Mind the gap: challenges of deep learning approaches to Theory of Mind

Jaan Aru · Aqeel Labash · Oriol Corcoll · Raul Vicente

Published online: 27 January 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract
Theory of Mind (ToM) is an essential ability of humans to infer the mental states of others. Here we provide a coherent summary of the potential, current progress, and problems of deep learning (DL) approaches to ToM. We highlight that many current findings can be explained through shortcuts. These shortcuts arise because the tasks used to investigate ToM in deep learning systems have been too narrow. Thus, we encourage researchers to investigate ToM in complex open-ended environments. Furthermore, to inspire future DL systems we provide a concise overview of prior work done in humans. We further argue that when studying ToM with DL, the research’s main focus and contribution ought to be opening up the network’s representations. We recommend researchers to use tools from the field of interpretability of AI to study the relationship between different network components and aspects of ToM.

Keywords Theory of Mind · Artificial intelligence · Reinforcement learning · Deep learning

1 Introduction
Rapid advances in deep learning (DL) have led to human-level performance on certain visual recognition and natural language processing tasks. Moreover, research has revealed shared computational principles in humans and DL models for vision (Güçlü and van Gerven 2015; Seibert et al. 2016; Cichy et al. 2016; Yamins and DiCarlo 2016; Kriegeskorte 2015) and language processing (Schrimpf et al. 2021; Caucheteux and King 2022; Goldstein et al. 2022). In particular, there is a hierarchical correspondence between the neural
activity at different levels of processing and the different layers of DL systems (Güçlü and van Gerven 2015; Seibert et al. 2016; Cichy et al. 2016). This work has illuminated how relatively simple transformations applied throughout a hierarchy of processing stages can account for some characteristics of object recognition (Yamins and DiCarlo 2016; Kriegeskorte 2015). Recently, it has been demonstrated that a similar hierarchical correspondence can also be observed during natural language processing (Schrimpf et al. 2021; Caucheteux and King 2022; Goldstein et al. 2022). These findings do not imply that DL has fully captured how these processes operate in the human brain. Still, DL has contributed to better characterizing the computational principles underlying them. Can DL similarly contribute to studying Theory of Mind (ToM)?

ToM is an essential ability of humans to infer the mental states of others, such as, for example, their perceptual states, beliefs, knowledge, desires, or intentions (for review Apperly 2010; Heyes and Frith 2014; Tomasello 2014; Wellman 2014). It is a field that has been studied extensively (Wellman et al. 2001; Wellman 2014; Rakoczy 2022), with thousands of research articles having investigated different aspects of ToM, its development (Apperly 2010; Wellman 2014; Rakoczy 2022) and its deficits in different psychiatric disorders such as autism (Apperly 2010; Baron-Cohen 2000). There are many tasks that have been designed for testing the different aspects of ToM in humans (Apperly 2010; Quesque and Rossetti 2020; Rakoczy 2022). For instance, one crucial milestone for ToM in human children is whether they understand that a person can have a false belief, i.e., the person believes that something is the case although in reality it is not (Wellman et al. 2001; Siegal 2008; Wellman 2014). However, as we will discuss below, acquiring ToM is not equivalent to passing a false-belief task, but rather is a complex skill that takes years to develop in humans (Tomasello 2014; Wellman 2014, 2020; Rakoczy 2022). For instance, a child who passes the false-belief task, is still not capable of understanding that a person might pretend to believe or feel one way but actually believes or feels the opposite (Wellman 2014, 2020). Given the complexity of ToM, it is unclear whether any of the success of DL in vision and language tasks would carry over to ToM.

Nevertheless, it is worth investigating ToM with DL for several reasons. First, there is a pressing concern to align the strategies discovered by DL models to human needs, desires, and values (Leike et al. 2018; Christian 2020; Kenton et al. 2021). Furthermore, a practical goal of DL systems is to build artificial agents that interact with, support, and understand humans. To achieve this, there might not be a way around studying ToM because, at least according to some prominent views about communication, ToM is necessary for the emergence of meaningful communication and language (Tomasello 2010, 2014; Scott-Phillips 2014; Mercier and Sperber 2017). According to this perspective, the only way to build agents that can communicate with humans or each other in a meaningful fashion is to understand and develop ToM capabilities.

Finally, given that there are still many open questions about ToM even in humans (Siegal 2008; Call and Tomasello 2011; Apperly 2010; Wellman 2014; Heyes and Frith 2014; Quesque and Rossetti 2020), modern DL tools could in principle help to understand ToM at an unprecedented level. For instance, while most researchers agree that language is important for the development of ToM (Siegal 2008; Wellman 2014), the study of the role of language in humans has its own constraints. For example, it would be unethical to deprive children of language input. In DL models, on the other hand, one can easily parametrically change the amount of language input and precisely manipulate the words and phrases the model experiences. DL models are also useful because one can open or modify individual components of the DL architecture. For instance, in DL systems, one can find the artificial neurons and neuron populations that model other agents and then manipulate
these components. This way it is possible to evaluate the necessity or sufficiency of specific elements of ToM in diverse tasks. In short, work on DL can illuminate aspects of ToM that are otherwise hard, if not impossible, to study in humans. However, this potential can be realized only if the DL models of ToM come close enough to natural ToM.

Our goal in this paper is to highlight common problems in DL-based ToM research and to provide a concise overview of prior work done in humans that could inspire future DL systems. In the next section, we start by comparing ToM to vision to demonstrate that it will be more difficult to develop systems that have ToM-like abilities. In Sect. 3, we will consider the problem of shortcuts: current DL-systems likely develop decision-rules that do not amount to ToM-skills. In Sect. 4, we explain that ToM is not a skill that emerges from training on a specific task. Section 5 looks at inherent biases that humans might have to develop ToM. Section 6 is focused on how DL might learn ToM with humans in the loop. In Sect. 7, we arrive at the question of how to evaluate ToM in future DL systems.

2 Will Theory of Mind be more challenging than vision for DL?

Four specific aspects have enabled DL to help make advances in visual recognition and in shedding light on the processes underlying vision in the brain. (1) Appropriate training data: DL models are trained on datasets that are directly relevant for biological vision (e.g., Imagenet Deng et al. 2009). (2) Built in and learned invariances: DL systems have achieved a certain level of robustness and generalization due to the invariances implemented in the network (e.g., convolutional kernels to achieve translational invariance) or learned from the variability in which an object is presented in the training dataset; (3) DL systems have a training objective that is similar to biological vision: machine vision has the straightforward goal to accurately recognize objects in a scene (or to segment or localize them), and it is reasonable to posit that at least one goal (if not the goal) of biological vision is also to recognize objects; (4) there exist appropriate neuroscientific comparison data: we have a reasonably good understanding of the areas involved in vision, how they are connected and organized. Decades of work on the visual system have revealed the visual processing hierarchy, which can and has been compared to the hierarchy of transformations learned by DL systems (Güçlü and van Gerven 2015; Seibert et al. 2016; Cichy et al. 2016; Yamins and DiCarlo 2016; Kriegeskorte 2015; Kuzovkin et al. 2018).

In ToM research, the connection between DL and biology is much more challenging to establish. In DL, ToM is mainly studied with deep reinforcement learning (DRL), where the data that the agent experiences and the objective are intermingled. The objective is only implicitly defined via the reward structure, which drives the agent’s actions. In turn, the agent’s actions determine the reward that it experiences and thus create constant feedback. Hence, in DRL, the reward structure of the task is a central factor that determines precisely what the agent does and learns. However, in the case of ToM, there might not exist a simple and explicit cost function or a “reward structure” that would necessarily lead to the emergence of ToM, although recent work has started to unravel the cost functions for social referential communication (Jara-Ettinger and Rubio-Fernandez 2021).

The problem is further exacerbated by the fact that there is no appropriate neuroscientific comparison data. In vision, it is relatively straightforward to present the same images to humans and deep neural networks and study the correspondences between these representations (Güçlü and van Gerven 2015; Seibert et al. 2016; Cichy et al. 2016; Kuzovkin et al. 2018). In particular, one can compare the activation of a specific layer of the deep
neural network with the activity patterns measured in the human brain and ask how much variance of the biological data is explained by the DL model. For ToM, such data does not yet exist and therefore it is hard to study the correspondence.

In addition, the datasets that are used to train machine vision algorithms have variability with regard to the exact position, angle, etc. of the objects, thus enabling invariances to arise. Even more than that, DL networks that can cope with visual recognition tasks have in-built structural biases such as convolutions. In contrast, we have currently very little understanding of invariances desired (to be implemented or gained through learning) to achieve a ToM that is robust and generalizes to situations different from the trained scenarios.

In summary, ToM is more complex than vision and thus it will be more challenging to develop systems that have ToM-like abilities. The success of DL in vision has benefited from factors that are not (yet) available for ToM.

3 Shortcuts in Theory of Mind tasks

When one wants to develop a DL system that plays Atari or GO, one lets the system play Atari games or GO. This is not so simple in ToM: the fundamental problem is that no task unequivocally taps into ToM (as explained in the next section). Therefore, it is impossible to optimize DL systems directly for ToM. Instead, researchers use tasks that, according to their intuition, need ToM. A significant problem with this approach is that even if we humans think that one or another particular task requires ToM-like skills such as perspective taking or intention understanding, it does not mean that the DL agent indeed acquires ToM-like skills when trained on this specific task. The only pressure the agents experience is to maximise the reward on the task. If the reward can be obtained differently, without any ToM-like skills, the agent can learn to do so.

It has been observed that DL models learn simple decision rules called shortcuts (Geirhos et al. 2020). Shortcuts are tricks that enable the DL agent to get a high score on a particular task without using the processes the researchers intended the DL system to have (Geirhos et al. 2020; Baker 2020; Lehman et al. 2020). For instance, in object recognition, the DL system might learn to classify the objects based on a background element instead of learning the features of the target objects (Geirhos et al. 2020). Given that shortcuts can arise even in vision, which is arguably simpler than ToM, then one should also expect these shortcuts in ToM, where the lack of a strong and specific pressure to infer hidden states of other agents can result in DL agents learning simpler decision rules.

Regarding ToM, this problem of a shortcut is not specific to DL. Namely, there is a long-lasting discussion about whether infants and chimpanzees possess some ToM-like skills or whether they use low-level cues (i.e., shortcuts) to solve these tasks. For instance, in experiments with chimpanzees, it was observed that the subordinate animal behaved as if they could take the perspective of the dominant animal (Hare et al. 2000, 2001). In particular, in a competitive setting, the animals seemingly could infer what the other chimpanzee could see (Hare et al. 2000) and know (Hare et al. 2001). This behavior was akin to the process of visual perspective-taking, which is an ability to appreciate the visual experience of another person and that has been thoroughly studied in human ToM research (Apperly and Butterfill 2009; Apperly 2010). However, these results were immediately followed by discussion whether chimpanzees use low-level cues or ToM to solve tasks like this.
Similarly, a well-known paper measured eye-movements (in the so-called anticipatory looking test) to claim that infants understand false beliefs (Onishi and Baillargeon 2005). This led to an ongoing discussion about the reliability of such measures and whether they really demonstrate ToM understanding in infants (Ruffman and Perner 2005; Heyes 2014; Rakoczy 2022). When researchers used the anticipatory looking test on chimpanzees and showed that based on this measure, chimpanzees pass a false-belief task (Krupenye et al. 2016), a discussion ensued about how much of this can be traced back to ToM (Heyes 2017; Krupenye et al. 2017; Horschler et al. 2020). Even more, the same argument can be made about adult humans: under certain circumstances, for instance when resources are occupied by another task, humans use more automatic responses, general knowledge and heuristics (i.e., shortcuts) to solve ToM tasks (Apperly and Butterfill 2009; Apperly 2010). Given that we cannot be sure whether adult humans, infants and animals with similar brains to us are using ToM or shortcuts, we should be wary of attributing ToM-like skills to artificial agents performing simpler tests.

This means that even if the researcher intends to study ToM, the DL agents might potentially find simple shortcuts. In Table 1 we highlight several cases where the agents potentially learned a shortcut. To bring one concrete example, Labash et al. (2020) were inspired by the perspective-taking experiments with chimpanzees described above (Hare et al. 2000, 2001). The authors (Labash et al. 2020) implemented two agents (“subordinate” and “dominant”) and investigated whether the behavior of the agents revealed some rudimentary skills of perspective-taking, similar to the experiments done with chimpanzees (Hare et al. 2000, 2001). Indeed, after training, a subordinate agent solved the task: go to food if the dominant is not observing; avoid the food if the dominant is observing (Fig. 1).

In humans, the inference of whether some item is visible from the point of view of another is usually accompanied by the possibility of imagining such a point-of-view itself (Apperly 2010). Hence, in this experiment, one might assign such ToM skill to an artificial agent that successfully solves the task (Labash et al. 2020). Yet, a simple “shortcut” based on the position and orientation of the other agent in relation to the food position is enough to successfully determine the agent strategy (Fig. 1). In short, given a reward scheme and a fixed environment, there are simple geometric combinations that allow the agent to succeed without representing other agents’ perceptual states. Likewise, Rabinowitz et al. (2018) and Nguyen and Gonzalez (2020) used an environment and task in which the DL agent could exploit simple combinations of geometrical features such as positions and distances between elements to solve the task successfully.

Further, in game-theoretic scenarios, it is known that there exist simple strategies such as tit-for-tat (repeat opponent’s action from the previous round) as in Foerster et al. (2017), Freire et al. (2019), and Robert et al. (1984) which can lead to significant payoffs. In games that seemingly required coordination or communication (Matiisen et al. 2018; Lowe et al. 2017), the initial configuration of targets and trajectories can determine the agent’s actions without the need to coordinate or infer each other’s goals. This problem cannot be mitigated simply by suggesting new or more complex tasks. For example, one recent prominent proposal has been that the card-game Hanabi is suitable for studying ToM with DL (Bard et al. 2020; Foerster et al. 2019; Fuchs et al. 2021). However, even in Hanabi there will be statistical regularities which can be exploited by DRL agents, making it difficult to prove that ToM-like abilities indeed contributed to solving the task.

The fact that DL agents use shortcuts in ToM tasks does not imply that this work bears no relevance at all to understanding human-like ToM. This is because, as illustrated above, humans also use shortcuts in ToM tasks. In psychological ToM-research, there is an important theoretical distinction between two different systems underlying ToM-skills (Apperly
### Table 1  DL agents potentially learn shortcuts instead of ToM-skills

| Task (Papers)                                                                 | ToM-like expected solution                                                                 | Potential shortcut                                                                 |
|-------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| Agent avoids consuming rewards observed by dominant agent (Labash et al. 2020) | Agent learns the perspective of the dominant agent                                           | Combination of position and orientation of dominant and subordinate in addition to food position |
| Agent observes a simulation of the false belief test (Rabinowitz et al. 2018; Nguyen and Gonzalez 2020) | Agent learns to differentiate between internal states of other agents to predict where the observed agents will go | Agent exploits a combination of features such as positions and distances between elements |
| Game-theoretic tasks (Foerster et al. 2017; Freire et al. 2019; Robert et al. 1984) | Predict the other agent’s policy to cooperate or defect                                     | Tit-for-tat strategy                                                               |
| Three agents cooperate to cover three landmarks (Matisen et al. 2018; Lowe et al. 2017) | Agents learn the landmarks that other agents intend to cover                                | Agents exploit the initial configuration of agents and landmarks                   |
| Hanabi card game (Bard et al. 2020; Foerster et al. 2019; Fuchs et al. 2021)    | Agents develop beliefs about the beliefs of other agents to provide the best hints          | Decisions are made based on the statistical regularities                           |

The first column briefly specifies the task used by the researchers. The second and the third column of the table highlight the difference between the ToM-like behavior expected by the agent compared to a potential shortcut that might lead to solving the task. Note that we are not saying that in these papers the DL agents indeed used these shortcuts. We are simply pointing out that these are potential shortcuts that the agents might have used.
and Butterfill 2009; Apperly 2010; Low and Perner 2012; Butterfill and Apperly 2013). One is automatic, implicit and inflexible, whereas the second one, that emerges later in development, can be used in a more flexible and explicit manner (Apperly and Butterfill 2009; Apperly 2010; Low and Perner 2012; Butterfill and Apperly 2013).

In summary, given that the problem of shortcuts is evident even for vision, one would expect this problem to be even more prominent for ToM tasks. However, in ToM, shortcuts are not only a nuisance, but rather also constitute one valid way how humans and other animals solve ToM tasks. Similarly to human ToM research, DL research into ToM could take into account that tasks can be solved via shortcuts and differentiate between different types of ToM. This would make the DL research more relevant for the study of ToM in humans and animals and would bring us closer to building algorithms that have more human-like ToM.

4 Towards Theory of Mind: beyond a task

All these previously highlighted works have tried to understand ToM as a skill that can be learned based on some particular task. However, perhaps ToM is not a skill emerging from some task. It is not clear whether there exists a simple and explicit cost function or set of rewards for biological systems that would necessarily lead to the emergence of ToM. It could be a complex cost function that cannot simply be optimized by training on a specific task.
In ToM research done on humans, there are tasks like the classic “Sally–Anne” task used to measure ToM. In this task, the child is presented with two dolls (Sally and Anne) that enact a scene wherein Sally hides a marble in the room before leaving, after which Anne removes it from its location. ToM is displayed when the child understands Sally’s false belief that the marble is still where she left it (when Sally returns). This task requires understanding that people behave according to what they believe, even if that belief is not aligned with reality. Thus, other people may be holding a false belief based on their own perspective. However, nobody would claim that children learn ToM by a repeated confrontation with the “Sally–Anne” task. ToM is not inherent to a task; it is a specific way humans deal with these types of situations. So, it cannot be assumed that simply because a task might require ToM in humans, it does so also in DRL agents. They might solve it differently, by learning a shortcut. To humans, Sally–Anne-like tasks serve as an evaluation platform of a larger and more complex system trained independently in an open-ended fashion.

Open-endedness departs from the single-task paradigm to an unbounded number of tasks (or even no task at all, simply a world with different possibilities). To the study of ToM, open-ended environments could provide a fruitful playground where agents coordinate, cooperate and compete to solve tasks and, possibly, learn similar strategies to ToM in humans. Like human children, DL agents might need to be and learn in an open-ended environment where ToM skills are necessary and might be acquired through interaction with other agents. Recent work (Guss et al. 2019; Wang et al. 2020; Samvelyan et al. 2021; Team et al. 2021; Hafner 2021) has shown how powerful open-endedness can be for learning complex behavior. Particularly, Team et al. (2021) introduces XLand, a vast environment where multiple agents learn from a spectrum of completely cooperative to fully competitive tasks. Agents trained on XLand learn complex strategies to solve any given task, but it is unknown whether ToM is one of these strategies. We encourage researchers to study whether the learning of ToM-like strategies can emerge from complex environments such as XLand.

Taken together, although there are tasks that are used to measure ToM in humans, there are reasons to think that they are not suitable as benchmarks for DL systems. Open-ended environments might be preferable to develop DL agents that have ToM-like capabilities.

5 Towards Theory of Mind: which biases are needed?

It is important to note, we are not claiming that ToM-like skills would “pop out” from DL agents playing in open-ended environments just like that. Developing ToM requires an open-ended environment, but it might take more than simply better data to acquire ToM. Specifically, there might be several inductive biases and constraints in the human brain which enable acquiring ToM. Many of these biases are still unknown, but here we list some of the possible venues for exploration.

First, there could be biases for attention. For example, recognizing and distinguishing other human beings is important and hence there is an innate bias for attending to faces (Johnson 2005; Reid et al. 2017). In particular, preferences for faces over similarly configured non-face objects are present in neonatal infants (Farroni et al. 2005) and even in fetuses in the third trimester of pregnancy (Reid et al. 2017). Similar biases likely exist for drawing an infant’s attention to speech (Perszyk and Waxman 2018), hands (Ullman et al. 2012), eyes and gaze-direction (Grossmann 2017), and biological motion (Simion et al.
2008). These early biases make sure that the child learns about the aspects of the world that are informative about the minds of other persons.

Second, some of these biases might be structural. For example, the human brain has special circuits devoted to ToM—areas where activity is selectively evoked by tasks that involve considering the minds of other people (Saxe and Young 2013; Koster-Hale et al. 2013). It is currently unknown whether these circuits underlie some computations or structural biases that are specific to ToM. Similarly to convolutions—structural building blocks in convolutional neural networks that help to achieve translational invariance in visual recognition tasks—these constraints would assist the human brain to extract features and information relevant for ToM. The trouble is that we have a very limited understanding of what these structural biases might even be. Importantly, this question might be better tackled with tools from DL than with methods from cognitive science and neuroscience.

Third, research with human children has described certain steps along the way to developing full-blown ToM (Wellman and Liu 2004; Wellman 2014, 2020). First, a child needs to understand that other people can have diverse desires (i.e., people desire different things). Next, a child comes to understand diverse beliefs (i.e., people have different beliefs, even about the same situation). According to this well-established framework, the third stage is the state of knowledge-access (where the child understands that “something can be true, but someone without access to it would be ignorant of it”, Wellman 2014). Only the fourth step is the stage where the child understands false belief (i.e., the child knows that something is true, but is aware that someone else might believe something different). Finally, the authors describe a state called hidden emotion (Wellman 2014) or hidden mind (Wellman 2020), according to which the child understands that desires, beliefs, and knowledge (i.e., internal states) might not be apparent in a person’s behavior.

This does not mean that DL definitely has to emulate these steps to acquire ToM. However, it might be informative to keep in mind that developing ToM takes time and usually progresses along this sequence in humans. One possibility is that these stages might constitute a curriculum for training DRL agents (Matiisen et al. 2019; Forestier et al. 2017). Alternatively, these steps might simply constitute points of comparison and useful benchmarks. Also, the progression of these steps might also indicate something about how easy or complex it is to acquire that stage, not only in humans but also in DL models. In particular, it is relatively easy to learn the stage of “diverse desires”, because desires are clearly evident in behavior and thus could be learned from visual input. On the contrary, it seems very hard if not impossible to learn the last step (hidden mind states) from visual input alone, as the overt behavior is dissociated from the internal states in the cases relevant to this step.

Based on the human ToM research, one can say that children represent the minds of others and they acquire it in a step-by-step manner. In other words, at least some aspects of the representation of the minds of others are learned. Furthermore, we also know that one crucial input for acquiring full-blown ToM is language (Milligan et al. 2007; Hale and Tager-Flusberg 2003; Wellman 2014, 2020; Peterson et al. 2005). For instance, deaf children born to parents who do not master sign language develop ToM much later and their development is dependent on learning sign language (Peterson et al. 2005; Wellman 2014, 2020). Furthermore, a recent meta-analysis demonstrated that language training enhances performance in ToM tasks (Hofmann et al. 2016). Currently, there is no evidence that the later steps of ToM (i.e., understanding of false belief and hidden mind states) can develop without language input. Thus, DL agents in these open-ended environments might need to be combined with the capabilities of large language models (Devlin et al. 2018; Radford et al. 2019; Brown et al. 2020; Fedus et al. 2021). Here also lies one potential avenue for
DL to significantly contribute to the ToM research. Namely, the assumption in the research on human ToM is that specific linguistic input drives the development of ToM skills. For instance, using verbs that relate to specific mental states (e.g., “know”, “think”, “believe”) could underlie the understanding of these mental states (Hofmann et al. 2016; Rakoczy 2022). One could provide the DL models with the exact or at least similar language input as infants and young children get and manipulate the amount of ToM-specific linguistic input. In DL models, one can not only establish the importance of language input, which most ToM-researchers would agree with, but specifically investigate which type of linguistic input (e.g., which exact phrases) drive specific ToM-skills.

Finally, there are still many unknowns about ToM in humans (Apperly 2010; Wellman 2014; Heyes and Frith 2014; Quesque and Rossetti 2020). In particular, the neural and computational basis of ToM is relatively unknown and unexplored (but see Baker et al. 2017; Jara-Ettinger 2019). Historically, one research direction has been about explaining ToM through the activity of so-called mirror neurons (Gallese and Goldman 1998; Rizzolatti and Sinigaglia 2008). However, strong criticisms of this view (Hickok 2014; Heyes 2010) have curbed the enthusiasm of these early claims and the relative interest in mirror neurons has decreased considerably (Heyes and Catmur 2022). As explained below, we see such controversies as an opportunity for DL researchers.

In summary, ToM is unlikely to simply arise from large amounts of training data. Rather, human brains likely have specific inductive biases that enable them to acquire ToM. The goal of future research is to understand these biases and implement them in DL systems.

6 Theory of Mind with humans in the loop

Instead of waiting for these biases to emerge from open interactions with other agents or for researchers to establish the correct structural biases needed for ToM, it might also be possible for humans to direct the DRL agents to learn the required biases for ToM. One aspiration in DRL is to learn from experts; these can be humans or even other DRL agents. Traditionally, methods like imitation learning (IL; Pomerleau 1991; Schaal 1999) or inverse reinforcement learning (IRL; Ng and Russell 2000) have been used to learn a policy or reward function, respectively. These methods require experts to demonstrate the desired behavior, which can be incredibly hard, arduous, or impossible in many interesting cases. Recent research (Christiano et al. 2017; Wu et al. 2021; Lampinen et al. 2021; Abramson et al. 2021; Team et al. 2022) has proposed variations of these methods capable of producing complex agents.

A technique that has been shown to scale with the complexity of the desired behavior is learning from human preferences (Christiano et al. 2017). This method does not rely on demonstrations by human experts; instead, this approach requires humans to rate some of the agent’s trajectories and trains a reward function via supervised learning using these ratings. In contrast to more traditional reward schemes based on handcrafted rules, the learned reward function evolves with the agent’s learned behavior in a curriculum-learning fashion and, to some extent, reduces the description of the desired behavior to a simple rating.

A similar but arguably more scalable approach to learning from humans is explanation-based learning. For instance, Lampinen et al. (2021) propose to use explanations as an auxiliary task. In other words, the agent needs to explain its behavior using natural language. Although the authors use synthetic explanations, their study shows that explanations can
bias the agent towards using features with more generalization power and, at least in some cases, to identify the world’s causal structure.

If ToM is not an emergent property of DRL agents, maybe aspects of ToM can be incorporated through explanations. Recent work (Akula et al. 2022) takes a step in this direction by proposing an image classification system that iteratively explains its decisions to a human in the hope of finding better decisions. Research has also demonstrated that real-time visualizations of the robot’s internal decisions were the most effective means for promoting human trust in the robot’s behavior (Edmonds et al. 2019).

Taken together, these approaches could provide a natural platform to facilitate the learning of human-like ToM by DL agents. Human ratings and explanations could encourage finding solutions that require ToM.

7 Evaluating Theory of Mind in DL agents

As of now, ToM in DL is usually evaluated through the performance of some task. However, as we have highlighted in this paper, DL systems make use of shortcuts (Geirhos et al. 2020), i.e., they learn to utilize decision rules that are simpler than the ones intended by the researchers.

We propose that when studying ToM with DL, the primary focus and main contribution to ToM research could be opening up the network’s representations. For instance, research on ToM through DL could provide insights on the debated role of mirror neurons (Heyes 2010; Hickok 2014). Does something akin to mirror neurons arise in the DL systems trained in open-ended environments? What would happen if one were to modify or remove these neurons from the system? Tackling these questions would be informative to all researchers studying ToM. Yet, up to now, most of the work done in DL that examines ToM has not inspected the trained model weights, activations, or correlations with specific aspects of ToM (see Triesch et al. 2007 for an early exception). If ToM representations of hidden variables of other agents are developed and used by the network, it should be possible to isolate such representations.

Recently, the field of interpretability has received much attention from the research community (Du et al. 2019; Olah et al. 2018; Molnar 2020; Alharbi et al. 2021), leading to novel and powerful methods that enable a better understanding of the learned representations. We recommend taking advantage of these developments for the study of ToM.

In particular, we bring to the reader’s attention the work done by Hilton et al. (2020) where the authors proposed to combine multiple interpretability techniques like feature visualization, attribution, and dimensionality reduction to understand vision in DRL agents. By applying these techniques to the agent’s neural network, they could identify failure cases like hallucinations of reward-leading features or even make the agent “blind” to specific high-level features, revealing how relevant these are to the task.

More straightforward methods like linear probing (Alain and Bengio 2016) can also be used to test if the explicit representations on the intermediate layers of the network encode some aspects of ToM. After a specific network component has been located that seems responsible for some aspect of ToM, we recommend performing an ablation (Meyes et al. 2019) of these parts of the network to check the necessity of those components for solving the task. If the agent can still perform the task after the network component is removed,
it is likely that the DL agent was using shortcuts or that this particular component is not required for performing the task.

Applying similar methodologies to agents trained in cooperative and/or competitive environments like XLand (Team et al. 2021), Hide and Seek (Baker 2020) or Capture the Flag (Jaderberg et al. 2019) could bring us closer to understanding ToM-like capabilities.

In summary, in DL systems it is possible to study the mechanisms of ToM in an unprecedented manner. However, to realize this potential, the environments, tasks and inductive biases in the DL systems need to support the emergence of ToM in DL agents.

8 Conclusion

ToM is a central facet of human intelligence (Apperly 2010; Tomasello 2010, 2014; Heyes and Frith 2014; Scott-Phillips 2014; Mercier and Sperber 2017). Inspired by the success of DL in understanding biological vision (Güçlü and van Gerven 2015; Seibert et al. 2016; Cichy et al. 2016; Kriegeskorte 2015; Yamins and DiCarlo 2016) and language processing (Schrimpf et al. 2021; Caucheteux and King 2022; Goldstein et al. 2022), over the last years a challenge has emerged to develop DL agents that can mimic aspects of ToM.

In this paper, we surveyed papers that have investigated ToM with DL. We have observed that DL models can develop shortcuts which means that although the researcher intends the DL system to learn ToM, the system might actually learn a much simpler decision rule. This is a problem and a challenge, but also an opportunity for future research into ToM.

DL architectures and their learning algorithms are not the ultimate brain-like learning system, but they do provide scientific models (Cichy and Kaiser 2019) that can guide our understanding of higher mental functions such as ToM. So far, DL remains our best source to understand the working algorithms in tasks similar to those that animals have to solve. Given the difficulty of monitoring all the relevant variables in real brains, the fact that we can open these artificial algorithms and analyze them in detail provides a source of inspiration that we can not afford to leave unexplored.

Acknowledgements We are thankful to Daniel Majoral, Roman Ring, Kadi Tulver and four anonymous reviewers for constructive comments on the manuscript.

Funding This research was supported by the European Regional Funds through the IT Academy Programme, the Estonian Research Council Grants PSG728 and PRG1604, the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No. 952060 (Trust AI), the Estonian Centre of Excellence in IT (EXCITE) Project Number TK148, and the Project CardioStressCI (ERA-CVD-JTC2020-015) from the European Union’s ERA-CVD Joint Transnational Call 2020.

Declarations

Conflict of interest We declare that none of the authors have competing financial or non-financial interests as defined by Nature Portfolio.

References

Abramson J, Ahuja A, Barr I, Brussee A, Carnevale F, Cassin M, Chhaparia R, Clark S, Damoc B, Dudzik A, Georgiev P, Guy A, Harley T, Hill F, Hung A, Kenton Z, Landon J, Lillicrap T,
Mind the gap: challenges of deep learning approaches to Theory...
Gallese V, Goldman A (1998) Mirror neurons and the simulation theory of mind-reading. Trends Cogn Sci 2(12):493–501
Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA (2020) Shortcut learning in deep neural networks. Nat Mach Intell 2(11):665–673
Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, Nastase SA, Feder A, Emanuel D, Cohen A et al (2022) Shared computational principles for language processing in humans and deep language models. Nat Neurosci 25:369–380
Grossmann T (2017) The eyes as windows into other minds: an integrative perspective. Perspect Psychol Sci 12(1):107–121
Güçlü U, van Gerven MA (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci 35(27):10005–10014
Guss WH, Houghton B, Topin N, Wang P, Codel C, Veloso M, Salakhutdinov R (2019) MineRL: a large-scale dataset of Minecraft demonstrations. arXiv preprint. arXiv:1907.13440
Hafner D (2021) Benchmarking the spectrum of agent capabilities. arXiv preprint. arXiv:2109.06780
Hale CM, Tager-Flusberg H (2003) The influence of language on theory of mind: a training study. Dev Sci 6(3):346–359
Hare B, Call J, Aagnosta B, Tomasello M (2000) Chimpanzees know what conspecifics do and do not see. Anim Behav 59(4):771–785
Hare B, Call J, Tomasello M (2001) Do chimpanzees know what conspecifics know? Anim Behav 61(1):139–151
Heyes C (2010) Where do mirror neurons come from? Neurosci Biobehav Rev 34(4):575–583
Heyes C (2014) False belief in infancy: a fresh look. Dev Sci 17(5):647–659
Heyes C (2017) Apes submentalise. Trends Cogn Sci 21(1):1–2
Heyes C, Catmur C (2022) What happened to mirror neurons? Perspect Psychol Sci 17(1):153–168
Heyes CM, Frith CD (2014) The cultural evolution of mind reading. Science 344(6190):1243091
Hickok G (2014) The myth of mirror neurons: the real neuroscience of communication and cognition. WW Norton & Company, New York
Hilton J, Cammarata N, Carter S, Goh G, Olah C (2020) Understanding RL vision. Distill 5(11):e29
Hofmann SG, Doan SN, Sprung M, Wilson A, Ebesutani C, Andrews LA, Curtiss J, Harris PL (2016) Training children’s theory-of-mind: a meta-analysis of controlled studies. Cognition 150:200–212
Horschler DJ, MacLean EL, Santos LR (2020) Do non-human primates really represent others’ beliefs? Trends Cogn Sci 24(8):594–605
Jaderberg M, Czarnecki WM, Dunning I, Marris L, Lever G, Castaneda AG, Beattie C, Rabinowitz NC, Morcos AS, Ruderman A et al (2019) Human-level performance in 3D multiplayer games with population-based reinforcement learning. Science 364(6443):859–865
Jara-Ettinger J (2019) Theory of mind as inverse reinforcement learning. Curr Opin Behav Sci 29:105–110
Jara-Ettinger J, Rubio-Fernandez P (2021) The social basis of referential communication: speakers construct physical reference based on listeners’ expected visual search. Psychol Rev 129(6):1394–1413
Johnson MH (2005) Subcortical face processing. Nat Rev Neurosci 6(10):766–774
Kenton Z, Everitt T, Weidinger L, Gabriel I, Mikulik V, Irving G (2021) Alignment of language agents. arXiv preprint. arXiv:2103.14659
Koster-Hale J, Saxe R, Dungan J, Young LL (2013) Decoding moral judgments from neural signals of representations of intentions. Proc Natl Acad Sci USA 110(14):5648–5653
Kriegeskorte N (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu Rev Vis Sci 1:417–446
Krupenye C, Kano F, Hirata S, Call J, Tomasello M (2016) Great apes anticipate that other individuals will act according to false beliefs. Science 354(6308):110–114
Krupenye C, Kano F, Hirata S, Call J, Tomasello M (2017) A test of the submetalizing hypothesis: apes’ performance in a false belief task inanimate control. Commun Integr Biol 10(4):e1343771
Kuzovkin I, Vicente R, Petton M, Lachaux JP, Baciu M, Kahane P, Rheims S, Vidal JR, Aru J (2018) Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. Commun Biol 1(1):1–12
Labash A, Aru J, Matilien T, Tampuu A, Vicente R (2020) Perspective taking in deep reinforcement learning agents. Front Comput Neurosci 14:69
Lampinen AK, Roy NA, Dasgupta I, Chan SCY, Tam AC, McClelland JL, Yan C, Santoro A, Rabinowitz NC, Wang JX, Hill F (2021) Tell me why! Explanations support learning of relational and causal structure
Lehman J, Clune J, Misevic D, Adami C, Altenberg L, Beaulieu J, Bentley PJ, Bernard S, Beslon G, Bryson DM et al (2020) The surprising creativity of digital evolution: a collection of anecdotes from the evolutionary computation and artificial life research communities. Artif Life 26(2):274–306
Leike J, Krueger D, Everitt T, Martic M, Maini V, Legg S (2018) Scalable agent alignment via reward modeling: a research direction. arXiv preprint. arXiv:1811.07871

Low J, Perner J (2012) Implicit and explicit theory of mind: state of the art. Br J Dev Psychol 30(1):1–13

Lowe R, Wu YI, Tamar A, Harb J, Pieter Abbeel O, Mordatch I (2017) Multi-agent actor-critic for mixed cooperative–competitive environments. In: Advances in neural information processing systems, vol 30

Matiisen T, Labash A, Majoral D, Aru J, Vicente R (2018) Do deep reinforcement learning agents model intentions? arXiv preprint. arXiv:1805.06020

Matiisen T, Oliver A, Cohen T, Schulman J (2019) Teacher–student curriculum learning. IEEE Trans Neural Netw Learn Syst 31(9):3732–3740

Mercier H, Sperber D (2017) The enigma of reason. Harvard University Press, Cambridge

Meyers R, Lu M, de Puiseau CW, Meisen T (2019) Ablation studies in artificial neural networks. arXiv preprint. arXiv:1901.08644

Milligan K, Astington JW, Dack LA (2007) Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. Child Dev 78(2):622–646

Molnar C (2020) Interpretable machine learning. Lulu.com

Ng AY, Russell S (2000) In: Proceedings of the 17th international conference on machine learning, 2000. Morgan Kaufmann, pp 663–670

Nguyen TN, Gonzalez C (2020) In: Proceedings of the 42nd annual meeting of the Cognitive Science Society (CogSci 2020), 2020

Olah C, Satyanarayana A, Johnson I, Carter S, Schubert L, Ye K, Mordvintsev A (2018) The building blocks of interpretability. Distill 3(3):e10

Onishi KH, Baillargeon R (2005) Do 15-month-old infants understand false beliefs? Science 308(5719):255–258

Perszyk DR, Waxman SR (2018) Linking language and cognition in infancy. Annu Rev Psychol 69:231–250

Peterson CC, Wellman HM, Liu D (2005) Steps in theory-of-mind development for children with deafness or autism. Child Dev 76(2):502–517

Pomerleau DA (1991) Efficient training of artificial neural networks for autonomous navigation. Neural Comput 3(1):88–97. https://doi.org/10.1162/neco.1991.3.1.88

Povinelli DJ, Vonk J (2003) Chimpanzee minds: suspiciously human? Trends Cogn Sci 7(4):157–160

Quesque F, Rossetti Y (2020) What do theory-of-mind tasks actually measure? Theory and practice. Perspect Psychol Sci 15(2):384–396

Rabinowitz N, Perbet F, Song F, Zhang C, Eslami SA, Botvinick M (2018) In: International conference on machine learning, 2018. PMLR, pp 4218–4227

Rakoczy H (2022) Foundations of theory of mind and its development in early childhood. Nat Rev Psychol 1(4):223–235

Reid VM, Dunn K, Young RJ, Amu J, Donovan T, Reissland N (2017) The human fetus preferentially engages with face-like visual stimuli. Curr Biol 27(12):1825-1828.e3

Rizzolatti G, Sinigaglia C (2008) Mirrors in the brain: how our minds share actions and emotions. Oxford University Press, Oxford

Robert A et al (1984) The evolution of cooperation. Basic Books, New York

Ruffin T, Perner J (2005) Do infants really understand false belief? Cogn Dev 9:377–395

Samvelyan M, Kirk R, Kurin V, Parker-Holder J, Jiang M, Hambro E, Petroni F, Kühler H, Grefenstette E, Rocktäschel T (2021) MiniHack the planet: a sandbox for open-ended reinforcement learning research. arXiv preprint. arXiv:2109.13202

Saxe R, Young L (2013) Theory of mind: How brains think about thoughts. Oxf Handb Cogn Neurosci 2:204–213

Schaal S (1999) Is imitation learning the route to humanoid robots? Trends Cogn Sci 3(6):233–242. https://doi.org/10.1016/S1364-6613(99)01327-3

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum JB, Fedorenko E (2021) The neural architecture of language: integrative modeling converges on predictive processing. Proc Natl Acad Sci USA 118(45):e2105646118

Scott-Phillips T (2014) Speaking our minds: why human communication is different, and how language evolved to make it special. Macmillan International Higher Education, New York
Seibert D, Yamins D, Ardila D, Hong H, DiCarlo JJ, Gardner JL (2016) A performance-optimized model of neural responses across the ventral visual stream. bioRxiv, 036475
Siegal M (2008) Marvelous minds: the discovery of what children know. Oxford University Press, Oxford
Simion F, Regolin L, Bulf H (2008) A predisposition for biological motion in the newborn baby. Proc Natl Acad Sci USA 105(2):809–813
Tomasello M (2010) Origins of human communication. MIT Press, Cambridge
Tomasello M (2014) A natural history of human thinking. Harvard University Press, Cambridge
Tomasello M, Call J, Hare B (2003) Chimpanzees understand psychological states—the question is which ones and to what extent. Trends Cogn Sci 7(4):153–156
Triesch J, Jasso H, Deák GO (2007) Emergence of mirror neurons in a model of gaze following. Adapt Behav 15(2):149–165
Ullman S, Harari D, Dorfman N (2012) From simple innate biases to complex visual concepts. Proc Natl Acad Sci USA 109(44):18215–18220
Wang R, Lehman J, Rawal A, Zhi J, Li Y, Clune J, Stanley K (2020) In: International conference on machine learning, 2020. PMLR, pp 9940–9951
Wellman HM (2014) Making minds: how theory of mind develops. Oxford University Press, Oxford
Wellman H (2020) Reading minds: how childhood teaches us to understand people. Oxford University Press, Oxford
Wellman HM, Liu D (2004) Scaling of theory-of-mind tasks. Child Dev 75(2):523–541
Wellman HM, Cross D, Watson J (2001) Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev 72(3):655–684
Wu J, Ouyang L, Ziegler DM, Stiennon N, Lowe R, Leike J, Christiano P (2021) Recursively summarizing books with human feedback
Yamins DL, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19(3):356–365

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.