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

A Preprint

Jaan Aru
Institute of Computer Science
University of Tartu
jaan.aru@gmail.com

Aqeel Labash
Institute of Computer Science
University of Tartu
aqeel.labash@gmail.com

Oriol Corcoll
Institute of Computer Science
University of Tartu
ocorcoll@gmail.com

Raul Vicente
Institute of Computer Science
University of Tartu
raulvicente@gmail.com

April 1, 2022

Abstract

Theory of Mind 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 approaches to Theory of Mind. We highlight that many current findings can be explained through shortcuts. These shortcuts arise because the tasks used to investigate Theory of Mind in deep learning systems have been too narrow. Thus, we encourage researchers to investigate Theory of Mind in complex open-ended environments. Furthermore, to inspire future deep learning systems we provide a concise overview of prior work done in humans. We further argue that when studying Theory of Mind with deep learning, 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 Theory of Mind.

***K*eywords** 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 [1–5] and language processing [6–8]. These findings do not imply that DL has fully captured how these processes operate in the human brain, but DL has definitely 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 [9–12]). 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 [12, 13]. 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 [11, 12, 14]. For instance, a child who passes the false-belief task, is still not capable to understand that a person might pretend to believe or feel one way but actually believes or feels the opposite [12, 14]. 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 [15–17]. 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 [11, 18–20]. 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 is still much controversy about ToM even in humans [9, 10, 12, 13, 21, 22], modern DL tools could in principle help to understand ToM at an unprecedented level. This is because one can open or modify individual components of the DL architecture. For example, suppose a given theory of ToM claims specific elements of ToM. In that case, one could try to disentangle these components in the deep neural network and selectively manipulate and remove them. 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.

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 [23]). 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 [1–5, 24].

In ToM research, the connection between DL and biology is much more difficult to establish. In DL, ToM is mostly 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.

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 [1–3, 24]. 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.

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 ToM is a complex and multifaceted ability [12, 22]). 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 somehow 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 [25]. 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 [25–27]. 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 chimpanzees possess some ToM-like skills or whether they are using low-level cues (i.e., shortcuts) to solve these tasks [28–31]. Given that experiments with chimpanzees are not conclusive on whether these 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.

To bring one concrete example from recent research, [32] were inspired by experiments done with chimpanzees, where a subordinate animal behaved as if she had the ability to take the perspective of the other animal [33, 34]. The authors 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 [33, 34]. 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).

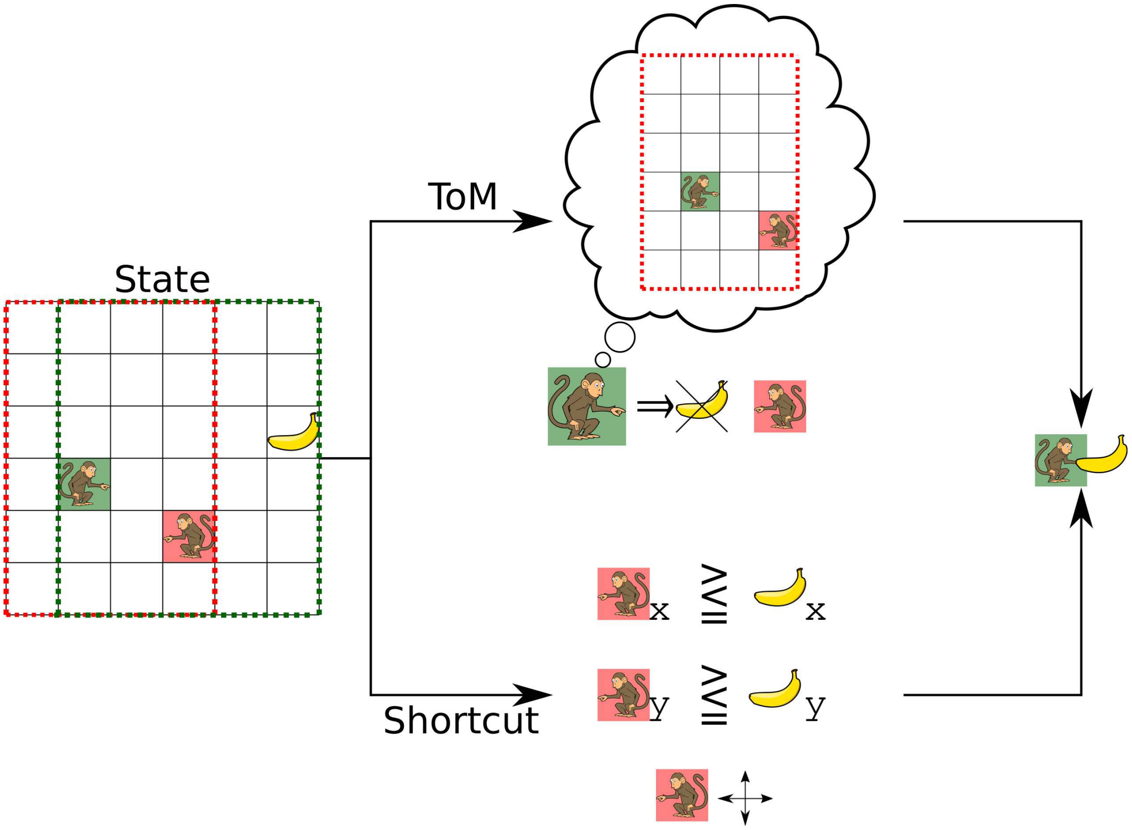

Figure 1: **ToM vs shortcut in artificial agents solving a perspective-taking task. Left** Example of an environment state in the task. The dominant agent is marked with the red square, the subordinate agent with the green square. The visual field of the dominant agent is surrounded by red dashed lines, the visual field of the subordinate with green dashed lines. **Right, top** Based on human behavior one could think that the agent on the green square infers that the dominant agent cannot see the banana and therefore goes for the banana. **Right, bottom** However, the shortcut solution is that the agent simply takes into account the orientation and the distance of the dominant agent without inferring anything about its perspective.

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 [9]. Hence, in this experiment, one might assign such ToM skill to an artificial agent that successfully solves the task [32]. 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, [35, 36] 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 [37–39] which can lead to significant payoffs. In games that seemingly required coordination or communication [40, 41], the initial configuration of targets and trajectories can determine the agent's actions without the need to coordinate or infer each other 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 [42–44]. 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.

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 (possibly requiring many cognitive skills) 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 to understand Sally's false belief that the marble is still where she left it (when Sally returns). 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 [45–49] has shown how powerful open-endedness can be for learning complex behavior. Particularly, [48] 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.

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 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 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 [50, 51]. In particular, preferences for faces over similarly configured non-face objects are present in neonatal infants [52] and even in fetuses in the third trimester of pregnancy [51]. Similar biases likely exist for drawing an infant's attention to speech [53], hands [54], eyes, and gaze-direction [55], and biological motion [56]. 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 [57, 58]. 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 the features and the 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 the tools from DL than with methods from cognitive science and neuroscience.

Third, research on human children has described certain steps along the way of developing full-blown ToM [12, 14, 59]. 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", [12]). 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 [12] or hidden mind [14], 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 [60, 61]. 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 [12, 14, 62–64]. Deaf children born to parents who do not master sign language develop ToM much later and their development is dependent on learning sign language [12, 14, 64]. 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 [65–68].

However, there are still many unknowns about ToM in humans [9, 10, 12, 22]. In particular, the neural and computational basis of ToM is relatively unknown and unexplored (but see [69, 70]). Historically, one research direction has been about explaining ToM through the activity of so-called mirror neurons ([71, 72]). However, strong criticisms of this view ([73, 74]) have curbed the enthusiasm of these early claims and the relative interest in mirror neurons has decreased considerably [75]. As explained in the next section, we see such controversies as an opportunity for DL researchers.

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) [76, 77] or inverse reinforcement learning (IRL) [78] 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 [79–83] 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 [79]. This method does not need demonstrations from experts; instead, it expects 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. [81] propose to use explanations as an auxiliary task. In other words, the agent needs to explain its behavior using natural language. This study shows that explanations can bias the agent towards using features with more generalization power and, in some cases, identify the world's causal structure.

The approaches mentioned above could provide a natural platform to prioritize strategies learned by the agent closer to ToM. Humans could describe tasks requiring ToM either by rating, explaining, or both, and if possible, encourage solutions where the agent needs to model others.

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 [25], 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 [73, 74]. Does something akin to mirror neurons arise in the DL systems trained in open-ended environments? What happens if one would 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 correlation with specific aspects of ToM (see [84] 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 [85–88], 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 [89] 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 [90] 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 ablation [91] of these parts of the network to check the necessity of those components to the task. If this network component is removed from the agent, will it still solve the task? If it can still perform the task, it is likely that the DL agent was using shortcuts or that this particular component is not required for performing the task.

Applying similar studies to agents trained in cooperative and/or competitive environments like XLand [48], Hide and Seek [26] or Capture the Flag [92] could reveal ToM-like capabilities present in these agents. Moreover, these approaches may show what pressures in the environment are important for ToM to emerge.

8 Conclusion

Theory of Mind is a central facet of human intelligence [9–11, 18–20]. Inspired by the success of DL in understanding biological vision [1–5] and language processing [6–8], 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 actually might 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 [93] that can guide our understanding of higher mental functions such as ToM. So far, DL remains our best source for working algorithms in large-scale 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.

Declarations

8.1 Funding

This research was supported by the European Regional Funds through the IT Academy Programme, the Estonian Research Council grants PSG728 and PRG1604 and the European Union's Horizon 2020 research and innovation programme under grant agreement No 952060.

8.2 Conflict of interest/Competing interests

We declare that none of the authors have competing financial or non-financial conflict of interests.

8.3 Acknowledgements

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

References

[1] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[2] Darren Seibert, Daniel Yamins, Diego Ardila, Ha Hong, James J DiCarlo, and Justin L Gardner. A performance-optimized model of neural responses across the ventral visual stream. *bioRxiv*, page 036475, 2016.

[3] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):1–13, 2016.

[4] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[5] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[6] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021.

[7] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10, 2022.

[8] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 2022.

[9] Ian Apperly. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press, 2010.

[10] Cecilia M Heyes and Chris D Frith. The cultural evolution of mind reading. *Science*, 344(6190), 2014.

[11] Michael Tomasello. *A natural history of human thinking*. Harvard University Press, 2014.

[12] Henry M Wellman. *Making minds: How theory of mind develops*. Oxford University Press, 2014.

[13] Michael Siegal. *Marvelous minds: The discovery of what children know*. Oxford University Press, USA, 2008.

[14] Henry Wellman. *Reading minds: How childhood teaches us to understand people*. Oxford University Press, USA, 2020.

[15] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

[16] Brian Christian. *The alignment problem: Machine learning and human values*. WW Norton & Company, 2020.

[17] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of language agents. *arXiv preprint arXiv:2103.14659*, 2021.

[18] Michael Tomasello. *Origins of human communication*. MIT press, 2010.

[19] Thom Scott-Phillips. *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Macmillan International Higher Education, 2014.

[20] Hugo Mercier and Dan Sperber. *The enigma of reason*. Harvard University Press, 2017.

[21] Josep Call and Michael Tomasello. Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5):187–192, 2008.

[22] François Quesque and Yves Rossetti. What do theory-of-mind tasks actually measure? theory and practice. *Perspectives on Psychological Science*, 15(2):384–396, 2020.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[24] Ilya Kuzovkin, Raul Vicente, Mathilde Petton, Jean-Philippe Lachaux, Monica Baciu, Philippe Kahane, Sylvain Rheims, Juan R Vidal, and Jaan Aru. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, 1(1):1–12, 2018.

[25] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[26] Bowen Baker. Emergent reciprocity and team formation from randomized uncertain social preferences. *Advances in Neural Information Processing Systems*, 33:15786–15799, 2020.

[27] Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, 26(2):274–306, 2020.

[28] Daniel J Povinelli and Jennifer Vonk. Chimpanzee minds: suspiciously human? *Trends in cognitive sciences*, 7(4):157–160, 2003.

[29] Michael Tomasello, Josep Call, and Brian Hare. Chimpanzees understand psychological states–the question is which ones and to what extent. *Trends in cognitive sciences*, 7(4):153–156, 2003.

[30] Cecilia Heyes. Apes submentalise. *Trends in cognitive sciences*, 21(1):1–2, 2017.

[31] Christopher Krupenye, Fumihiro Kano, Satoshi Hirata, Josep Call, and Michael Tomasello. A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. *Communicative & Integrative Biology*, 10(4):e1343771, 2017.

[32] Aqeel Labash, Jaan Aru, Tambet Matiisen, Ardi Tampuu, and Raul Vicente. Perspective taking in deep reinforcement learning agents. *Frontiers in Computational Neuroscience*, 14:69, 2020.

[33] Brian Hare, Josep Call, Bryan Agnetta, and Michael Tomasello. Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, 59(4):771–785, 2000.

[34] Brian Hare, Josep Call, and Michael Tomasello. Do chimpanzees know what conspecifics know? *Animal behaviour*, 61(1):139–151, 2001.

[35] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018.

[36] Thuy Ngoc Nguyen and Cleotilde Gonzalez. Cognitive machine theory of mind. In *Proceedings of the 42nd annual meeting of the cognitive science society (cogsci 2020)*, 2020.

[37] Jakob N Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017.

[38] Ismael T Freire, Xerxes D Arsiwalla, Jordi-Ysard Puigbò, and Paul Verschure. Modeling theory of mind in multi-agent games using adaptive feedback control. *arXiv preprint arXiv:1905.13225*, 2019.

[39] Axelrod Robert et al. The evolution of cooperation, 1984.

[40] Tambet Matiisen, Aqeel Labash, Daniel Majoral, Jaan Aru, and Raul Vicente. Do deep reinforcement learning agents model intentions? *arXiv preprint arXiv:1805.06020*, 2018.

[41] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[42] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

[43] Jakob Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, and Michael Bowling. Bayesian action decoder for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 1942–1951. PMLR, 2019.

[44] Andrew Fuchs, Michael Walton, Theresa Chadwick, and Doug Lange. Theory of mind for deep reinforcement learning in hanabi. *arXiv preprint arXiv:2101.09328*, 2021.

[45] William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: A large-scale dataset of minecraft demonstrations. *arXiv preprint arXiv:1907.13440*, 2019.

[46] Rui Wang, Joel Lehman, Aditya Rawal, Jiale Zhi, Yulun Li, Jeffrey Clune, and Kenneth Stanley. Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions. In *International Conference on Machine Learning*, pages 9940–9951. PMLR, 2020.

[47] Mikayel Samvelyan, Robert Kirk, Vitaly Kurin, Jack Parker-Holder, Minqi Jiang, Eric Hambro, Fabio Petroni, Heinrich Küttler, Edward Grefenstette, and Tim Rocktäschel. Minihack the planet: A sandbox for open-ended reinforcement learning research. *arXiv preprint arXiv:2109.13202*, 2021.

[48] Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, et al. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*, 2021.

[49] Danijar Hafner. Benchmarking the spectrum of agent capabilities. *arXiv preprint arXiv:2109.06780*, 2021.

[50] Mark H Johnson. Subcortical face processing. *Nature Reviews Neuroscience*, 6(10):766–774, 2005.

[51] V. M. Reid, K. Dunn, R. J. Young, J. Amu, T. Donovan, and N. Reissland. *The human fetus preferentially engages with face-like visual stimuli.* Current Biology, 2017.

[52] Teresa Farroni, Mark H Johnson, Enrica Menon, Luisa Zulian, Dino Faraguna, and Gergely Csibra. Newborns' preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences*, 102(47):17245–17250, 2005.

[53] Danielle R Perszyk and Sandra R Waxman. Linking language and cognition in infancy. *Annual review of psychology*, 69:231–250, 2018.

[54] Shimon Ullman, Daniel Harari, and Nimrod Dorfman. From simple innate biases to complex visual concepts. *Proceedings of the National Academy of Sciences*, 109(44):18215–18220, 2012.

[55] Tobias Grossmann. The eyes as windows into other minds: An integrative perspective. *Perspectives on Psychological Science*, 12(1):107–121, 2017.

[56] Francesca Simion, Lucia Regolin, and Hermann Bulf. A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences*, 105(2):809–813, 2008.

[57] Rebecca Saxe and Liane Young. Theory of mind: How brains think about thoughts. *The Oxford handbook of cognitive neuroscience*, 2:204–213, 2013.

[58] Jorie Koster-Hale, Rebecca Saxe, James Dungan, and Liane L Young. Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, 110(14):5648–5653, 2013.

[59] Henry M Wellman and David Liu. Scaling of theory-of-mind tasks. *Child development*, 75(2):523–541, 2004.

[60] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher–student curriculum learning. *IEEE transactions on neural networks and learning systems*, 31(9):3732–3740, 2019.

[61] Sébastien Forestier, Rémy Portelas, Yoan Mollard, and Pierre-Yves Oudeyer. Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv preprint arXiv:1708.02190*, 2017.

[62] Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, 78(2):622–646, 2007.

[63] Courtney Melinda Hale and Helen Tager-Flusberg. The influence of language on theory of mind: A training study. *Developmental science*, 6(3):346–359, 2003.

[64] Candida C Peterson, Henry M Wellman, and David Liu. Steps in theory-of-mind development for children with deafness or autism. *Child development*, 76(2):502–517, 2005.

[65] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[66] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[67] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[68] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2021.

[69] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10, 2017.

[70] Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019.

[71] Vittorio Gallese and Alvin Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501, 1998.

[72] Giacomo Rizzolatti and Corrado Sinigaglia. *Mirrors in the brain: How our minds share actions and emotions*. Oxford University Press, USA, 2008.

[73] Gregory Hickok. *The myth of mirror neurons: The real neuroscience of communication and cognition*. WW Norton & Company, 2014.

[74] Cecilia Heyes. Where do mirror neurons come from? *Neuroscience & Biobehavioral Reviews*, 34(4):575–583, 2010.

[75] Cecilia Heyes and Caroline Catmur. What happened to mirror neurons? *Perspectives on Psychological Science*, 17(1):153–168, 2022.

[76] Dean A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

[77] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6):233–242, 1999.

[78] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *in Proc. 17th International Conf. on Machine Learning*, pages 663–670. Morgan Kaufmann, 2000.

[79] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017.

[80] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback, 2021.

[81] Andrew K. Lampinen, Nicholas A. Roy, Ishita Dasgupta, Stephanie C. Y. Chan, Allison C. Tam, James L. McClelland, Chen Yan, Adam Santoro, Neil C. Rabinowitz, Jane X. Wang, and Felix Hill. Tell me why! – explanations support learning of relational and causal structure, 2021.

[82] Josh Abramson, Arun Ahuja, Iain Barr, Arthur Brussee, Federico Carnevale, Mary Cassin, Rachita Chhaparia, Stephen Clark, Bogdan Damoc, Andrew Dudzik, Petko Georgiev, Aurelia Guy, Tim Harley, Felix Hill, Alden Hung, Zachary Kenton, Jessica Landon, Timothy Lillicrap, Kory Mathewson, Soňa Mokrá, Alistair Muldal, Adam Santoro, Nikolay Savinov, Vikrant Varma, Greg Wayne, Duncan Williams, Nathaniel Wong, Chen Yan, and Rui Zhu. Imitating interactive intelligence, 2021.

[83] DeepMind Interactive Agents Team, Josh Abramson, Arun Ahuja, Arthur Brussee, Federico Carnevale, Mary Cassin, Felix Fischer, Petko Georgiev, Alex Goldin, Mansi Gupta, Tim Harley, Felix Hill, Peter C Humphreys, Alden Hung, Jessica Landon, Timothy Lillicrap, Hamza Merzic, Alistair Muldal, Adam Santoro, Guy Scully, Tamara von Glehn, Greg Wayne, Nathaniel Wong, Chen Yan, and Rui Zhu. Creating multimodal interactive agents with imitation and self-supervised learning, 2022.

[84] Jochen Triesch, Hector Jasso, and Gedeon O Deák. Emergence of mirror neurons in a model of gaze following. *Adaptive Behavior*, 15(2):149–165, 2007.

[85] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

[86] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

[87] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[88] Raed Alharbi, Minh N Vu, and My T Thai. Learning interpretation with explainable knowledge distillation. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 705–714. IEEE, 2021.

[89] Jacob Hilton, Nick Cammarata, Shan Carter, Gabriel Goh, and Chris Olah. Understanding rl vision. *Distill*, 5(11):e29, 2020.

[90] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[91] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*, 2019.

[92] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

[93] Radoslaw M Cichy and Daniel Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317, 2019.