we could plug into the architecture at line (2) above, we could use any conventional ap-

proach based on the traditional STRIPS representation for actions and events [2, 8]. This would represent the action to reject the offer, say, as having precondition that the proposer has made the offer of a certain percentage *offer(p)*, and postconditions that both players get no money (so *gets(proposer,0)* and *gets(i,0)*, where the agent refers to itself with the personal pronoun *i*).

Without going into more detail of how the agent's planner works, it would arrive at the plan to maximise the profit to the agent itself. This is the intuition that economists have regarding the UG, namely that the rational thing to do is to accept any money offered. We can therefore call that response rational (according to economists' typical views about rationality as maximising utility).

In addition we may assume that the planner deals with the possibilities of other events occurring in the world, that are not its own actions, by making predictions about their likelihood. Without specifying how this might be done, let us say that for our case the agent arrives at the reasonable expectation that the proposer will be "fair" and offer an approximately even split.

- The plan is to wait for the offer, and accept it.
- Expecting an offer around 50%.

With the architecture implied by this algorithm, the agent would perform as follows.

Run through:-

1. event perceived is that I have been offered 20%
2. offer was lower than expected, but still within the plan so continue without replanning
3. next action is thus to accept the offer
4. plan and execution and game terminated: I accepted 20%.

The resulting decision is considered the rational one by the rational actor position in economics. If the agent is offered only 1% or 2% it should accept it, as its aim is to maximise its financial gain. Let us now turn to an emotional variant of this architecture, and see if it might behave otherwise.

4 The Emotional Algorithm

We add in a capacity for (supposedly emotional) *reaction* to the architecture by inserting an extra step, which is (2) below. It occurs before the planner, but could also be after it, and before the plan actions are executed.

The emotional step considers the observed event as potentially relevant to its suite of possible reactions, and reacts accordingly. The reaction rules may be expressed in a similar language to the STRIPS language used above for other planning actions. However the difference is that the reactions rules are not planned; they are triggered, or activated by certain kinds of stimulus events.

An example of an emotional reaction would be for the agent to retaliate when it is hurt by another agent. How it knows that it has been hurt is an interesting problem left on one side here. This is the rule that is exemplified in the execution run below.

1. event perceived
2. maybe react to event
   - if I appraise the event in context as emotionally significant
     - then execute the relevant emotional reaction (in context)
   - maybe break and repeat from (1), to perceive action as new event.
3. maybe replan
   - if no current plan, then
     - maybe construct one from current state and goals
   - else (have current plan), so
     - maybe replan it if the new event was unexpected
4. execute next action in the plan
5. repeat from (1)

Just as with the rational algorithm, the plan is to accept the offer. The planner works in just the same way, even with the emotional component, because in this design, the emotions only occur as reactions to events. In advance of any events (including the offer made by the opponent), then, the same decisions are made as before.

- The plan is to wait for the offer, and accept it.
- Expecting an offer around 50%.

Run through:-

1. event perceived is that I have been offered 20%
2. that is much lower than expectated 50%, so feel pain
   - appraised that action of opponent has hurt me
     - general emotion of "anger" requires retaliation
     - to hurt opponent in context is achieved by rejecting offer
     - therefore reject it
     - and maybe continue to plan, but in this case we have ended.
3. game over, so no replanning
4. and neither is there any need to continue executing the current plan
5. plan and execution and game terminated: I rejected the 20% offer.

The addition of an emotional capacity into the architecture has changed the behaviour to what we would call irrational. The agent itself would have agreed with that assessment, at any time before its own emotional reaction.

Notice that the emotional agent has the same plan as before, and thus the same intentions to accept any offer. But the occurrence of an emotional reaction has upset its plans, presumably to its own consternation afterwards. Later, after punishing the opponent in this way, the agent may repent at leisure: "Oh, but I should have taken the money!"

5 On the reality of cognitive models

We have considered two alternative algorithms, one named rational and the other emotional. The emotional one gives a better account of human behaviour, and in that sense it is a better model. How realistic is it though, and can it be said to be a true model of the cognitive mechanisms inside the human brain?

The matter of models and realism is an interesting issue in the philosophy of science (or the methodology of cognitive science). An influential trichotomy was put forward by David Marr [6], in which he distinguished three levels of analysis which a model could inhabit. The top level is the *computational* level, where models emulate what the natural system (such as a human subject) is doing; how it behaves, and the ultimate (evolutionary) purposes for that behaviour. The middle level is the *algorithmic* one, where the way that the computation is performed is also intended or claimed to be an accurate model of how the natural organism does it. The lowest level is the *implementation* level, where the mechanisms that execute the specified algorithms are also intended to be authentic.

For the human case then, a cognitive model at the implementation level would need to be implemented in some kind of artificial neural network architecture. Artificial intelligence models, and models in cognitive science, are generally pitched at the computational or algorithmic levels. Dennett has described the general methodological approach of the cognitive sciences as a descent down these three levels, from an initially accurate computational model, down through the lower levels by specifying particular algorithms and then mechanisms that in turn should be verified by eventual experiments. This approach toward "reverse engineering" the human mind is what he has called the "intentional stance" [1].

These matters are still debated to this day in cognitive science. See, for example, an interesting discussion by Zednik and Jäkel in 2014 [10].

For an example of a similar sort of argument as the one put forward here, see the interesting account of wishful thinking given by Neumann et al [7]. In that study, the authors propose a model that accounts for some human behaviour by limiting cognitive resources. In other words, they put forward an algorithmic model to explain the phenomenon of wishful thinking. They claim not to have found the unique best algorithmic model, but only an interesting one that would be fruitful for further research. That is the sort of claim that I am making in this paper.

In relation to these levels of analysis then, where do the algorithms here stand? Firstly, they count as computational models, in which the emotional one is found to be superior because it matches human data better. But then: is the emotional algorithm also an accurate model at the algorithmic level of analysis? Not necessarily: that is not the claim in this paper.

The point about the emotional algorithm is that it is a *possible* algorithm that would account for the correct behaviour at computational level. To further validate it as the *only possible* algorithm would require further experimental work, of the type often found in cognitive science. But the fact that it is possible (i.e. consistent with human behaviour) does mean that it excludes claims of alternative algorithms to be uniquely accurate models. In particular, any alternative scheme in which parallel processes for cognition and for emotion (to be crude about it for now) cannot claim to be the best models, if a sequential model like the emotional algorithm presented above can also model behaviour.

Kahneman's dichotomy [5] into system-1 and system-2 types of thinking, that is fast and slow, is a scheme of the above sort. This is what leads me to conclude that the algorithms presented here show that his scheme is not necessarily correct. Rather than speed of thinking processes, in order to explain emotional behaviour as the winner in some cognitive race, we can use the priority or precedence of the two processes, in a sequential algorithm instead. In the emotional algorithm shown earlier, its relative speed had nothing to do with the behaviour patterns shown. Instead, it was that emotional process were simply consulted first, and took precedence over less emotional cognition.

It is not such a significant result as to change research directions in cognitive science; and it does not necessarily invalidate Kahneman's views in any crucial sense. However, it is a curious reminder of how easily we might overstep the mark in our interpretations of mental mechanisms.

This perspective also happens to be consistent with Frijda's notion of *control precedence*, [3], [4]. It was partly because of his term that I have referred to the emotion's precedence; and why I wrote the algorithm out so that the emotion would literally *precede* the later cognitions. What Frijda means by control precedence is not only that emotion takes priority over other cognition; but that it can do so even in the agent's knowledge that it is acting against its own interests. In that sense emotion takes priority over rational preference, as demonstrated in our simple examples earlier.

Some readers may wonder if that is always the case. An example of a scientist giving his research a high priority, although it is only a cognitive goal, might seem to contradict. However, in my personal experience as such a scientist with that high priority goal in life, I can attest to the irritating fact that my own efforts to do research are frequently interrupted and often ruined by emotions of all sorts. While I might say and believe that science is a high priority for me, the evidence is clear that it is not as high as even mundane emotions.

6 Conclusions

The simple architecture outlined here has demonstrated how a component that provides for a kind of *emotional reaction* can issue in behaviour that more realistically resembles human behaviour in the UG experiments. In contrast, the purely "rational" version of the architecture does not behave like a human when it is offered a tiny percentage. Real people reject such unfair offers, possibly because of a sense of unfairness; but in any case because of an emotional reaction.

The emotional version of the model here also rejects the tiny offers, if it has the appropriate rule to do so (which we might call "anger" or "retaliation").

One interesting issue that has been left out here is the matter of how the rule (which is presumably evolved in humans) becomes related to a specific context (like the UG, which can only be learned).

Regarding the matter of rationality, the architecture(s) give an account for why emotion is often seen as irrational, even by the agents that feel them and act upon them. The crux of the matter is that the emotions are unplanned; and that only the agents plans are to be regarded as rational. (Otherwise, why plan them in the first place? The intention to be rational is implicit in the act of planning.)

Regarding Kahneman's dichotomy between system-1 and system-2, it is clear that the planner is system-2 (along with most traditional, symbolic reasoning AI systems). The new entrant here is the emotion subsystem, which falls in the category of system-1 thinking. The emotional reaction shown is not deliberate (it was not planned), but instead rather automatic (when triggered by appropriate events). It is also relatively impenetrable to consciousness or subconscious, although it has a conscious facet in the experience of feeling, for those organisms that can feel their emotions.

The dichotomy between system-1 and system-2 holds up fairly well therefore; but for one surprising exception. In this case (at least) there is no great computational cost in the plans that are constructed, as the plan can only have one action in it. The search algorithm needed to construct the plan is therefore trivial in our example; and so we may reasonably take it that the planning process runs off about as fast as the emotional reaction does, and thus might even direct the agent's next action before the emotion does. But if so, why does the agent react emotionally? The answer is clear from the algorithm: the emotion step occurs earlier in the algorithm's cycle.

This is why the new (emotional) step was introduced at step (2), and not merely added onto the end. If it had been put after the plan execution step in our linear model, then emotions would never occur, as the algorithm would return to repeat at the first step immediately after performing an action (in order to observe its own behaviour). The emotional step takes priority because it literally precedes the other cognitive processes. This is consistent with Frijda's term of

"control precedence" which is one of his defining characteristics of emotion.

We are thus lead to the conclusion, from the architectures here, that the more fundamental distinction between system-1 and system-2 thinking is priority, or control precedence, and not speed as such (despite the title of Kahneman's beautiful book [5]).

7 Acknowledgements

Two anonymous reviewers raised some interesting queries that I have attempted to answer above. One was the nice paradox about the scientist with a high priority goal to do research.

REFERENCES

[1] Daniel Dennett, *The intentional stance*, MIT Press, 1987.
[2] R. Fikes and Nils Nilsson, 'Strips: a new approach to the application of theorem proving to problem solving', *Artificial Intelligence*, (2), 189–208, (1971).
[3] Nico H. Frijda, *The emotions*, CUP Press, 1986.
[4] Nico H. Frijda, *The laws of emotion*, Erlbaum, 2007.
[5] Daniel Kahneman, *Thinking, Fast and Slow*, Macmillan, 2011.
[6] David Marr, *Vision*, Henry Holt Co., 1982.
[7] Rebecca Neumann, Anna N. Rafferty, and Thomas L. Griffiths, 'A bounded rationality account of wishful thinking', in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, ed., P. Bello et al. Cognitive Science Society, (2014).
[8] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach (2nd ed.)*, Prentice Hall, 2003.
[9] Paul J. Zak, Angela A. Stanton, and Sheila Ahmadi, 'Oxytocin increases generosity in humans', *PLoS ONE*, **2**(11), e1128, (11 2007).
[10] Carlos Zednik and Frank Jäkel, 'How does bayesian reverse-engineering work?', in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, ed., P. Bello et al. Cognitive Science Society, (2014).

The Search for Computational Intelligence

Joseph Corneli [1] and **Ewen Maclean** [2]

Abstract. We define and explore in simulation several rules for the local evolution of generative rules for 1D and 2D cellular automata. Our implementation uses strategies from conceptual blending. We discuss potential applications to modelling social dynamics.

1 Introduction

This paper takes a local approach to studying the evolution of cellular automata (CA), following on the global approach of "PICARD" [24].

> *Like a traditional one-dimensional CA, PICARD executions move from one iteration to another by some rule. However, whereas traditional CA's require the rule to be static and externally specified, PICARD infers the iteration rule from the current state of the CA itself.* [24, pp. 1–2]

PICARD's inferred rule is derived from the current state of the CA by a global characteristic such as the number of 1's in the CA's current state (modulo 256), or the density ρ of 1's (normalised as $\rho/256$). These global criteria are similar to Van Valen's theory of resource density as an "incompressible gel" [29].

In the current paper we introduce the notion of a MetaCA, in which CA rules are derived locally at each cell within the CA as it runs. Examples appear in Figure 1. Here, each colour represents one of the 256 standard one-dimensional CA rules. States evolve locally, according to globally-defined dynamics.

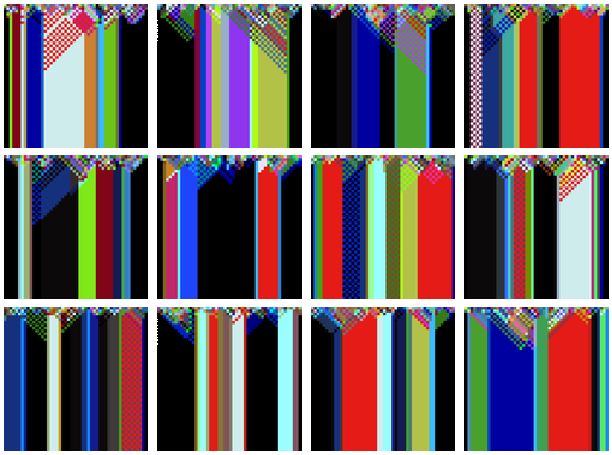


Figure 1. An illustration of MetaCA evolution

[1] Department of Computing, Goldsmiths College, University of London
✉ j.corneli@gold.ac.uk
[2] School of Informatics, University of Edinburgh
✉ ewenmaclean@gmail.com

2 Background

2.1 Cellular Automata

Each elementary 1D CA rule defines a mapping from all eight triples formed of 0's and 1's to the set $\{0,1\}$. Thus, for example the rule **01010100** is defined as the following operation:

$$\begin{array}{cccc} 0 & 0 & 0 & \mapsto \mathbf{0} \\ 0 & 0 & 1 & \mapsto \mathbf{1} \\ 0 & 1 & 0 & \mapsto \mathbf{0} \\ 0 & 1 & 1 & \mapsto \mathbf{1} \\ 1 & 0 & 0 & \mapsto \mathbf{0} \\ 1 & 0 & 1 & \mapsto \mathbf{1} \\ 1 & 1 & 0 & \mapsto \mathbf{0} \\ 1 & 1 & 1 & \mapsto \mathbf{0} \end{array}$$

The rules determine the next generation of a 1D CA locally, from three "parents". In the example, $\boxed{0}\boxed{0}\boxed{0} \mapsto \boxed{0}$ and $\boxed{0}\boxed{0}\boxed{1} \mapsto \boxed{1}$ and so on. There are 256 of these rule tables; the example above is Rule 84 in Wolfram's standard enumeration [31]. A crucial development in the history of CA research was the proof [5] that certain CA rules are Turing complete (in particular, Rule 110 enjoys this property).

Earlier classic works [14, 17, 23] exploring related "edge of chaos" effects. In [23, 17, 16], genetic algorithms are used to search the space of CA rules via crossover and mutation. This sort of evolution is global and is connected with the CA rule by a derived parameter, "Langton's λ" (cf. [14]). An overview of the "EvCA" programme is presented in [12]. CAs are also explored in two (and more) dimensions and with irregular topologies [7, 6]; in this paper, we develop both 1D and 2D examples. Closest to the work presented here is [26], which introduces the paradigm of *cellular programming*. As the name indicates, this approach is a fusion of ideas from cellular automata and genetic programming.

> *As opposed to the standard genetic algorithm, where a population of independent problem solutions globally evolves, our approach involves a grid of rules that coevolves locally.* [26, p. 74]

In cellular programming, local evolution of the CA rule makes use of a "fitness" metric ([26, pp. 79–81]), as the systems are evolved to perform certain global computational tasks. In the current effort system evolution is not directly guided by a specific fitness criterion. This paper defers any detailed *post hoc* analysis of MetaCA behaviour, although we hope to explore this further in a sequel, possibly following in the footsteps of the EvCA project [9, 10, 13].

2.2 Modelling social dynamics

Previous researchers have looked at CAs "as multi-agent systems based on locality with overlapping interaction structures" [6]. An

early application of cellular programming was to evolutionary game theory, a field with natural parallels (cf. [22]). We are inspired by recent work in this area on the evolution and failures of cooperation [1, 21, 27, 28] but we do not use a game theoretic approach. George Mead extends the term *social* to describe any scenario exhibiting emergent coevolution; this becomes central to our discussion.

> *What is peculiar to intelligence is that it is a change that involves a mutual reorganization, an adjustment in the organism and a reconstitution of the environment; for at its lowest terms any change in the organism carries with it a difference of sensitivity and response and a corresponding difference in the environment. . . . Now what we are accustomed to call social is only a so-called consciousness of such a process, but the process is not identical with the consciousness of it, for that is an awareness of the situation. The social situation must be there if there is to be consciousness of it.* [15, pp. 4, 48]

2.3 Conceptual Blending

One of our inspirations for working with cellular automata is that we are involved with a research project that studies computational blending [25], and cellular automata seem to offer a very simple example of blending behaviour. That is, they consider the value of neighbouring cells, and produce a result that "combines" these values (in some suitably abstract sense) in order to produce the next generation. We were also inspired by the idea of "blending" ordered and chaotic behaviour to produce edge-of-chaos effects. We propose to exploit existing formalisms of blending (in the style of Goguen [11]) in the context of cellular automata to investigate emergent and novel behaviours. The fundamental building blocks used in calculating concept or theory blends are:

Input Concepts are the concepts or theories which are understood have some degree of commonality (syntactic or semantic).

Signature Morphism is a definition of how symbols are mapped between theories or concepts.

Generic Space is the space which contains a theory which is common to both input theories.

Blend is the space computed by combining both theories. The computation is computed using a "pushout" from the underlying categorical semantics [18].

Once a blend has been computed, it may represent a concept which is in some way inconsistent. Equally it may represent a concept which is in some way incomplete. We can then either weaken an input theory, or refine the blend:

Weakening Given an inconsistent blend it is possible to weaken the input concept in order to produce a consistent blend. Weakening means removing symbols or axioms from the input concept.

Refinement Given a blend which represents a concept which is in some way incomplete, it is possible to refine the concept by adding symbols or axioms.

In this paper the primary examples have input concepts expressed in the same language, and indeed have the same specification. This means that the morphisms are not interesting and the calculated pushout could be computed without utilising the full machinery of category theory. Planned extensions will explore the idea of combining rules for cellular automata which may have entirely different techniques for expressing propagation (and we provide one example). For this reason, we target the Heterogeneous Tool Set (HETS) system [19] as an infrastructure for computing blends. We describe our current approach to blending in the context of cellular automata in Sections 3.2 and 3.3.

3 Implementation

3.1 Generating Genotypes

A MetaCA evolves a CA with 256 possible states – rather than the traditional $\boxed{0}$ and $\boxed{1}$ – where each state now corresponds to a "1D CA rule". By positioning three CA rules next to each other, we define a multiplication by applying the central rule bitwise across the alleles. For example, here is the result of "multiplying" $01101110 \times 01010100 \times 01010101$. In the context of such an operation, we refer to the central term as the "local rule." This example uses Rule 84 as the local rule, highlighted in bold.

0	**0**	0		0	*Apply local rule to "000"*
1	**1**	1		0	*Apply local rule to "111"*
1	**0**	0		0	*Apply local rule to "100"*
0	**1**	1	$\mapsto$	1	*Apply local rule to "011"*
1	**0**	0		0	*Apply local rule to "100"*
1	**1**	1		0	*Apply local rule to "111"*
1	**0**	0		0	*Apply local rule to "100"*
0	**0**	1		1	*Apply local rule to "001"*

Realised in a simulation with random starting conditions, the results of this operation are not particularly impressive: they stabilise early and do not produce any interesting patterns (Figure 2).

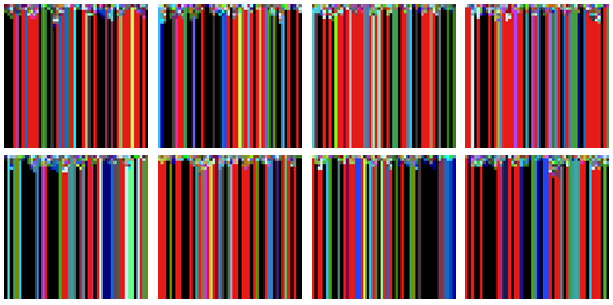


Figure 2. Under evolution according to the local rule without blending dynamics, a barcode-like stable pattern forms quickly

3.2 Introducing Blending

The blending variant says to first compute the "generic space" by noting the alleles where the two adjacent neighbours are the same, and where they differ. Only when the generic space retains some ambiguity (indicated by $\{0, 1\}$) do we apply the local rule (again recorded on the centre cell at left and highlighted in bold) in a bitwise manner across each allele, to arrive at the final result.

0	**0**	0		0		0	*Neighbours are both 0*
1	**1**	1		1		1	*Neighbours are both 1*
1	**0**	0		$\{0,1\}$		$\boxed{0}$	*Apply local rule to "100"*
0	**1**	1	$\mapsto$	$\{0,1\}$	$\mapsto$	$\boxed{1}$	*Apply local rule to "011"*
1	**0**	0		$\{0,1\}$		$\boxed{0}$	*Apply local rule to "100"*
1	**1**	1		1		1	*Neighbours are both 1*
1	**0**	0		$\{0,1\}$		$\boxed{0}$	*Apply local rule to "100"*
0	**0**	1		$\{0,1\}$		$\boxed{1}$	*Apply local rule to "001"*

For illustrative purposes, this blend has been formalised in the HETS system by introducing CASL files to represent the 8 bit encodings (Listing 1, and corresponding development graph shown in Figure 3). In this example, the first computed blend is inconsistent as there is not a unique value representing the output value of each function. In order to resolve this, we weaken the input rules in CASL by removing the function values which cause conflict. Note that purposes of efficiency, we have implemented our 1D experiments in LISP rather than in HETS/CASL. We've put the working code on Github[3].

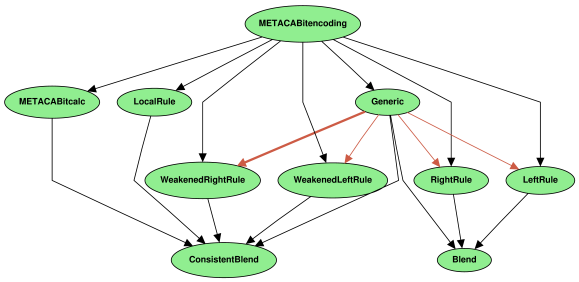
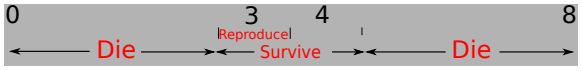


Figure 3. The development graph for calculating a blend of 8 bit encodings

3.3 2D Experiments

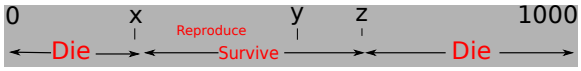
In order to extend the ideas presented so far in the 1D case, let us consider a variant of Conway's Game of Life [7], in which a global rule exists defining whether a square is alive or dead. We extend this by introducing the notion of a local rule at each square – a genotype, which governs the propagation of the phenotype.

In Conway's Game of life, one can view the rules for propagation as partitions on a finite interval $[0, 8]$.



The number on the line corresponds to the number of alive neighbours adjacent, in cardinal and inter-cardinal directions, to a given square. If the square is dead then it becomes alive (labelled reproduce) if the number of alive neighbours is exactly three. If there are five or more alive neighbours the square dies from overcrowding. If there are fewer than three alive neighbours the square dies from underpopulation. In all other cases the square maintains its status.

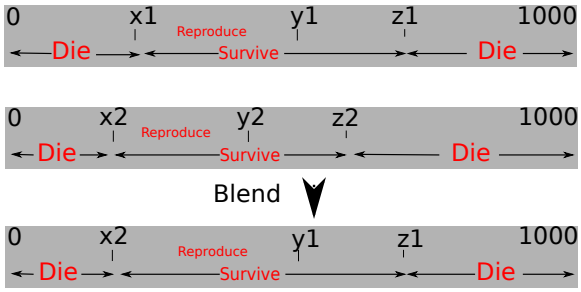
This can be generalised to partitions within a more finely grained line, for example from 0 to 1000, one creates a genotype (x, y, z):



We introduce the corresponding notion of a *weight* for each cell. The *phenotype* of the cell is then a pair $(alive, weight)$ which denotes whether the cell is alive, and what weight is has. In this paper we always calculate a newly propagated weight as the average of the neighbours' weights.

[3] https://github.com/holtzermann17/metaca

The notion of local propagation is introduced by allowing the genotypes to be blended at each point where a cell remains or becomes alive. As we have represented the genotype as a partitioned line, we can, for example perform a blend where the partition is blended in such a way as the minimise the lowest bound and maximise the highest bound, and maximise the interval for reproduction. Given two genotypes (x_1, y_1, z_1) and (x_2, y_2, z_2), the blend is $(\min\{x_1, x_2\}, \max\{y_1, y_2\}, \max\{z_1, z_2\})$:



Note that this is just one of several possible blending strategies, which we refer to as a *union* blend, since it maximises the partitions which pertain to survival. We consider alternative blends in our experiments.

4 Results

4.1 1D CAs

One of the first things we noticed was that even though the blending dynamic creates more interesting "CA-like" patterns than simple evolution according to the local rule (as illustrated in Figure 1), it also forms stable bands after this interesting initial period. In Figure 4, this is illustrated in a CA running with 500 cells over 500 generations. Figure 4 also includes a phenotype (in black and white) which is driven entirely by the genotype: that is, if the local genotype is $\boxed{\alpha}\boxed{\beta}\boxed{\gamma}$ where $\alpha, \beta, \gamma \in \{0, 1\}^8$ and the local phenotype is $\boxed{a}\boxed{b}\boxed{c}$ where $a, b, c \in \{0, 1\}$, then the genotype evolves locally according to the meta-rule $\alpha \times \beta \times \gamma$ (in the blending variant) while the phenotype evolves by applying the local rule β to the data "abc."

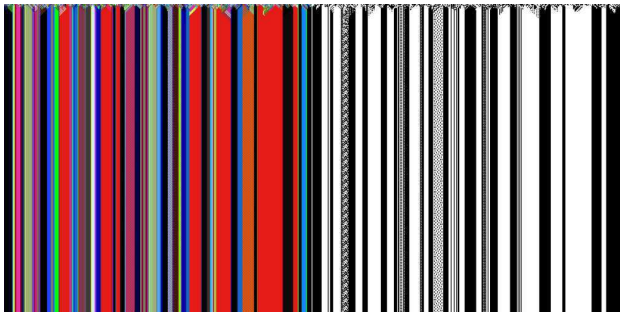


Figure 4. Phenotype with behaviour determined by genotype

In the phenotype layer, we see a few bands with interesting patterns, where the MetaCA at left has stabilised locally into one of the more interesting CA rules. However, at this scale we see that the long term evolution in the genotype layer is uninteresting: the structure observed in Figure 1 disappears quickly.

We therefore decided to introduce random mutations to the genotype, illustrated in Figures 5–7. With a high mutation rate, both genotype and phenotype are almost reduced to confetti. If we reduce the mutation rate sufficiently, some degree of stability is preserved, and the vertically striped bands are transformed into intermingling swaths of colour (Figure 6). We also see areas with more finely-grained structure in the phenotype layer.

In Figure 7, the colour-coded genotype layer has been replaced with a greyscale coding, and we see more clearly how the phenotype behaviour follows that of the genotype. That is, genotypes similar to Rule 0 (00000000) or Rule 256 (11111111) tend to produce 0 or 1, respectively, in the phenotype layer. Rules that output a blend of 0's and 1's are mapped to grey shades. Several interesting rules (Rule 110, Rule 30, Rule 90, Rule 184, and their reversals, bitwise inverses, and inverted-reversals) are highlighted in colour. In particular, Rule 110 variants are highlighted in red.

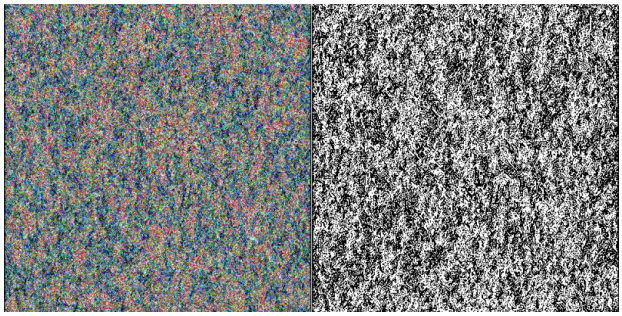


Figure 5. A high rate of mutation produces tantalising random structures

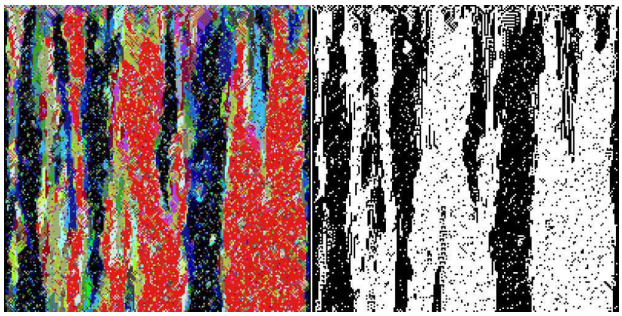


Figure 6. Throttling down the mutation rate preserves some of the large-scale stability while making room for variability

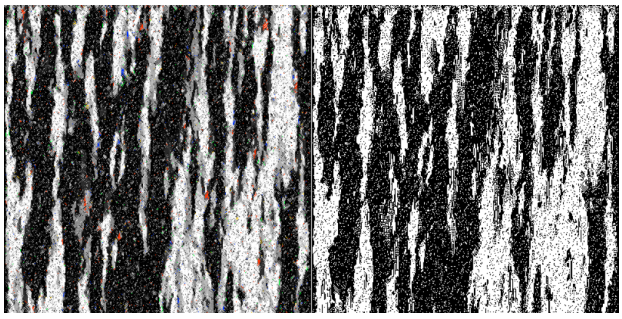


Figure 7. The search for intelligent life in the computational universe

We observe that Rule 0 and Rule 256 behaviour tends to predominate. Grey areas appear to be semi-stable. Red patches appear and disappear, as if independent planets evolve intelligent life and are then extinguished. With this physics, "intelligent life" seems inevitable, but also inevitably short-lived. One would have to look for another overall physics for intelligent behaviour to predominate.

A potential indication of the direction to look in is presented in Figure 8, which presents CAs generated by adjusting the typical blending evolution pattern by an (erroneously-programmed) mutation rule that only flips the first bit. We see that long-term behaviour in the genotype flutters randomly between Rule 0 (00000000) and Rule 128 (10000000). The short-term behaviour in the phenotype is nevertheless quite interesting, exhibiting many of the familiar lifelike edge-of-chaos patterns before ultimately succumbing to a version of Newton's First Law.

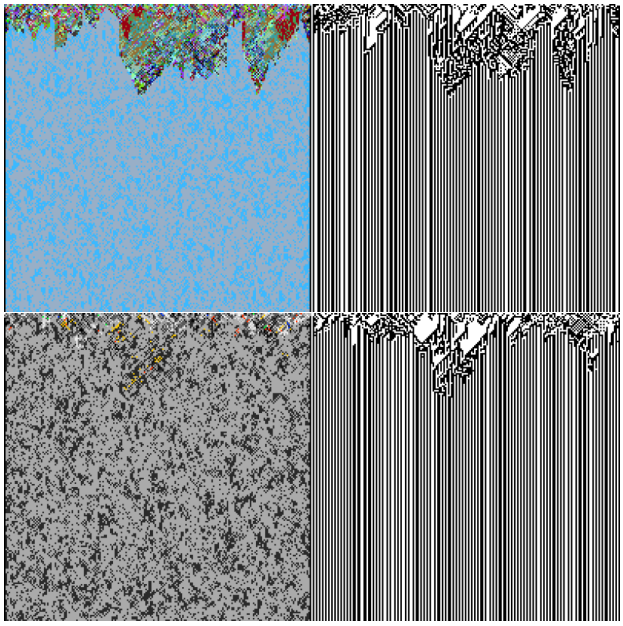
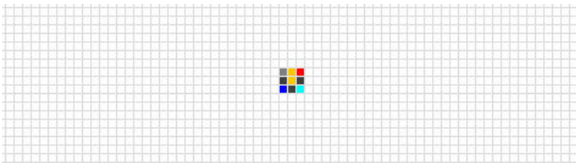


Figure 8. A skewed mutation pattern

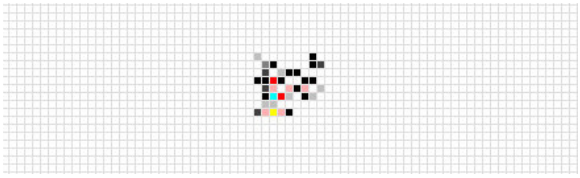
4.2 2D CAs

To see the behaviour of the union blend in action consider an initially populated grid, where colours represent the weights of alive cells:

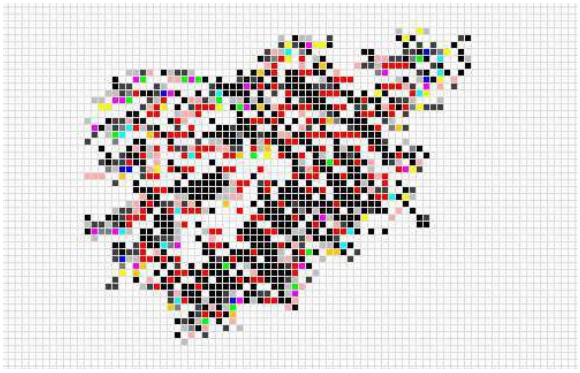


For this example, we initially restrict the computation of the blend for a particular cell to take place when the cell is alive in the next iteration. Also we compute the blend of genotype for all neighbours, whether dead or alive.

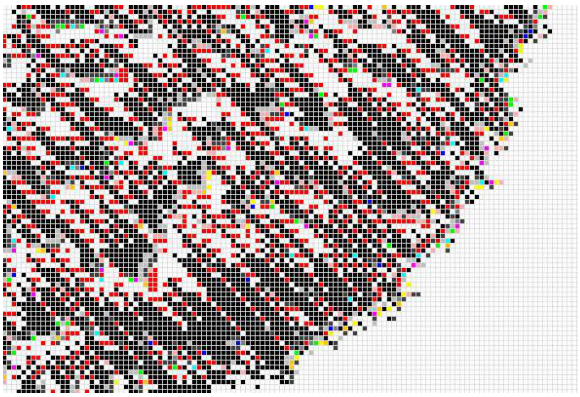
After 300 iterations the colony has grown a small amount:



Over time, the population continues to grow, with large patches of low-weight (black) cells:



Finally some structure starts to appear in the clustering:



The propagation that follows shows a population of cells which grows slowly over time. The majority of the members have low weight (represented by black squares), but interspersed within the population are chains of squares with high weight (represented by red squares) adjacent to dead cells (white).

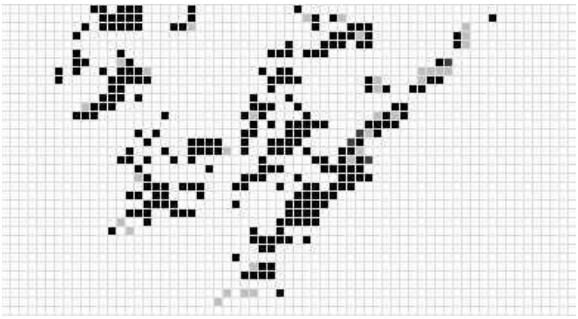
4.2.1 Modified Blends

So far we have only showed the union blend working on the genotype. However, it is possible to use different blending techniques:

- Consider blending only the genotypes of alive neighbours, or all neighbours;
- Consider only blending genotypes for cells which are alive after propagation;
- Consider an *intersection* blend, where the partition sizes for survival are minimised;
- Consider an *average* blend, where the values of each genotype (x_i, y_i, z_i) are summed and divided by either the number of alive neighbours, or the total number of neighbours.

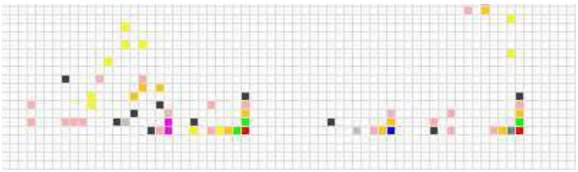
As an example of different observed emergent behaviour consider a union blend where the blend is only computed from alive neighbours, and as before we compute only for cells which are alive at the next iteration. We start with an initial state:

and observe a changing, but relatively steady pattern (resembling the motion of a flame) which does not grow in size using the union blend:

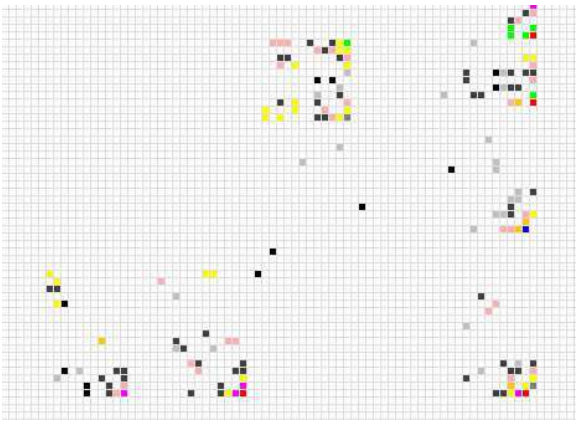


where the weight characteristic of the phenotype of each cell has fallen to very low.

Finally, consider applying instead an average blend under the same initial conditions:



Then we see a less steady but more active growth, with populations moving in triangular shapes away from population centres, leaving very small but steady and inactive populations behind:



The quickly-moving populations do not have a convergent weight characteristic in their phenotypes, as in the case with the union blend for the same initial conditions.

```
library metaca

logic CASL

spec METACABITENCODING =
    free type Bit ::= 0 | 1
    sort  Triple
    ops   t : Bit × Bit × Bit → Triple;
          bitop__ : Triple → Bit
end

spec METACABITCALC = % Calculate a blend given three 8-bit genotypes
    METACABITENCODING
then op   blend____ : Triple × Triple → Bit
    ∀ t1, t2, t3 : Triple
    • bitop t1 = bitop t2 ⇒ blend t1 t2 = bitop t1
    • ¬ bitop t1 = bitop t2 ⇒ blend t1 t2 = bitop t3
end

spec LEFTRULE =
    METACABITENCODING
then • bitop t(0, 0, 0) = 0
    • bitop t(0, 0, 1) = 1
    • bitop t(0, 1, 0) = 1
    • bitop t(0, 1, 1) = 0
    • bitop t(1, 0, 0) = 1
    • bitop t(1, 0, 1) = 1
    • bitop t(1, 1, 0) = 1
    • bitop t(1, 1, 1) = 0
end

spec RIGHTRULE =
    METACABITENCODING
then • bitop t(0, 0, 0) = 0
    • bitop t(0, 0, 1) = 1
    • bitop t(0, 1, 0) = 0
    • bitop t(0, 1, 1) = 1
    • bitop t(1, 0, 0) = 0
    • bitop t(1, 0, 1) = 1
    • bitop t(1, 1, 0) = 0
    • bitop t(1, 1, 1) = 1
end

spec LOCALRULE =
    METACABITENCODING
then • bitop t(0, 0, 0) = 0
    • bitop t(0, 0, 1) = 1
    • bitop t(0, 1, 0) = 0
    • bitop t(0, 1, 1) = 1
    • bitop t(1, 0, 0) = 0
    • bitop t(1, 0, 1) = 1
    • bitop t(1, 1, 0) = 0
    • bitop t(1, 1, 1) = 0
end

spec GENERIC = % Common between left and right
    METACABITENCODING
then • bitop t(0, 0, 0) = 0
    • bitop t(0, 0, 1) = 1
    • bitop t(1, 0, 1) = 1
end

view LEFT : GENERIC to LEFTRULE % Morphism from Generic to Left
end

view RIGHT : GENERIC to RIGHTRULE % Morphism from Generic to Right
end

spec BLEND = % This will be inconsistent
    combine Left, Right
end

spec WEAKENEDLEFTRULE =
    METACABITENCODING
then • bitop t(0, 0, 0) = 0
    • bitop t(0, 0, 1) = 1
    • bitop t(0, 1, 0) = 1
    • bitop t(0, 1, 1) = 0
    • bitop t(1, 0, 0) = 1
    • bitop t(1, 0, 1) = 1
    • bitop t(1, 1, 0) = 1
end

spec WEAKENEDRIGHTRULE =
    METACABITENCODING
then • bitop t(0, 0, 0) = 0
    • bitop t(0, 0, 1) = 1
    • bitop t(1, 0, 1) = 1
    • bitop t(1, 1, 1) = 1
end

view WEAKENEDLEFT : GENERIC to WEAKENEDLEFTRULE
end

view WEAKENEDRIGHT : GENERIC to WEAKENEDRIGHTRULE
end

spec CONSISTENTBLEND = % A consistent blend as new 8 bit encoding
    combine WeakenedLeft, WeakenedRight
and METACABITCALC
and LOCALRULE
end
```

Listing 1. CASL source code listing calculating the running example 01101110 × 01010100 × 01010101 via the blending meta-rule

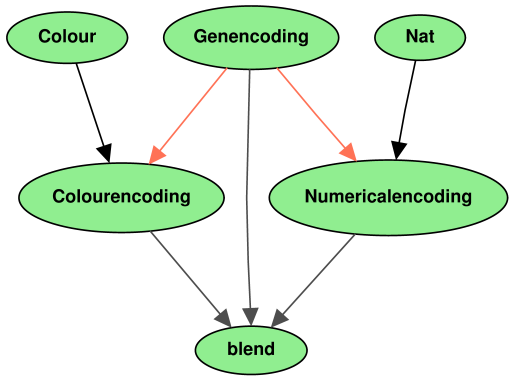


Figure 9. Blending different 2d genotypes

```
library metaca2d

logic CASL

spec NAT =
    sort  Nat
    op    max : Nat × Nat → Nat
    op    min : Nat × Nat → Nat
end

spec COLOUR =
    sort  Colour
    op    maxhue : Colour × Colour → Colour
end

% a 2−d cellular automaton with numerical Genotype
spec NUMERICALENCODING =
    NAT
then sort  NGenotype
    ops   genotype : Nat × Nat × Nat → NGenotype;
          t : Nat × Nat × Nat → NGenotype;
          numblend : NGenotype × NGenotype → NGenotype
    ∀ g1, g2 : NGenotype; x1, y1, z1, x2, y2, z2, x3, y3, z3 : Nat
    • g1 = t(x1, y1, z1) ∧ g2 = t(x2, y2, z2)
      ⇒ numblend(g1, g2)
        = t(min(x1, x2), min(y1, y2), max(z1, z2))
end

% A colour CA Genotype
spec COLOURENCODING =
    COLOUR
then sort  CGenotype = Colour
    op    hueblend : CGenotype × CGenotype → CGenotype
    ∀ g1, g2 : CGenotype
    • hueblend(g1, g2) = maxhue(g1 as Colour, g2 as Colour)
end

% A generic space
spec GENENCODING =
    sort  S
    sort  Genotype
    op    blend : Genotype × Genotype → Genotype
end

% A signature morphism from Generic to Numerical
view NUMERICALSM :
    GENENCODING to NUMERICALENCODING =
    S ↦ Nat, Genotype ↦ NGenotype, blend ↦ numblend
end

% A signature morphism from Generic to Colour
view COLOURSM :
    GENENCODING to COLOURENCODING =
    S ↦ Colour, Genotype ↦ CGenotype, blend ↦ hueblend
end

spec BLEND =
    combine NumericalSM, ColourSM
end
```

Listing 2. CASL source code using signature morphisms and pushout calculation to blend genotypes with different languages

5 Discussion

5.1 Research Contribution

The motivation for combining a notion of blending with cellular automata was to investigate ways in which cellular automata could be used to model processes, where propagation rules, or genotypes, were locally defined. The main research contributions in the field of two dimensional cellular automata are

- We built and implemented a framework where local propagation experiments can be performed;
- We used the HETS system to show that the notion of blending can be used to invent new propagation rules for different genotypes;
- We invented simply definable genotypes and blends of these genotypes to show proof of concept;
- Finally, we shared the results of simulations that illustrate qualitative behaviour in one and two dimensional MetaCAs.

The primary limitation of this work is that our results are purely observational at present. For example, the early experiments seemed to provide visual evidence that blending is useful: Figure 1 is more interesting than Figure 2. The robustness of our qualitative findings have been supported by developing a range of different experiments, for example, some analogy could be drawn between the "grey areas" observed in Figure 7 for the 1D case and the red-and-white chains that develop in the 2D case under union blending.

Our results confirm the basic finding of CA research: interesting global behaviour can arise from simple rules governing local interactions, with the added twist that these rules can also arise locally. The MetaCA setting seems to offer fertile ground for further computational research into evolutionary and co-evolutionary effects.

5.2 Social Interpretation

One can view the propagation of cells and patterns in a 1D or 2D MetaCA as a social process, and blending as a knowledge exchange. In the 2D case, we can think of the generated diagrams as illustrations of interactions between individuals with high knowledge, skill, or social impact (high weight), and those with less (low weight). The propagation in the "union" blend shows how large numbers of individuals with low social impact outnumber those with high social impact, but those with high social impact impose the emergent structure and determine the growth of the group of individuals.

In a fundamental respect our blending rules seem to embody a thought-provoking blend of two very different kinds of "ethics." Specifically, blending seems to introduce a dynamic similar to Carol Gilligan's *ethic of care* [8], which seeks to defend the relationships that obtain in a given situation. Here this is manifested by the question "Have my neighbours already formed a consensus?" This behaviour complements the local rule, which would correspond to Lawrence Kohlberg's *ethic of justice* (cf. [3]).

As we saw in Section 4.1, we would have to work harder to find meta-rules that give rise to an "intelligent universe" or in which life (considered as symbolic computation) plays an obvious negentropic role (*après* Bergson [4]).

One strategy that has not been developed here would be to make use of a "Baldwin effect" [2, 30], to use "learning" (considered as entropy) in the phenotype layer to drive (co)evolution. More specifically, $\boxed{0}\boxed{0}\boxed{0} \mapsto \boxed{0}$ and $\boxed{1}\boxed{1}\boxed{1} \mapsto \boxed{1}$ seem to be relatively uninteresting behaviours, but they are also hard to resist under the blending dynamics as we've defined them (compare Figures 4 and 7). Can we find ways to select against them?

5.3 Planned extensions

One observes that under our blending rule, the two non-entropic behaviours listed above are actually selected for, not against, because they are examples of the "neighbours match" condition. Indeed, reviewing the essential features of blending in the 1D case, we can use our basic principles:

> "*If neighbours match: use their shared value as the result.*
> *If neighbours don't match: use local logic to get the result.*"

to define a 1D CA rule, if we interpret "local logic" to mean "substitute my own value as the result." Here's how we would then define blending for triplets:

0	0	0	$\mapsto$	0	*Neighbours match*
0	0	1	$\mapsto$	0	*Local logic*
0	1	0	$\mapsto$	0	*Neighbours match*
0	1	1	$\mapsto$	1	*Local logic*
1	0	0	$\mapsto$	0	*Local logic*
1	0	1	$\mapsto$	1	*Neighbours match*
1	1	0	$\mapsto$	1	*Local logic*
1	1	1	$\mapsto$	1	*Neighbours match*

This is Wolfram's Rule 23: and as it happens, its evolutionary behaviour is not particularly interesting. Of course, for blending at the genotype level, "local logic" can be determined by any CA. Even so, when we use blending bitwise on alleles, we only ever run the local logic on half of the cases, and moreover it always the same half, determined by a "censored" version of Rule 23.

0	$*$	0	$\mapsto$	0	*Neighbours match*
0	$*$	1	$\mapsto$	$*$	*Local logic*
0	$*$	0	$\mapsto$	0	*Neighbours match*
0	$*$	1	$\mapsto$	$*$	*Local logic*
1	$*$	0	$\mapsto$	$*$	*Local logic*
1	$*$	1	$\mapsto$	1	*Neighbours match*
1	$*$	0	$\mapsto$	$*$	*Local logic*
1	$*$	1	$\mapsto$	1	*Neighbours match*

Rather than using Censored Rule 23 as our template, we could instead have the template determined by phenotype data, thereby involving the phenotype as a "hidden layer" in the computation.

The standard template could be understood to be generated by locking in $\boxed{0}\boxed{0}\boxed{0} \mapsto \boxed{0}$ along with a "variation"[4] $\boxed{0}\boxed{1}\boxed{0} \mapsto \boxed{0}$ and the bitwise inverses of these. A wider class of templates could be calculated from arbitrary phenotype data by the same operations. What we would lose in abandoning the intuition associated with local blending, we may be repaid through a much more abstract but richer procedural blend, operating at the level of genotype+phenotype coevolution. At the very least, we can point to a generic space, namely the locked-in local rule which would be carried over (along with its variants) from the phenotype to the corresponding alleles.

As a simple example of cross-domain blending consider a genotype defined as in §3.3, and another which is defined by comparing the hue of just one neighbour. Their blend is a richer theory combining elements from both genotypes. CASL code expressing these concepts is given in Listing 2, and the resulting categorical diagram can be seen in Figure 9. Experimentation with more sophisticated genotypes and blends is ongoing.

[4] $\boxed{0}\boxed{1}\boxed{0} = \boxed{0}\boxed{0}\boxed{0} + \boxed{\ }\boxed{1}\boxed{\ }$

5.4 Future work

Coevolution has been understood to be relevant from both a philosophical [15] and empirical perspective [29]. Finding patterns that allow us to exploit Baldwin effects to drive the co-evolution of genotype and phenotype in the direction of intelligent behaviour is an interesting computational project. The MetaCA domain may help to show how to systematise some aspects of the search for the principles and techniques that underlie broader computational intelligence.

Expanding on the semantically simple domain of CAs, we would like to use HETS to formalise the mechanisms of social knowledge sharing and problem solving in fields like mathematics. It may be possible to encode mathematical problems in a MetaCA or cellular program and involve a group of agents in finding solutions to these problems as a society, in an emergent manner. This would be informed by ongoing empirical analysis of real problem-solving activities [20] developed in parallel to the simulation work presented here.

6 Conclusion

This research was inspired by the aim to build an example of computational blending that matched, to some extent, the way blending might work in social settings. One person suggests an idea, and another offers a variant of that, a third brings in another idea from elsewhere and some combination is made. The next day, things head in another direction completely. Our progress in this research project has followed this sort of trajectory: from an initial critique of blending theory ("it's not dynamic enough to be social!") to some tentative examples showing how large-scale system dynamics can be driven by local behaviour in an emergent manner. Perhaps the most interesting aspect of this research is the relationship between these emergent dynamics and the meta-rules. Whereas previous CA research has shown that complex global behaviour can be generated from a set of simple, local rules, this project gives an enticing glimpse of a future research programme that carries out a computational search for those very rules (out of the many possible) that lead to system behaviour we would recognise as "intelligent."

7 Acknowledgements

This research has been funded by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open Grant number 611553 (COINVENT). We would like to thank Timothy Teravainen, Raymond Puzio, and Cameron Smith for helpful conversations and pointers to literature, and Christian Guckelsberger and an anonymous reviewer for comments that improved the draft.

REFERENCES

[1] Robert Axelrod, *The Evolution of Cooperation*, Basic Books, 1984.
[2] James Mark Baldwin, 'A New Factor in Evolution', *American Naturalist*, **30**, 441–451, 536–553, (1896).
[3] Seyla Benhabib, 'The Generalized and the Concrete Other: The Kohlberg-Gilligan Controversy and Feminist Theory', *PRAXIS International*, **4**, 402–424, (1985).
[4] Henri Bergson, *Creative evolution*, Henry Holt & Co., 1911 [1907].
[5] Matthew Cook, 'Universality in elementary cellular automata', *Complex Systems*, **15**(1), 1–40, (2004).
[6] Andreas Flache and Rainer Hegselmann, 'Do irregular grids make a difference? Relaxing the spatial regularity assumption in cellular models of social dynamics', *Journal of Artificial Societies and Social Simulation*, **4**(4), (2001).
[7] M. Gardner, 'The fantastic combinations of John Conway's new solitaire game "life"', *Scientific American*, **223**, 120–123, (October 1970).
[8] Carol Gilligan, *In a different voice*, Harvard University Press, 1982.
[9] Georg M. Goerg and Cosma Rohilla Shalizi, 'LICORS: Light cone reconstruction of states for non-parametric forecasting of spatio-temporal systems', *arXiv preprint arXiv:1206.2398*, (2012).
[10] Georg M. Goerg and Cosma Rohilla Shalizi, 'Mixed LICORS: A Nonparametric Algorithm for Predictive State Reconstruction', *arXiv preprint arXiv:1211.3760*, (2012).
[11] Joseph Goguen, 'Mathematical models of cognitive space and time', in *Reasoning and Cognition: Proc. of the Interdisciplinary Conference on Reasoning and Cognition*, eds., D. Andler, Y. Ogawa, M. Okada, and S.Watanabe, pp. 125–148, Tokyo, (2006). Keio University Press.
[12] Wim Hordijk, 'The EvCA project: A brief history', *Complexity*, **18**(5), 15–19, (2013).
[13] Wim Hordijk, Cosma Rohilla Shalizi, and James P Crutchfield, 'Upper bound on the products of particle interactions in cellular automata', *Physica D: Nonlinear Phenomena*, **154**(3), 240–258, (2001).
[14] Chris G Langton, 'Computation at the edge of chaos: phase transitions and emergent computation', *Physica D: Nonlinear Phenomena*, **42**(1), 12–37, (1990).
[15] George H. Mead, *The philosophy of the present*, Open Court, 1932.
[16] Melanie Mitchell, James P Crutchfield, and Peter T Hraber, 'Evolving cellular automata to perform computations: Mechanisms and impediments', *Physica D: Nonlinear Phenomena*, **75**(1), 361–391, (1994).
[17] Melanie Mitchell, Peter Hraber, and James P Crutchfield, 'Revisiting the edge of chaos: Evolving cellular automata to perform computations', *Complex Systems*, **7**, 89–130, (1993).
[18] Till Mossakowski, Christian Maeder, and Klaus Lüttich, 'The Heterogeneous Tool Set', in *TACAS 2007*, eds., Orna Grumberg and Michael Huth, volume 4424 of *Lecture Notes in Computer Science*, pp. 519–522. Springer-Verlag Heidelberg, (2007).
[19] Till Mossakowski, Christian Maeder, and Klaus Lüttich, 'The heterogeneous tool set, HETS', in *Tools and Algorithms for the Construction and Analysis of Systems*, eds., Orna Grumberg and Michael Huth, 519–522, Springer, (2007).
[20] Dave Murray-Rust, Joseph Corneli, Alison Pease, Ursula Martin, and Mark Snaith, 'Synchronised multi-perspective analysis of online mathematical argument', in *Proc. of 1st European Conference on Argumentation: Argumentation and Reasoned Action, 9-12 June 2015, Lisbon, Portugal*, eds., Sally Jackson, Dima Mohammed, Lilian Bermejo-Luque, and Steve Oswald, (2015). To appear.
[21] M.A. Nowak, 'Five rules for the evolution of cooperation', *Science*, **314**(5805), 1560–1563, (2006).
[22] M.A. Nowak and R.M. May, 'Evolutionary games and spatial chaos', *Nature*, **359**(6398), 826–829, (1992).
[23] Norman H. Packard, 'Adaptation toward the edge of chaos', in *Dynamic Patterns in Complex Systems*, eds., J.A.S. Kelso, A.J. Mandell, and M.F. Shlesinger, 293–301, World Scientific, (1988).
[24] Theodore P. Pavlic, Alyssa M. Adams, Paul C.W. Davies, and Sara Imari Walker, 'Self-referencing cellular automata: A model of the evolution of information control in biological systems', in *Artificial Life 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, 522–529, The MIT Press, (2014).
[25] Marco Schorlemmer, Alan Smaill, Kai-Uwe Kühnberger, Oliver Kutz, Simon Colton, Emilios Cambouropoulos, and Alison Pease, 'COINVENT: towards a computational concept invention theory', in *Proc. of the 5th International Conference on Computational Creativity*, eds., Dan Ventura, Simon Colton, Nada Lavrac, and Michael Cook, (2014).
[26] Moshe Sipper, *Evolution of parallel cellular machines*, Springer Heidelberg, 1997.
[27] Alexander J. Stewart and Joshua B. Plotkin, 'From extortion to generosity, evolution in the iterated prisoner's dilemma', *Proc. of the National Academy of Sciences*, **110**(38), 15348–15353, (2013).
[28] Alexander J. Stewart and Joshua B. Plotkin, 'Collapse of cooperation in evolving games', *Proc. of the National Academy of Sciences*, **111**(49), 17558–17563, (2014).
[29] Leigh Van Valen, 'A new evolutionary law', *Evolutionary theory*, **1**, 1–30, (1973).
[30] *Evolution and learning: The Baldwin effect reconsidered*, eds., Bruce H. Weber and David J. Depew, MIT Press, 2003.
[31] Stephen Wolfram, *Cellular automata and complexity: Collected papers*, Addison-Wesley Reading, 1994.