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.

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]

[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.

[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.

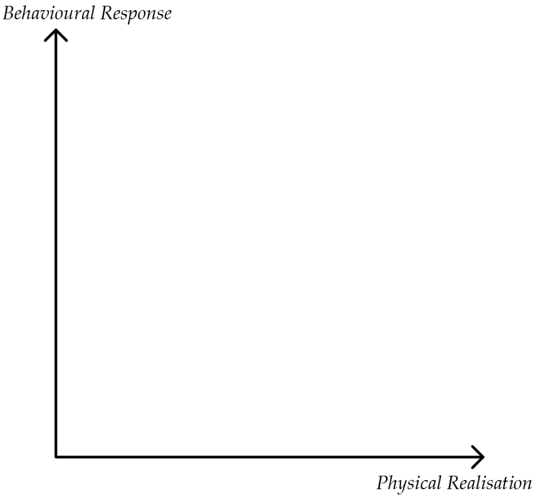

Figure 1. This graph represents a naïve visualisation of the idea of merging the behavioural response towards the environment and the physical realisation of the agent. Note that this visualisation is not meant to represent a mathematical function, but it is rather a supporting element for comprehending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

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-

[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?".

[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.

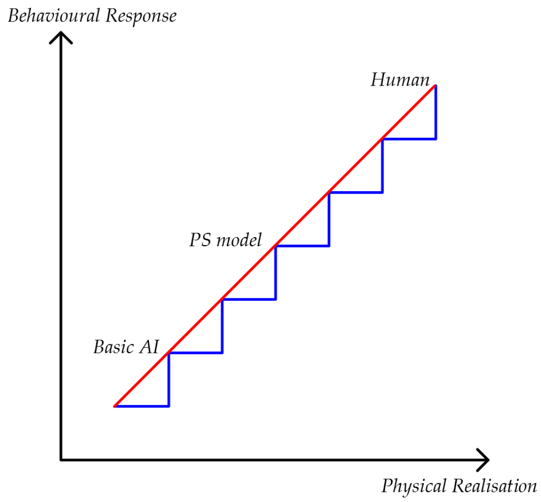

Figure 2. A representation of the comparison of non-projecting AI, PS model and human agent. Note that many patterns allow to connect those three distinct points, leaving open the question whether this should be a continuous or discrete "evolution".

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.