A Historical Perspective on Cognitive Science and Its Influence on XAI Research

Chapter · September 2019

DOI: 10.1007/978-3-030-30391-4_12

3 authors:

Marcus Westberg
Umeå University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Amber Zelvelder
Umeå University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Amro Najjar
University of Luxembourg
34 PUBLICATIONS **117** CITATIONS
SEE PROFILE

How Cognitive Science Impacts AI and What We Can Learn From It

Marcus Westberg[1], Amber Zelvelder[2], and Amro Najjar[2]

[1] Department of History, Philosophy and Religion, Oxford Brookes University, Oxford, UK
westberg.m@gmail.com
[2] Computer Science Department, Umea University, Sweden
amberz@cs.umu.se, najjar@cs.umu.se

Abstract. Cognitive science and artificial intelligence are interconnected in that developments in one field can affect the framework of reference for research in the other. Changes in our understanding of how the human mind works inadvertently changes how we go about creating artificial minds. Similarly, successes and failures in AI can inspire new directions to be taken in cognitive science. This article explores the history of the mind in cognitive science in the last 50 years, and draw comparisons as to how this has affected AI research, and how AI research in turn has affected shifts in cognitive science. In particular, we look at explainable AI (XAI) and suggest that folk psychology is of particular interest for that area of research. In cognitive science, folk psychology is divided between two theories: theory-theory and simulation theory. We argue that it is important for XAI to recognise and understand this debate, and that reducing reliance on theory-theory by incorporating more simulationist frameworks into XAI could help further the field. We propose that such incorporation would involve robots employing more embodied cognitive processes when communicating with humans, highlighting the importance of bodily action in communication and mindreading.

Keywords: XAI · Cognitive science · Folk psychology.

1 Introduction

Philosophy has had many influences on cognitive science, especially in regards to theories of mind. How we understand the mind affects how we seek to construct artificial intelligence. Jerry Fodor's computational theory of mind [22] has served as a platform for AI research as it states that the workings of the human mind are fundamentally algorithmic manipulation of symbols and thus perfectly possible to recreate in an artificial environment. Similarly, embodied approaches to cognition have had positive effects on robotics, as theories of the mind focused on world-agent interaction have inspired reactive bottom-up systems allowing robots to navigate the world through much less data-hungry decision making. So, what are the fundamental assumptions of the mind that

can be found in machine learning and explainable AI, and could philosophy or cognitive science help improve these assumptions? Since the field is concerned with human-robot interactions and understanding, there has to be one or several assumptions about how mindreading occurs and what role different types of interactions play in the cognitive processes leading to forming such understanding. There would also be underlying theories about how we rationalise and structure our internal models of the world (if at all) and how this comes into play when describing our intentions and actions. This in turn highlights the problem of the evidentiary boundary between mind and world, both robotic and human, and how we overcome it. Philosophy offers many different solutions to this problem, some elevate the role of action and behaviour in perception over internal representation in order to bypass the problem altogether, such as enactivism, while others remain strictly internalist but propose predictive models in order to unify world and agent through Bayesian best-performing hypotheses, something that can be found in predictive processing and predictive theories of mind. By looking at robots/AI, we do not only see assumptions about how robot minds should work, but also about how human minds interact with the world and other agents within it. In order for AI to be understandable in such a way that humans can cooperate with it, we need these theories to be compatible so that we can bypass the evidentiary boundary between what we see in the world (an agent's behaviour and outward communication) and what is going on with the agent's mental states (intentions, beliefs, possible future actions).

This article highlights the intertwined relationship between cognitive science and AI , and reviews how the history of philosophy of mind can be echoed in the history of developments in approaches to AI research (Section 2), then it identifies some examples of cognitive science-inspired AI application (Section 3), discusses folk psychology (Section 4) and presents a roadmap describing how XAI can profit from the recent evolution in theories of mind (Section 5). Finally, Section 6 concludes this article.

2 Tracing the history of Philosophy of Mind and Cognition

The history of AI research is often directly or indirectly linked to the history of cognitive science and theories of mind. Theories about how the human mind works can inform AI research in what to model their work on, and successes and failures within AI can teach us lessons about of the human mind itself is likely to work. If we were ever to create a general-level artificial intelligence that is perfectly indistinguishable from a human in cognitive capacities, we would likely want to say that we have also gained some insight into a plausible model of the human mind. Similarly, should theorized models of how the human mind works from a field such as philosophy fail as models for creating efficient and functional AI, then there is an equally compelling argument to be made that such results discredit said theorized models. An example of this crossover effect can be seen in the history of the Computational Theory of Mind (CTM). The Computa-

tional Theory of Mind was born out of the emergence of computing machines, most prominently the abstract Turing machine, invented by Alan Turing [60] as a model to, among other purposes, find a solution to the decision problem in formal logic. A line of thought came about that if machines can perform calculations and solve problems, could they be considered intelligent? Could a machine think? Turing himself asked these questions [61] and developed what would be known as the Turing test, a test where an interrogator would pass written questions to two other anonymous participants (one human, one machine) and would receive answers from both. The interrogator would then determine which participant is the machine, and which is the human. The purpose of the test was for the machine to exhibit their capacity of intelligent behaviour. If the machine could, even after rigorous questioning, appear to the interrogator as the human participant, then the machine had a claim for being an intelligent thinker.

This pursuit of intelligent machines had a reverse side to it: If these machines could solve problems and make decisions in a way that approaches cognition, could cognition itself in fact be computational? Such a thought offered a robust and structured approach to modelling the mind, while providing an analogous relationship with the emerging science of creating better and more complex computers. The rise of a computational theory of mind took place in this landscape, beginning in the earliest stages with a suggestion by McCulloch & Pitts [52] but was properly put into theory by Putnam [57] and was further developed by Fodor [22].

2.1 Fodor and cognitive science

The comparison between human intelligence and the workings of computers offered a solid argument for how the mechanisms of our mind and thoughts were structured, and provided a bridge between the abstract mind models of philosophers with the physical form which these cognitive faculties inhabit. If machines could think, then human thinkers could also be like machines. CTM served as a very influential founding theory in cognitive science [53]. While the theory started with Putnam [57], it was further developed and popularized by Fodor [22]. The theory is grounded in the idea that the mind works in many ways like a digital computer; the mind is parsing internal representations (symbols) in algorithmic ways, forming an internal computational language that is used to process input data into output. Fodor saw the symbol-dependent processing of the mind as a language and referred to this internal syntax as "the language of thought" and "mentalese." This interpretation placed mentalese as essentially a computational language of the mind, physically realised in the brain. By comparison, we could compare this to how a computer language works, and how the symbols and syntax of programming languages are realised in software (thinking), but physically situated in hardware (the brain). In fact, Fodor's theory claimed that this is exactly what is going on, suggesting that the mind is a physically realized computational environment where information processing occurs. That is, our minds are a formalized system parsing a language based on information-carrying representations, structured by syntactic and semantic rules, the upshot

of this idea being that the mind and the world are connected through our understanding of what is effectively a second simulated world within our mind. Information about the world enters the mind as sensory data; the things we see, hear, taste, smell and feel all enter our mind as raw data, which is then used to form mental representations, starting simple and building in complexity to form concepts. These representations are the components of our thought processes, which in turn are algorithmic in nature; our thought processes are problem-solving operations using an internal rule set which determines how the symbols (representations) are to be manipulated by the system. This representation in turn can vary in complexity and structure, such that they may be structurally atomic or molecular, which then carries down to their syntactic constituents [24].

CTM became a very popular theory of mind and gave new fuel to cybernetics (which had already been around since the 1930s) which led to the formation of modern cognitive science, as well as the resurgence of artificial intelligence research. As a result of this, throughout the 70s, 80s and 90s, much of the research in cognitive science and AI followed these computational mind models. In neuroscience as well, CTM was adopted by David Marr as a computational theory of vision [44][45]. However, the 90s were a decline in CTM's era as the leading theory of mind. One of the reasons was the rise in popularity of connectionism, and the people adopting it generally went against the idea of a language of thought. Connectionist models of the mind are built upon neural networks of interconnected nodes rather than the more linguistically inspired CTM. In particular, eliminative connectionism sought to move away from the idea of computationalism and mental representation in thought [15][38]. Thus the 90s was a transformative era where philosophers started to move away from the idea of classical CTM and arrived in a new paradigm era where the classical computational theory of mind as offered by Putnam and Fodor had decreased in popularity.

2.2 Embodied cognition and robotics

While the linguistic aspects of the classical version of CTM declined in the face of connectionism and research into neural network architectures, CTM's input-output model of mind also faced criticism for its inefficiency in how it modeled our interaction with the world. Classical CTM relied heavily on large amounts of data being collected in order for the mind to structure and learn about the world. This was also true of Marr's theory of vision although there were discussions in place that addressed the issue of dealing with uncertainty in identifying partial objects [45]. CTM had traditionally constructed our interaction of the world as being input-output, that is we perceive the world as it enters our senses (input), the sensory data transforming into a mental image of the world within our mind, which we then act upon (output). This has sometimes been referred to as the input-output sandwich, the idea being that cognition is filling, trapped between the bread of perception and action. Such a model, in order to connect with the outside world, first has to gather enough data via perception in order to create a working understanding of the world before being able to make decisions about

actions to be taken within that world. Such systems are information-hungry in ways that human-like animals often could not afford to be, which was considered a flaw in classical CTM's plausibility as a model for how our minds actually work. Similarly, AI that relied on intensive representation processing was still far away from achieving something akin to what a human mind could do, and often required quite complex top-down systems to guide their decision-making. Meanwhile, robotics was starting to see new successes by using motion and the world integrate with their cognitive systems [46][47] and roboticist Rodney Brooks presented a new type of computational architecture that was light on representation crunching and more focused on world-driven processes [10][9]. Philosopher Andy Clark [16] identified this new movement as the rise of Embodied Cognition, proposing a model of the mind where bodily action was more integrated with the classically introverted and isolated cognition presented in CTM. This theory of mind allowed for cognition to offload processes onto the world, and to use the environment for cognitive scaffolding. This paradigm shift made cognitive research more focused on mind-world-body interplay and interaction rather than complex internal architectures. For example, Collins et al. performed research into passive dynamic walker robots that use the natural pendulum motions of legs to aid walking, creating a smoother gait than the computation heavy and precision-demanding alternatives [19]. In philosophy, these developments led to both CTM and connectionism losing popularity and attention in favor of the now dominant EEEE theories: Embodied, Embedded, Extended and Enactive cognition.

2.3 Predictive processing against radical enactivism

As classical CTM loses influence, philosophers in the early 2000s and onwards become increasingly interested in describing the mind through world-driven processes and much interest is given to where we draw the border between mind and world. Classical CTM was traditionally very brain-focused in this regard, but the new EEEE theories all have in common that they attempt to expand this view, or reject it entirely. Embodied cognition [16] as mentioned before pays attention to how the body can play a causal role in cognition; embedded cognition involves the usage of external tools to facilitate cognitive processes; extended cognition is a functionalist stance that argues that external processes, if they fulfil the same functional role as the processes our heads, would constitute as part of our mind [18]; enactivism proposes the idea that experience of the world is conceived by interaction between brain, body and world, thus making cognition a dynamic activity rather than a passive intake of information [63]. The elements of these theories have branched off in many ways, and although the general emphasis within philosophy these days lies in 'active' cognition, i.e. focused on world-engaging processes rather than the more 'passive' traditionalist model, many positions have appeared that challenge each other.

The discovery of backward connections in the brain and an increased interest in Bayesian prediction models has given rise to theories of mind based on prediction error minimization in perception called predictive processing, where

the mind meets the world by predicting future input. Some use this theory to defend the traditionalist view of a brain-centric mind [37] while others seek to bridge the gap between embodied, enactive and representational models [17]. Either of these approaches to predictive processing still support the idea of mental representation, something that was core to classical CTM, the effective change being that predictive processing has taken care of the problem of inefficient input ecology. On the other hand, branches of enactivism such as radical enactivism argues against the existence of mental representations or content [39], claiming that enactive processes make contentful representations functionally obsolete. This creates a new conflict within philosophy of mind where our fundamental understanding of the mind has changed from a passive observer to an active one, while the debate about representations in cognition continues much like it did during the rise of eliminative connectionism.

2.4 Influences in XAI

As noted above, modeling an artificial intelligence can be influenced by how we as a scholastic community understand minds to work, be that as classical CTM, as predictive or embodied. eXplainable AI (XAI), however, is not only concerned with how minds work but also with mindreading and a cognitive agent's ability to explain itself to other minds. How this explanation is structured is related to what theory of mind we endorse. An AI modeled after classical CTM would be primarily concerned with reading and communicating its cognition in terms of representational mental states, while an AI incorporating embodied cognitive processes could make use of more world-driven processes in order to both read its audience and explain itself. For example, this could involve not only analyzing motion, facial expressions and other external signs of mental states, but in turn using these to both in performing cognitive processes and communicating said cognitive processes. Additionally, XAI is in the position where we are not only invested in how to structure the mind, but how we as agents understand other minds to work. In this way, XAI is influenced by and concerned with folk psychology, which has an extensive history within philosophy of mind and cognitive science. Concerns to be raised about folk psychology are thus also concerns relevant to XAI. In this paper, the term folk psychology is used specifically to refer to the cognitive capacity to a) attribute and explain mental states in other cognitive agents, b) predict future behaviour of other cognitive agents and c) to manipulate or coordinate behaviour with other cognitive agents through this attribution and prediction of mental states [36]. Broadly, possessing these capacities means that we describe the actions of others in intentional terms, highlighting the fact that we view other agents not as mere objects of causality, but as minds with their own beliefs and desires [58]. While these capacities are traditionally talked about in the context of relating to other human beings, there is a broader sense in which such capacities could be directed toward reading other functionally similar minds, for example reading and attributing mental states to non-human animals, or an AI.

3 Cognitive science inspiring AI Applications

In order to highlight the impact of developments in cognitive science of AI, this section presents examples of recent AI applications relying on enactivism (Section 3.1), and discusses how the renewed interest in XAI has relied on theories of mind (Section 3.2).

3.1 Enactive & Developmental AI

Enactive AI is inspired from the works on enaction developed by the cognitive biologists Maturana and Varela [48] [49]. In contrast to the cognitivism of the traditonal approaches, which involves a view of cognition that requires the representation of a given objective pre-determined world [62, 64], enaction relies on the assumption that existence of a cognitive agent are *enacted* (*i.e.* co-determined) by the agent as it interacts with its environment within which it is embedded. Thus, nothing is predetermined, and cognition becomes the process whereby an autonomous system becomes viable and effective in its environment [64]. Co-determination means that, on the one hand, the agent is specified by the environment and, on the other hand, it is the cognitive process itself which determines what in the environment is real and meaningful for the agent.

Since Enaction considers that the construction of cognition is undertaken on the basis of interactions between the agent and their physical and social environments, it supports constructivism, self-organization and developmental agents [20].

In contrast to the classical approach and the computational theory of mind, AI architectures based on enaction allow to overcome well known problems such as the *frame problem* [51], the *symbol grounding* problem [34], and *modeling of common-sense* problem [51].

Moreover, Enaction based AI is appealing because it allows Enactive-AI system to retain the following three characteristics [20]: *(i) no need for a-priori representations*: the agent does not need to have a pre-given model of its world. Instead, it can learn its environment when it interacts with it. *(ii) plasticity*: the agent is capable of adapting to the its environment even when significant distributions take place. This plasticity is located both at the physical (bodily) interactions (e.g. a robot capable of adapting its movement to an unforeseen slippery ground), and at the nerve level of higher interactions (*cerebral plasticity*). *(iii) co-evolution*: A modification of the world by the agent in return imposes a modification of that agent.

The characteristics mentioned above made enactive and developmental approaches to AI appealing for many AI applications, including developmental robotics [4], smart environment [54, 50], and road-traffic control [32].

3.2 The case for Explainable AI

Since the turn of the century, intelligent applications are becoming more and more pervasive in our daily lives. As these applications get more and more sophisticated, there is an impelling need to make them explainable. This tendency

is accentuated by the rise of black-box machine learning mechanisms [33] (e.g. deep learning) and their, sometimes, intriguing results [59]. To overcome these problems, recent legislation in the EU, emphasized the right of explanation [65]. Furthermore, evidence from user studies suggest that humans tend to anthropomorphize intelligent systems and attribute them with a State of Mind (SoM) [36]. This tendency is known as the *intentional stance* [21], and it pushes humans to explain the behavior of these systems in terms of beliefs, goals, and sometimes emotions [36]. For these reasons, recent research on XAI gained a significant new momentum [2]. XAI aims to offer explanations that would help the user to understand intelligent system and would lead to better human-agent collaboration and incite the user to understand the capabilities and the limits of the system, thereby improving the levels of trust and safety, and avoiding failures, since the lack of appropriate mental models and knowledge about the system may lead to failed interactions [12][3]. XAI is still in its early stages of development, for this reasons, most existing works are either carried out at the conceptual front [2]. Systems based on the Belief-Desire-Intention (BDI) architecture constitute a significant portions the few applied works [5]. The fact that the BDI agent architecture is inspired from Folk Psychology [56] makes it suitable to explain the agent intentions to the lay user [8]. Next section presents folk psychology and discusses its relationship with XAI.

4 Folk psychology and XAI

In philosophy and cognitive science there are two main theories for how mindreading (i.e. folk psychology) takes place: theory-theory and simulation theory. In Section 4.1 we will introduce theory-theory, followed by simulation theory in Section 4.2. The basic distinction to be made is that theory-theory holds the view that folk psychology operates from a set of rules that we hold in our mind, a distinct theory about mental states in others that informs our interpretation. Simulation theory on the other hand claims that our mindreading is performed by mental simulation, changing our perspective to that of others so that we may come to understand the underlying mental states for their behaviour.

4.1 Theory-theory and classical CTM

Theory-theory proposes that our understanding of other minds is built upon a tacit theory that we possess about how fears, desires and other mental states operate in other human beings, including the causal relationship these mental states have in a social context (i.e. how anger in one person can cause another to become sad, what it means to be sad and what kind of behaviours signify a mental state of sadness, how this sadness can be alleviated etc.). When we interact with or observe other cognitive agents, we employ this theory as a framework through which we form an understanding of the mental states in these agents based on what we can perceive in their behaviour. For example, person A sees person B reach out their hand toward an apple hanging on a branch. Person A

employs their theory that people who reach out their hands do so out of a desire to grab what their hand is reaching toward. Since an apple is an edible object, person A further projects that person B is likely to be hungry, since hunger drives desire for edible things. As such, person A has now mindread a state of hunger in person B, as well as their desire for an apple. Person A can then also predict future behaviour in person B, in that once they get the apple are likely to proceed with eating it in the near future.

There is significant overlap and compatibility between the theory-theory and classical CTM, especially when considering Fodor's theories on modularity of mind, where similarly tacit cognitive subsystems, i.e. modules, operate on a contextual basis for specific purposes - for example vision and language acquisition [23]. In a similar vein, theory-theory can be described as a folk psychology module in the human mind - a special capacity that is normally not part of our central processing, but within the right context receives the input (the perception or information of actions and behaviour of external agents), interprets the information through our folk psychology ruleset and produces an interpretation of what mental states we are perceiving. Older versions of theory-theory also tend toward a sentence-based representational structure closely resembling that of classical CTM, which is structured as a language of thought [22], but there are also those criticising this idea and proposing connectionist structure, while still maintaining the essence of a theory-based folk psychology [14][58].

Since theory-theory posits a set of rules or hypotheses informing the interpretations produced by our mindreading, this ruleset has to at some point be developed in the agent. This is an intersection where some supporters of theory-theory embrace nativism while others argue for empiricism. The nativist stance proposes that the framework for folk psychology is innately present at birth, and that its formation is part of a development pattern present in our genes [11]. Empiricists in contrast argue that the development of folk psychology is a process based on evidence-based theory formation throughout childhood and onward [30][29]. This debate can be compared to the debate of nativism versus empiricism in language acquisition, with similar arguments to be made, such as nativism's critique of the poverty of stimulus [13].

4.2 Simulation theory

Simulation theory contrasts itself with theory-theory in claiming that our process of mindreading is largely based on our ability to put ourselves in the shoes of others, in essence we simulate the behaviour of others in our own mind as if from our own perspective. There are two components to this that need to be unpacked: 1) how mental states in other persons are simulated in our own mind, 2) how these simulated mental states serve mindreading and prediction.

When it comes to (1), some supporters of simulation theory have latched onto the discovery of mirror neurons as signifying a mirroring capacity that could be linked to mental state simulation [25]. However, the modern version of simulation theory came about before this discovery, in the mid-1980s [31]. Thus for the first decade of its popularity, and still ongoing for those who do not subscribe to

the idea of mirroring, this capacity is instead described in terms of empathy or imagination.

Regarding (2), simulating mental states is not enough to predict behaviour, as there needs to be a process in place to explain how these mental states relate to possible future behaviour. What follows is thus a simulated decision-making process, where the agent asks themselves what they would do given the simulated premises. Essentially, the agent asks themselves *"what would I do if X?"* where X is the simulated mental states and context of the subject of the mindreading. The answer to that question becomes the theory used to predict the subjects future actions.

Simulation theory is thus frugal in the sense that it does not require an information-rich theory of how other minds work. Instead, all that is required is the capacity for the agent to place their frame of mind in someone elses situation through empathic simulation, and what follows is no different from the kinds of decision-making processes that take place in the agents own actions. This simulated process greatly reduces the amount of cognitive capital spent on mindreading in comparison to theory-theory, since mindreading becomes primarily process-driven rather than theory-driven [26].

5 A roadmap for XAI

As mentioned before in Section 2.4, folk psychology is relevant to XAI research. As XAI involves both humans explaining themselves to AI and, more importantly, AI explaining itself to humans, mindreading becomes an important aspect of this exchange of information and understanding. Understanding robots as subjects of mindreading helps creating a framework where the mental states of an AI can be explained, and thus aids the strategies employed by the AI in explaining itself to humans. These influences of folk psychology can already be found in XAI. For example, belief-desire-intention (BDI)-based agents [7][40] are an application of theory-theory in XAI: These agents have their actions explained through beliefs and goals, generating a theory of future behaviour based on an understanding of these beliefs and goals, and how they relate to actions taken. BDI agents are widely used for social simulations [1]. BDI agents offer a reasoning formalization inspired by human mentality based on intuitive concepts that allow for a straightforward implementation in IT systems. Hence, the BDI architecture has been highlighted as a practical solution to model humans and create human-like behavior in simulated environments [56].

However, the way these are communicated to humans, and thus the method through which a human agent is invited to mindread the BDI-based agent, is through either natural language explanation, or an understanding of the BDI-based agent's goal hierarchy tree (GHT). This kind of explanation involves no amount of simulation or putting oneself in the robot's shoes, but rather becomes a task of piecing together the robot's reasoning through empirical questioning and investigation, thus making it a clear case of theory-theory.

While this may be functionally sufficient in allowing the robot to explain itself within the context of its design, it is limiting in the sense that underutilises the human capacity for empathetic mentalisation, specifically when it comes to mindreading through interpreting action - something that both human children and chimpanzees have been shown capable of [35]. This appeal to empathetic mentalisation is a trait of simulation theory. Applying simulation theory as a framework for XAI could thus upon up a broader scope for communication by capturing more of the human experience.

Simulation theory, as presented in Section 4.2, possesses a greater advantage over theory-theory in that it can bypass the nativist versus empiricist debate present in theory-theory. The ability to learn mindreading from simulation provides an innate system for learning where our knowledge and understanding of other minds does not have to come pre-constructed at birth, nor does the learning itself involve the construction of a rigorous ruleset. Instead, any constructed rulesets would be created through best-making hypotheses much in the manner of a prediction-error minimization model for perception [37], where the more advanced facets of our knowledge of human behaviour are constructed models from a much more basic and frugal process.

However, it is important to note that theory-theory is improving by making adjustments to incorporate Bayesian prediction models, just like predictive processing revitalised the representational aspects of CTM in the face of increasingly non-representational alternatives [17]. In theory-theory's case, this turn toward prediction is pertaining only to the empiricist part of the divide, and thus not a nativist strategy [28]. This, however, may not be enough for standalone theory-theory to win out over simulation theory.

Instead, what is becoming increasingly popular are hybrid theories, incorporating simulationist elements, but falling back on theory-theory wherever simulation alone is insufficient [27] or vice versa [55]. Such hybrid theories would still go well together with predictive processing: Theories that are primarily simulationist with elements of theory-theory would benefit from prediction error minimization's explanation of how theory can be constructed from continuous and updating simulation, while primarily theory-based theories with elements of simulation will find an increase in cognitive frugality that explains away the poverty of stimulus argument while avoiding nativism, thus promoting bottom-up over top-down learning. Thus even if standalone simulation theory does not stand as a clear superior theory, there is still a strong argument to be made that including elements of simulation theory presents a stronger theory for how we mindread over a standalone theory-theory. If this holds true, then it is even more important that XAI going onward incorporates more elements of understanding the AI as an embodied agent.

Some XAI research already shows an adaptation and understanding of the importance of a robot's physical movements and how nuances in this affect how a human observer interprets its mental states [36]. Experiments involving robots communicating their intentions with bodily movements [43] or their mental state through eye movements and posture [6] show positive results in human inter-

action with robots and opens up a new path of interpretation. Furthermore, the importance of emotions in explanations has been studied by recent works in XAI. Yet, most of these works (e.g. [41] [42]) rely on BDI agents (theory-theory) and address simplified scenarios. This research shows an opening for simulation theory in XAI since simulation theory helps promote emphatic behavior (by putting ourselves in the others' shoes), and allows the agents to explain its behavior using body cues and non-verbal communications.

6 Conclusion

In this article we have shown how intertwined the relationship between cognitive science and AI is, and have reviewed how the history of philosophy of mind can be echoed in the history of developments in approaches to AI research. For these reasons, it is important for XAI to pay attention to developments in cognitive science, particularly regarding folk psychology, as these developments could inform better approaches for XAI in the future. Conversely, the successes and failures of XAI could in turn influence how the strategies employed are viewed in discussions on folk psychology.

There is a clear shift in cognitive science and philosophy of mind toward world-driven and embodied explanation of cognition, be that in regards to perception, folk psychology, decision-making or action, and this shift does not seem to be going back. Thus, the question is no longer about the input-output sandwich and how isolated thinking is from the world. Rather, the new focus is on the relationship between thinking and world. With this in mind, our prediction is that human-robot interaction will increasingly involve embodied cognition as a tool for communication. Embodied processes in robots not only act as cognitive scaffolding in navigating and interacting with the world, but it also opens them up for mindreading as they present their mental states through behaviour that humans recognise and can understand, thus allowing humans to tacitly recognise robots as cognitive agents with beliefs and desires through the same methods as they would other humans and non-human animals.

References

1. Adam, C., Gaudou, B.: Bdi agents in social simulations: a survey. The Knowledge Engineering Review **31**(3), 207–238 (2016)
2. Anjomshoae, S., Najjar, A., Calvaresi, D., Framling, K.: Explainable agents and robots: Results from a systematic literature review. In: 18th International Conference on Autonomous Agents and Multiagent Systems (2019), to appear
3. Bethel, C.L.: Robots Without Faces: Non-verbal Social Human-robot Interaction. Ph.D. thesis, Tampa, FL, USA (2009), aAI3420462
4. Blank, D., Kumar, D., Meeden, L., Marshall, J.B.: Bringing up robot: Fundamental mechanisms for creating a self-motivated, self-organizing architecture. Cybernetics and Systems: An International Journal **36**(2), 125–150 (2005)
5. Bratman, M.: Intention, Plans, and Practical Reason, vol. 10. Harvard University Press Cambridge, MA (1987)

6. Breazeal, C., Fitzpatrick, P.: That certain look: Social amplification of animate vision. AAAI Technical Report (11 2001)
7. Broekens, J., Harbers, M., Hindriks, K., Bosch, K., Jonker, C., Meyer, J.j.: Do you get it? user-evaluated explainable bdi agents. vol. 6251, pp. 28–39 (09 2010)
8. Broekens, J., Harbers, M., Hindriks, K., Van Den Bosch, K., Jonker, C., Meyer, J.J.: Do you get it? user-evaluated explainable bdi agents. In: German Conference on Multiagent System Technologies. pp. 28–39. Springer (2010)
9. Brooks, R.A.: Intelligence without reason. In: Proceedings of the Twelveth Intarnationl Joint Conference on Artificial Intelligence. pp. 569–595 (1991)
10. Brooks, R.A.: Intelligence without representation. Artificial Intelligence **47**, 139–159 (1991)
11. Carruthers, P.: Simulation and self-knowledge: A defence of the theory-theory. In: Carruthers, P., Smith, P.K. (eds.) Theories of Theories of Mind, pp. 22–38. Cambridge University Press (1996)
12. Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., Parikh, D.: It takes two to tango: Towards theory of ai's mind. arXiv preprint arXiv:1704.00717 (2017)
13. Chomsky, N.: A review of b. f. skinner's verbal behavior. Language **35**(1), 26–58 (1959)
14. Churchland, P.M.: Folk psychology and the explanation of human behavior. Philosophical Perspectives **3**(n/a), 225–241 (1989)
15. Churchland, P.M.: A Neurocomputational Perspective: The Nature of Mind and the Structure of Science. MIT Press (1989)
16. Clark, A.: Being There: Putting Brain, Body, and World Together Again. MIT Press (1997)
17. Clark, A.: Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press (2016)
18. Clark, A., Chalmers, D.J.: The extended mind. Analysis **58**(1), 7–19 (1998)
19. Collins, S.H., Wisse, M., Ruina, A.: A three-dimensional passive-dynamic walking robot with two legs and knees. I. J. Robotics Res. **20**(7), 607–615 (2001)
20. De Loor, P., Manach, K., Tisseau, J.: Enaction-based artificial intelligence: Toward co-evolution with humans in the loop. Minds and Machines **19**(3), 319–343 (2009)
21. Dennett, D.C.: The intentional stance. MIT press (1989)
22. Fodor, J.A.: The Language of Thought. Harvard University Press (1975)
23. Fodor, J.A.: The Modularity of Mind. Cambridge, MA: MIT Press (1983)
24. Fodor, J.A., Pylyshyn, Z.W.: Connectionism and cognitive architecture. Cognition **28**(1-2), 3–71 (1988)
25. Gallese, V., Goldman, A.: Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences **2**(12), 493–501 (1998)
26. Goldman, A.: Interpretation psychologized. Mind and Language **4**(3), 161–85 (1989)
27. Goldman, A.I.: Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford University Press USA (2006)
28. Gopnik, A.: The theory theory 2.0: Probabilistic models and cognitive development. Child Development Perspectives **5**(3), 161–163 (2011)
29. Gopnik, A., Meltzoff, A., Kuhl, P.: The scientist in the crib: Minds, brains and how children learn. Journal of Nervous and Mental Disease - J NERV MENT DIS **189** (03 2001)
30. Gopnik, A., Wellman, H.M.: The theory theory. In: Hirschfeld, L.A., Gelman, S.A. (eds.) Mapping the Mind: Domain Specificity in Cognition and Culture. p. 257293. Cambridge University Press (1994)

31. Gordon, R.M.: Folk psychology as simulation. Mind and Language **1**(2), 158–71 (1986)
32. Guériau, M., Armetta, F., Hassas, S., Billot, R., El Faouzi, N.E.: A constructivist approach for a self-adaptive decision-making system: application to road traffic control. In: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 670–677. IEEE (2016)
33. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) **51**(5), 93 (2018)
34. Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena **42**(1-3), 335–346 (1990)
35. Harris, P.F.: From simulation to folk psychology: The case for development. Mind and Language **7**(1-2), 120–144 (1992)
36. Hellström, T., Bensch, S.: Understandable robots : What, why, and how. Paladyn - Journal of Behavioral Robotics **9**(1), 110–123 (2018)
37. Hohwy, J.: The Predictive Mind. Oxford University Press (2013)
38. Horgan, T.E., Tienson, J.L.: Connectionism and the Philosophy of Psychology. MIT Press (1996)
39. Hutto, D.: Enactivism: Why be radical? In: Sehen und Handeln, pp. 21–44. De Gruyter Akademie Forschung (01 2011)
40. Kaptein, F., Broekens, J., Hindriks, K., Neerincx, M.: Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In: 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp. 676–682 (Aug 2017)
41. Kaptein, F., Broekens, J., Hindriks, K., Neerincx, M.: The role of emotion in self-explanations by cognitive agents. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). pp. 88–93. IEEE (2017)
42. Kaptein, F., Broekens, J., Hindriks, K., Neerincx, M.: Self-explanations of a cognitive agent by citing goals and emotions. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). pp. 81–82. IEEE (2017)
43. Kobayashi, K., Yamada, S.: Motion overlap for a mobile robot to express its mind. JACIII **11**, 964–971 (01 2007)
44. Marr, D.: Visual information processing: the structure and creation of visual representations. Philosophical Transactions of the Royal Society of London. B, Biological Sciences **290**(1038), 199–218 (1980)
45. Marr, D.: Vision : a computational investigation into the human representation and processing of visual information / David Marr. W. H. Freeman (1982)
46. Mataric, M.J.: Navigating with a rat brain: A neurobiologically-inspired model for robot spatial representation. In: Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats. pp. 169–175. MIT Press, Cambridge, MA, USA (1990)
47. Mataric, M.J.: Integration of representation into goal-driven behavior-based robots. IEEE Transactions on Robotics and Automation **8**(3), 304–312 (June 1992)
48. Maturana, H., Varela, F.: Autopoiesis and cognition: The realization of the living (boston studies in the philosophy of science) (1991)
49. Maturana, H.R.: The organization of the living: A theory of the living organization. International journal of man-machine studies **7**(3), 313–332 (1975)

50. Mazac, S., Armetta, F., Hassas, S.: On bootstrapping sensori-motor patterns for a constructivist learning system in continuous environments. In: Artificial Life Conference Proceedings 14. pp. 160–167. MIT Press (2014)
51. McCarthy, J.: Programs with common sense. RLE and MIT computation center (1960)
52. Mcculloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Journal of Symbolic Logic **9**(2), 49–50 (1943)
53. Miller, G.A., G.E.P.K.: Plans and the structure of behavior. Holt, New York (1967)
54. Najjar, A., Reignier, P.: Constructivist ambient intelligent agent for smart environments. In: 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). pp. 356–359. IEEE (2013)
55. Nichols, S., Stich, S.P.: Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds. Oxford University Press (2003)
56. Norling, E.: Folk psychology for human modelling: Extending the bdi paradigm. In: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1. pp. 202–209. IEEE Computer Society (2004)
57. Putnam, H.: Brains and behavior. In: Butler, R.J. (ed.) Analytical Philosophy: Second Series. Blackwell (1963)
58. Stich, S.P., Nichols, S.: Folk psychology: Simulation or tacit theory? Mind and Language **7**(1-2), 35–71 (1992)
59. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
60. Turing, A.: On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society **2**(42), 230–265 (1936)
61. Turing, A.: Computing machinery and intelligence. Mind **59**, 433–460 (1950)
62. Van Gelder, T., Port, R.F.: Its about time: An overview of the dynamical approach to cognition. Mind as motion: Explorations in the dynamics of cognition **1**, 43 (1995)
63. Varela, F., Thompson, E., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience. MIT Press (1991)
64. Vernon, D., Furlong, D.: Philosophical foundations of ai. In: 50 years of artificial intelligence, pp. 53–62. Springer (2007)
65. Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing (2017)