AISB Convention 2015, 20-22nd April, Canterbury

The Society for the Study of Artificial Intelligence and Simulation of Behaviour

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.

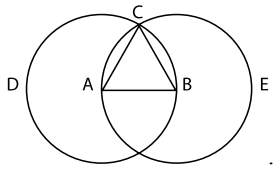
[13] It is worth noting in this context that our symbolic expression is arbitrary along two different dimensions. First, each of the letters 'a', 'b', 'c', and so on stand in for some natural number not further specified. The letter 'n' is different insofar as it is *also* arbitrarily large. My thanks 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 $imp(fv_1)$, …, $imp(fv_{Card(FlValues)})$ be the respective importance of the values $fv_1,\ldots, fv_{card(FlValues)}$ 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 $imp(fsg_1),\ldots, imp(fsg_{card(FrGoals)})$ be the respective importance of the values $fsg_1,\ldots, fsg_{card(FrGoals)}$ of FrGoals.

Let ObGoals be the values of the attribute Social Goals, that represent goals **obstructed** in the situation. Let $imp(osg_1),\ldots, imp(osg_{card(ObGoals)})$ be the respective importance of the values $osg_1, \ldots, osg_{card(ObGoals)}$ of ObGoals.

Let AcValues be the values of the attribute Values, that represent cultural values **accentuated** in the situation. Let $imp(av_1),\ldots, imp(av_{card(AcValues)})$ be the respective importance of the values $av_1,\ldots, av_{card(AcValues)}$ 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):

$$AIF = (AIFValues + AIFNorms + AIFArtif. + AIFRituals + AIFInstit. + AIFSymb. + AIFObGoals/70 \quad (1)$$

$$AIFValues = \sum_{i=1}^{card(FlValues)} imp(fv_i) \Big/ card(FlValues) \quad (2)$$

- ***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.

$$AIF = (AIFFrGoals + AIFAcValues)/20 \quad (3)$$

$$AIFFrGoals = \sum_{i=1}^{card(FrGoals)} imp(fsg_i) \Big/ card(FrGoals) \quad (4)$$

$$AIFAcValues = \sum_{i=1}^{card(AcValues)} imp(av_i) \Big/ card(AcValues) \quad (5)$$

In both cases (negative or positive emotion), the arousal of the emotion is then defined as follows (it varies between 0 and 1):

$$A = (AIF + 1) \times 0.5 \quad (6)$$

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 trade-off) 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.

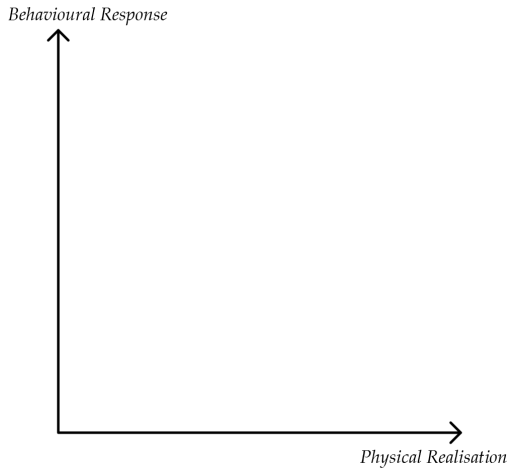


Figure 1. This graph represents a naïve visualisation of the idea of merging the behavioural response towards the environment and the physical realisation of the agent. Note that this visualisation is not meant to represent a mathematical function, but it is rather a supporting element for comprehending 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.

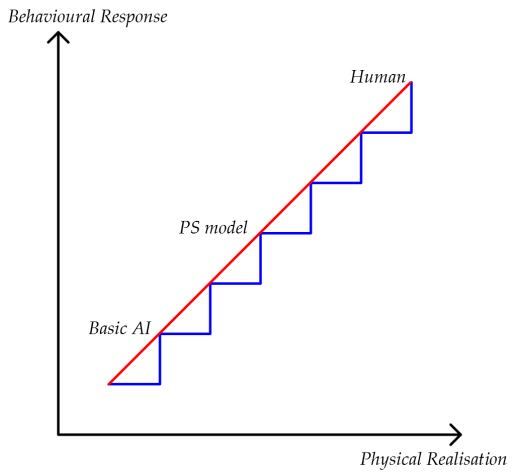


Figure 2. A representation of the comparison of non-projecting AI, PS model and human agent. Note that many patterns allow to connect those three distinct points, leaving open the question whether this should be a continuous or discrete "evolution".

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*")

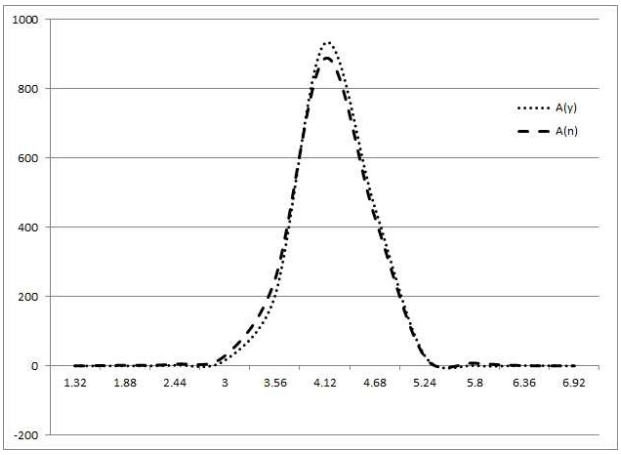


Figure 1. Openness to experience for A(y) (dotted) versus non-A (dashed)

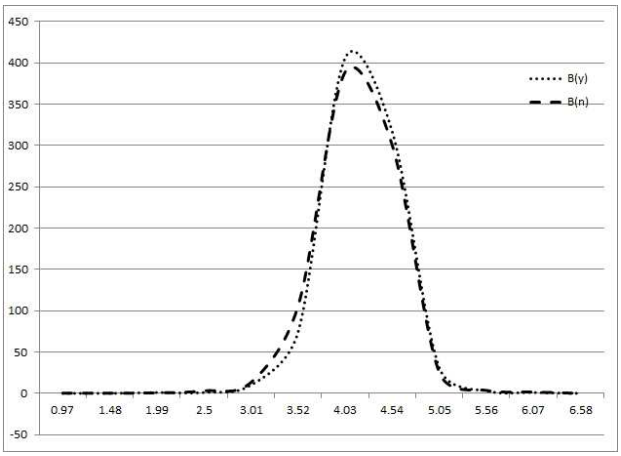


Figure 2. Openness to experience for B(y) (dotted) versus non-A (dashed)

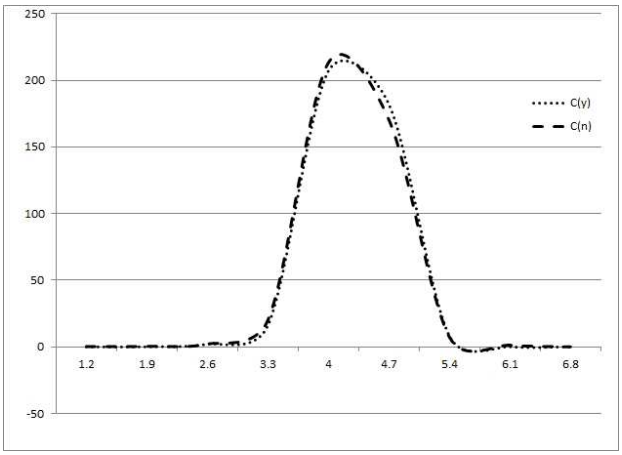


Figure 3. Openness to experience for C(y) (dotted) versus non-A (dashed)

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/

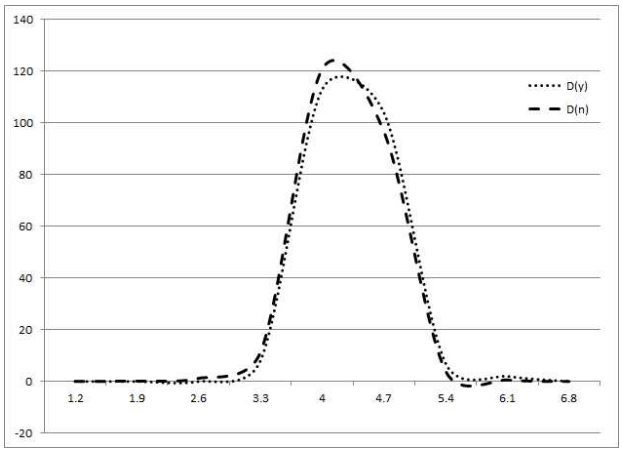


Figure 4. Openness to experience for D(y) (dotted) versus non-A (dashed)

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').

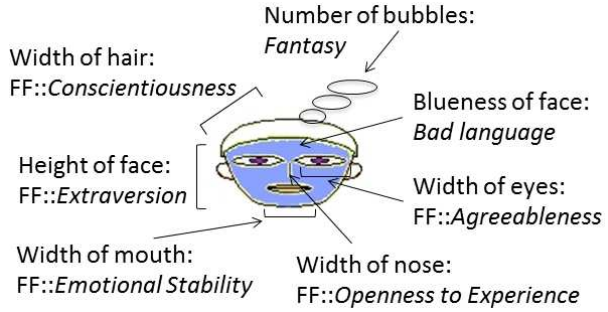


Figure 6. Traits and behaviours. Represented on a Chernoff face are the Five Factors (prepended by FF::) plus the additional behaviours for swearing level (darkness of blue colour on face) and fantasy level (amount of 'thought bubbles').

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.

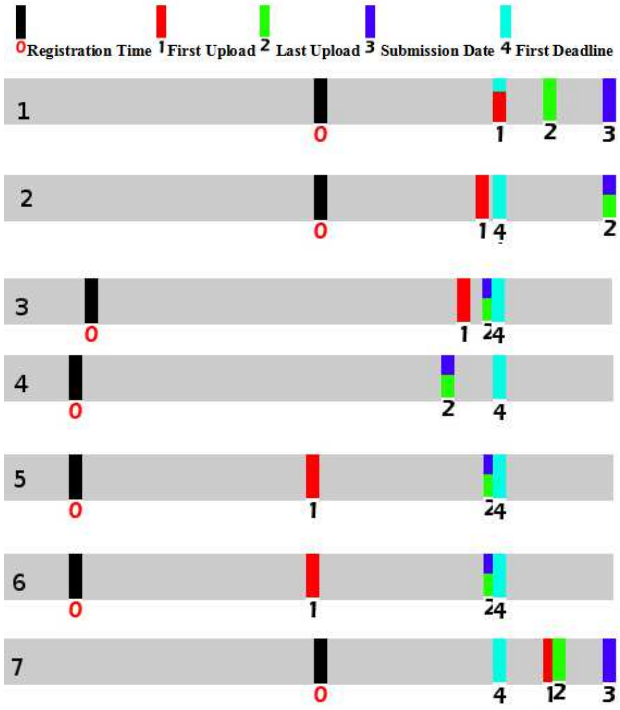


Figure 7. Seven user timelines. *T0* (black bar) is when the applicant first registered with the call. *T1* (red bar) represents when the applicant uploaded their first document, or First Action. *T2* (green bar) represents an applicants' Last Action. *T3* (blue bar) represents the applicants' Submission. *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.

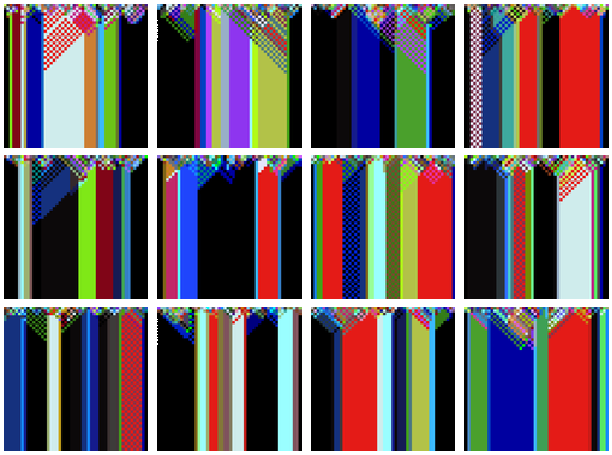


Figure 1. An illustration of MetaCA evolution

[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).

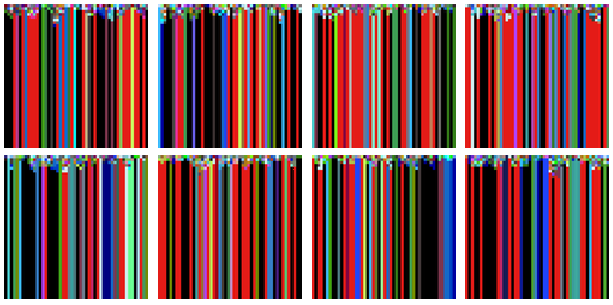


Figure 2. Under evolution according to the local rule without blending dynamics, a barcode-like stable pattern forms quickly

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].

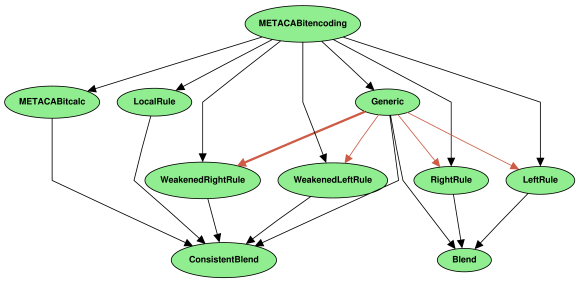
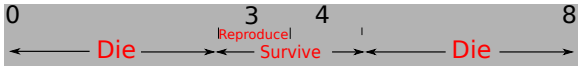


Figure 3. The development graph for calculating a blend of 8 bit encodings

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 $[0, 8]$.



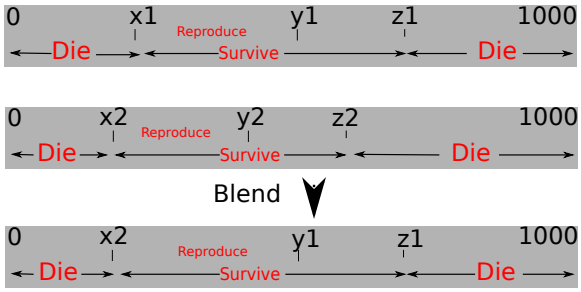
The number on the line corresponds to the number of alive neighbours adjacent, in cardinal and inter-cardinal directions, to a given square. If the square is dead then it becomes alive (labelled reproduce) if the number of alive neighbours is exactly three. If there are five or more alive neighbours 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."

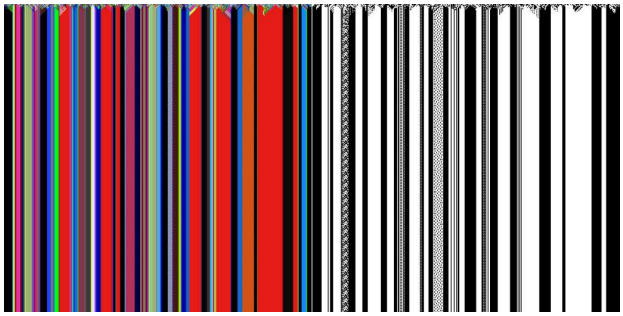


Figure 4. Phenotype with behaviour determined by genotype

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.

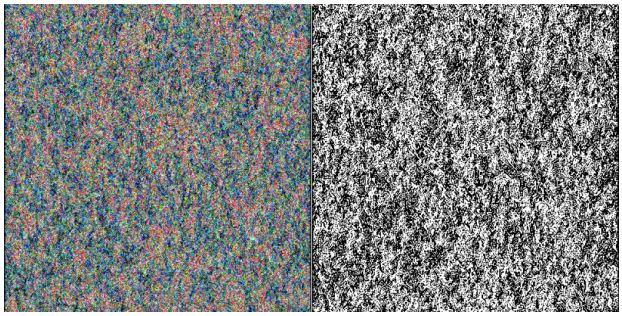


Figure 5. A high rate of mutation produces tantalising random structures

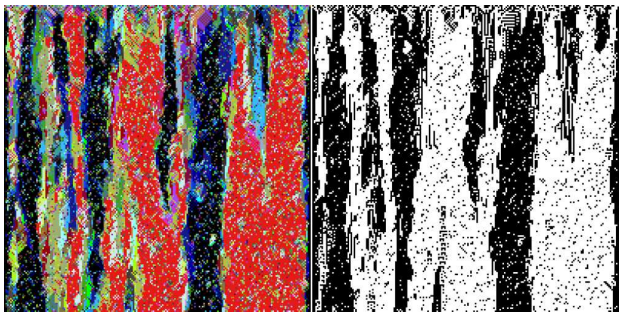


Figure 6. Throttling down the mutation rate preserves some of the large-scale stability while making room for variability

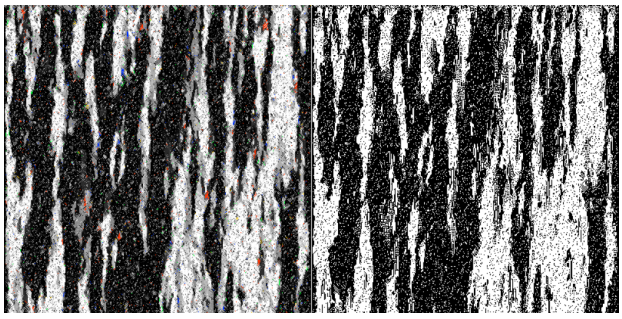


Figure 7. The search for intelligent life in the computational universe

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.

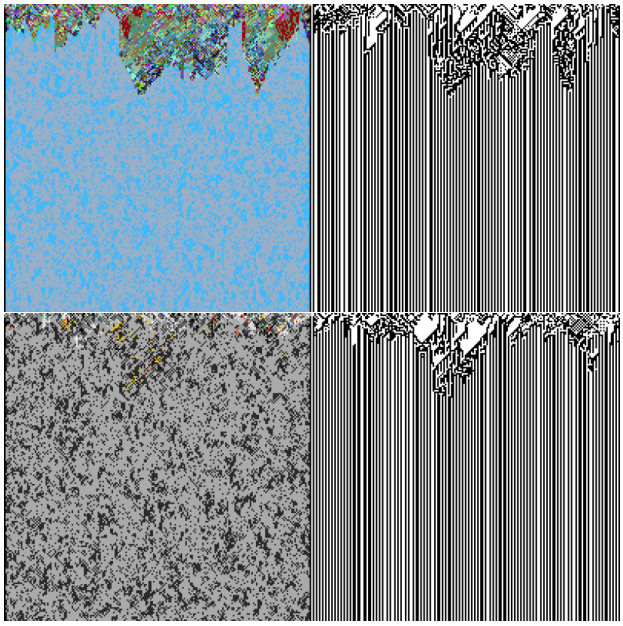
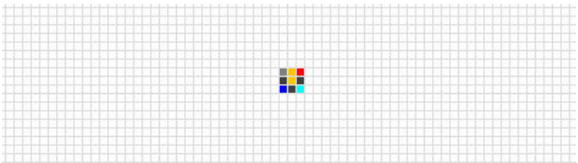


Figure 8. A skewed mutation pattern

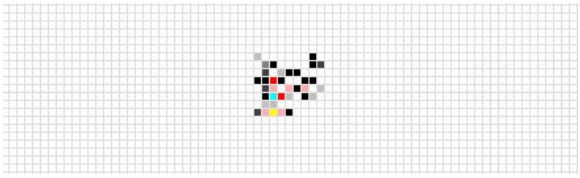
4.2 2D CAs

To see the behaviour of the union blend in action consider an initially populated grid, where colours represent the weights of alive cells:

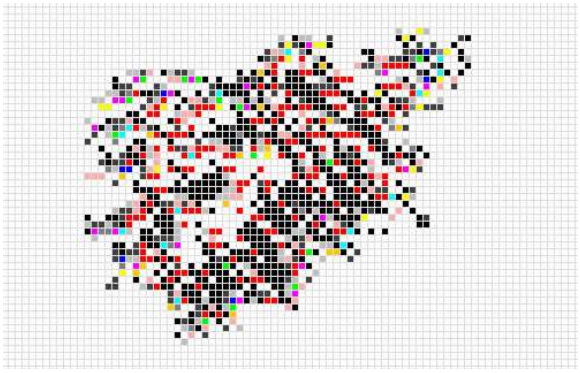


For this example, we initially restrict the computation of the blend for a particular cell to take place when the cell is alive in the next iteration. Also we compute the blend of genotype for all neighbours, whether dead or alive.

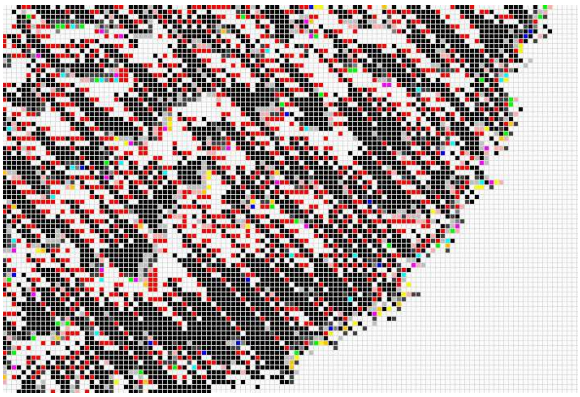
After 300 iterations the colony has grown a small amount:



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).

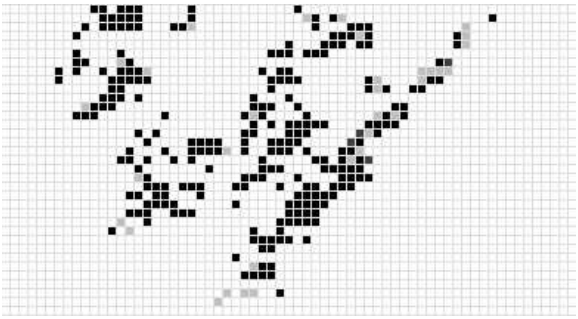
4.2.1 Modified Blends

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.

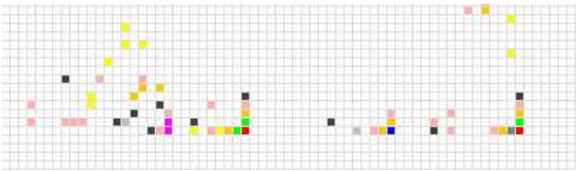
As an example of different observed emergent behaviour consider a union blend where the blend is only computed from alive neighbours, and as before we compute only for cells which are alive at the next iteration. We start with an initial state:

and observe a changing, but relatively steady pattern (resembling the motion of a flame) which does not grow in size using the union blend:

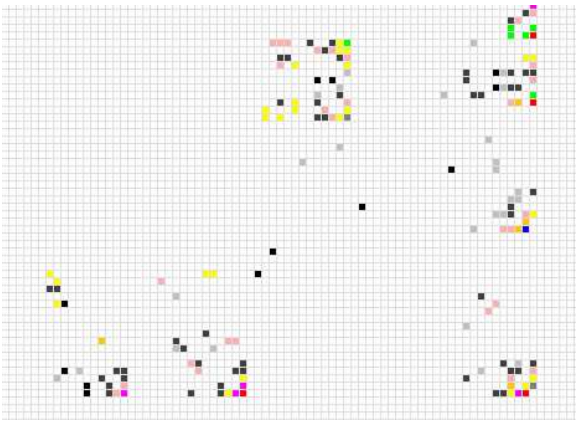


where the weight characteristic of the phenotype of each cell has fallen to very low.

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 $01101110 \times 01010100 \times 01010101$ via the blending meta-rule

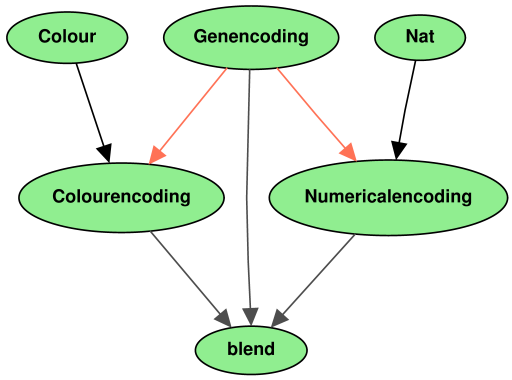


Figure 9. Blending different 2d genotypes

```
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.