Towards Properly Implementing Theory of Mind in AI: An Account of Four Misconceptions

Ramira van der Meulen, Rineke Verbrugge, Max van Duijn

February 2025

This is the draft version of a paper that we expect to submit in April 2025. While the contents in the paper are not binding, this draft version should provide a fairly accurate depiction of the final contents of the paper. For questions about the contents of the paper, please contact the first author (A.van.der.Meulen@leidenuniv.liacs.nl).

Keywords: Theory of Mind, Human-AI Collaboration, Artificial Intelligence, Computing Sciences, Literature Review

1 Introduction

The search for effective collaboration between humans and computer systems is one of the biggest challenges in Artificial Intelligence. One of the more effective mechanisms that humans use to coordinate with one another is theory of mind (ToM). ToM can be described as the ability to 'take someone else's perspective and make estimations of their beliefs, desires and intentions, in order to make sense of their behaviour and attitudes towards the world'. If leveraged properly, this skill can be very useful in Human-AI collaboration.

This introduces the question how we implement ToM when building an AI system. Humans and AI Systems work quite differently, and ToM is a multifaceted concept, each facet rooted in different research traditions across the cognitive and developmental sciences. We observe that researchers from artificial intelligence and the computing sciences (AI & CS), ourselves included, often have difficulties finding their way in the ToM literature. In this paper, we identify four common misconceptions around ToM that we believe should be taken into account when developing an AI system. We have hyperbolised these misconceptions for the sake of the argument, but add nuance in their discussion. We end each section by providing tentative guidelines on how the misconception can be overcome.

1.1 Introducing the Four Misconceptions

The first misconception we discuss concerns modularity. Human ToM is likely the result of multiple brain processes working together, whereas in AI & CS it is often conceptualised as a single module that adds a separate reasoning component to the system. In other words: (1) **"Humans Use a ToM Module, So AI Systems Should As Well"**.

The second misconception concerns when to use ToM to begin with, as we sometimes fall into the trap of thinking that (2) **"Every Social Interaction Requires (Advanced) ToM"**. We know that adding a 'stronger' reasoner to a system can improve its performance, but it is by far not always the more realistic option, also not when looking at the literature on human problem-solving. In fact: In everyday interaction, humans are more likely to use reasoning shortcuts than to overanalyse the situation. A system designed to understand human perspectives should take this into account.

The third misconception concerns limited universality. ToM might be a universal skill among humans, but its expression has differences depending on the human in question (for example, neurological, but also cultural factors play a part). These differences in expression are even stronger when dealing with an entity that does not perceive nor process the world in the same way as a human (such as an AI system). It is important to realise, that in fact, not (3) **"All ToM is the Same"**.

Finally, we address a recent popular claim. It is, for lack of a better term, often said that (4) **"Current Systems Already Have ToM"**. While some systems certainly are able to evaluate ToM on a level of specific tasks, recent models, especially large language models (LLMs), have evoked claims of generalisable ToM. These claims stem from their strong performance at many ToM benchmark tests. While not underestimating recent achievements, we nuance such claims by discussing generalisability of performance beyond linguistic platforms, keeping a clear view of the limitations and challenges ahead.

2 Definitions

While a brief background to most of the concepts and terms is provided in the course of our discussion, there are a few definitions that we highlight beforehand. Apart from 'Theory of Mind' and its relatives 'Mindreading', 'Perspective-Taking', and 'Mentalisation', these are the 'Beliefs, Desires, Intentions Architecture (BDI)', and a distinction between the 'Behavioural' and 'Mechanistic' level of analysis.

Theory of Mind (ToM) The capacity central in this paper has its roots in debates in philosophy of mind [1] and ethology [2] from the 1970s. It goes

back to the basic observation that humans and some other animals, in particular great apes, are able to predict how another individual will act in the near future, and factor this into the planning of their own behaviour. In order to do this, they must reason from the other's perspective and thus have an understanding (a 'theory') of the way they perceive and think about the world (cq. the other's 'mind'). In the ensuing decades, ToM was studied extensively in psychology, where it was soon linked to atypical developmental patterns such as autism [3], a view that was later challenged (see Section 5.1 below). While various perspectives on and definitions of ToM have emerged (such as [4, 5, 6, 7], there are two core aspects that we also include in our definition here: ToM is i) making sense of behavioural elements that one *can* observe by ii) reasoning in terms of what one *cannot* observe: i.e. mental states including beliefs, desires, intentions, motivations, and so on. Such sense-making can serve the purpose of predicting what someone else will do, but also help to deepen one's understanding of the other's perspective without immediate behavioural consequences—or of oneself, given that ToM can also be applied self-reflectively. The terms **Mindreading** and **Perspective-Taking** are used synonymously with ToM in this paper.

Mentalisation One's internal picture of someone else's mental states, and why they act the way they do (also: to mentalise, the action of forming this picture). In other words: The result of having applied ToM to someone.

Beliefs, Desires, Intentions Architecture (BDI) A popular AI model that can store and leverage the beliefs, desires and intentions of the entities it encounters [8]. The model is occasionally also leveraged to represent additional mental states, such as explicit goals.

Behavioural vs. Mechanistic Level We distinguish actions on a mechanistic and on a behavioural level. On a behavioural level, it is difficult to distinguish between ToM-style reasoning, and a pre-learned response (f.e. by an associative reasoner). Imagine you open a door for a colleague. Whether opening this door happens because you have explicitly taken their perspective and reasoned 'this is what they want, and I would like to stay in their good graces', or because 'it feels right, and is what I always do' is visibly indistinguishable to both the colleague and any outside observer. Mechanistically, they are vastly different, however. In the first case, the mechanism is 'ToM'. In the second case, it's an experience-based heuristic.

3 Humans Use a ToM Module, So AI Systems Should As Well

For the discussion of our first misconception, suppose we are building a robot assistant in healthcare. It needs to be able to move around, move its arms, respond to patients, see the world, and so on. Suppose each of these functions

is regulated through the robot's brain, its controller. We could be tempted to add a 'module' that regulates (pro)social behaviour, enabling the assistant to understand and anticipate other entities (humans) around it, and call this its 'ToM module'. This ToM module can be used for *(1) Finding out a hospital patient's specific want, and asking relevant questions* (other-inquiry, empathy, thought evaluation), *(2) realising that the patient moved to the right to get closer to an object they desire* (spatial reasoning, desire evaluation), *(3) working with a patient to get them to take the right medicine and evaluate their health status* (collaboration, motivation understanding, lie evaluation), and so on. While each of these functions benefits from ToM, they are quite different expressions of what seems to be coined as the same phenomenon. This is because, in actuality, Human ToM is the result of a multitude of complex processes.

3.1 ToM Modularity in Robotics

While our example focuses on a future application with high requirements, the treatment of ToM as modular is abundant in the current robotics literature. Examples include Peppers being 'equipped with a simple ToM module' [9], introducing a ToM Manager module to deal with the representation of other's mental states [10], or representing ToM through a human-robot response planner [11]. Many papers use the tri-modular 'Leslie-model' [12], translated to the field Robotics in Scassellati's foundational work 'Theory of mind for a humanoid robot' [13]. This work presents a model that combines a 'movement/mechanical' processor, with two ToM-like systems – one system dealing with intents and goals of agents (i.e. actions), and one system dealing with attitudes and beliefs of agents, somewhat reminiscent of BDI, but as an extra system rather than as an integral part of the machine. This happens in for example 'Theory of mind improves human's trust in an iterative human-robot game' [14], where it is shown that adding processes described as part of human ToM (i.e. the attitudes/beliefs processor), positively corresponds to how trustworthy a robot comes across.

The approaches presented by Scassellati, Ruocco, Kirtay, Devin and Görür have one thing in common: They emulate an interpretation of beliefs, desires, intentions, and so on, but they do it through a central executive unit specifically for ToM, implying this executive unit is the representation of ToM as a whole. This sells short the complexity of the phenomenon. While the reasoning capabilities are enough for the problems set out in the papers, they moreso capture a reasoning ability about specific BDI-components and problems, into one 'theory of mind'.

3.2 ToM Modularity in Other Fields

Research in developmental psychology has taught us that basic skills in ToM, such as the ability to recognise that others may have false beliefs, develop during childhood and early adolescence [15, 16, 17]. The exact mechanisms that

control how senses such as false belief recognition develop is unclear, but literature in developmental psychology theorises that ToM itself is influenced by at least two mechanisms: A mechanism driving subconscious snap decisions, and a mechanism that is used for elaborate, explicit, planning [4][1]. Humans make explicit representations of the mental states of others (System II), but even when unprompted and actively inhibited, humans still consider the perspective of the other (System I) [20]. The literature also refers to this as two-systems theory, but it is argued there might be more systems that similarly emerge from ToM-style decision-making.

We also know from the field of linguistics that ToM is closely tied to the development of language [21, 22, 23, 24]. This effect is strong enough to apply bidirectionally: Experience with ToM-style reasoning also improves one's linguistic abilities [25]. Reversely, being deprived of social/language input because of, for example, a hearing impairment, also appears to influence the ability to solve traditional ToM problems [26]. This ties ToM to (spontaneous) conversation, painting ToM as more of a social skill, once again indicating that it is influenced by situations that draw on processes that require multiple brain regions.

3.3 Neurological Correlates of ToM

A direct argument against a singular ToM module comes from the many, many studies in Neuroscience that have tried to identify the brain region that is 'responsible' for the human social brain and by extension, ToM. Many literature reviews find that papers in the field have varying opinions on where ToM happens in the brain, often concluding that the answer must be 'in many places'[27, 28], or that we need a stricter definition of ToM if we wish to assign it any specific region [27, 29].

For example, an overview study by Carrington and Bailey, spanning 40 neurological studies, found that while both the Medial Prefrontal Cortex/Orbitofrontal Cortex (Decision-making, Expectation Management) and Superior Temporal Sulcus (Facial Recognition, Language, Voice) are 'core' regions that activate for ToM-related tasks, the Temporoparietal Junction (Self-other Distinction) and Anterior- and Posterior-cingulate Cortices (Attention, Motivation, Anticipation, Learning) were also active in over half the ToM-related tasks. Many other brain regions were active for approximately 20-40% of the tasks, indicating that solving tasks with what is considered ToM requires a large part of the brain (even if some specific regions do get more consistent activation than others).

That said, Mahy et al.'s 'How and where: Theory-of-mind in the Brain'[29] finds that different definitions in the field greatly contribute to the question of

[1] This work is very reminiscent of earlier work on biases in human decision making, see [18, 19]

'where' ToM is located, since different definitions lead to different brain regions. They similarly conclude that both the definitions and the paradigms (e.g. evaluating cognitive vs affective states) are not precise enough. We are inclined to agree and feel these are not necessarily problems, but rather show that ToM is diverse, and that what this study encounters, characterises ToM's status as a social skill applicable in many situations. In a similar vein, Schurz et al. [28] argue for an approach that models ToM as a multilevel construct, given that ToM tasks often reflect diverse real-world tasks that draw on different skill sets. Even for more defined tasks, such as the ToM-typical ability to recognise oneself and others' False Beliefs [15], it is difficult to predict how other sociocognitive situations draw on the particular skill.

The difficulties in both of the aforementioned overview studies, combined with the in-depth analysis by Carrington et al., lead us to once more conclude that ToM is a distributed process in the brain. This leaves us with an argument to see ToM as a behavioural phenomenon, underpinned by a set of mechanisms and processes that varies between tasks, contexts, and individuals, instead of a single modular function.

3.4 Conclusion: Forgoing The Artificial Module

This does not mean that specific implementations of ToM are a no-go. What we mainly aim to do here is create awareness. Revisiting the example of the healthcare robot, we see through the evidence in the neurosciences that humans use different parts of the brain for the robot's functions we have labelled as ToM. If we wish to reflect this in a modular fashion, the robot would have to rely on different 'modules' to achieve this result in a human-like fashion. It would mean building a module for each behaviour that is functionally distinct enough to qualify as a unique expression of ToM (e.g., a 'lying-recognition module', an 'other-movement-reasoning' module, and so on).

This said, we may not *need* a specific (set of) module(s) labelled as 'ToM'. The solution here can lie in treating ToM as a result of social skills coming together, without any regions being specifically built to represent a general or specialised ToM skill. ToM is diverse, and treating it as socially emergent might better resemble what humans do than expecting one module to be able to fulfil all of these functionalities at the same time.

This solution does come with its own difficulties: Knowing when to address a specific aspect of ToM. Always monitoring someone's beliefs, desires and intentions is doable, but knowing how to use them (which is also ToM) is tricky. This is a discussion we consider out-of-scope, as it comes with its own set of new challenges.

4 Every Social Interaction Requires (Advanced) ToM

Imagine you are in a hot office room. The air-conditioning is broken, you are hard at work, fully focused, and you are personally not too bothered by the heat. Your colleague at the desk to the right of you asks: "Could you open the window for me, please?". You do acknowledge that it's hot, and snapping out of your focus, you move yourself towards the window handle automatically, as if out of social convention – *It's hot in here, so one opens the window, it's the right thing to do*, not a single second thought. It is then that you realise that your office mate is closer to the window than you are and you have to *pass* them to open the window. What was social convention and nicety before is now more of a social puzzle. *Why did they ask this of me? Are they too lazy to get up themselves? Does our company policy say something about opening windows, and do they know something about this that I do not?* It is likely that we only *now* have started using ToM. Initially, we only only ran the script in our brain to fulfil the favour.

To clarify our example from a scientific point of view: There is ample evidence that we humans use scripts and heuristics in many of our daily interactions [30, 31, 32]. We run in an almost automatic fashion, using mental shortcuts wherever we go, until it is necessary to take a more active look at the situation. For ToM specifically, this dynamic of humans being 'lazy mindreaders' by default has been described by [5], who argues that we usually rely on flow and social synergy, until a situation demands that we 'scale up' our processing efforts. This implies that we continuously (subconsciously) monitor for 'hitches' that the default script/heuristic may run into, but are not always in 'full-on reasoning mode'. This is understandable from a cognitive resources perspective: Using advanced ToM right-away is a major resource drain as it is quite costly to use [33, 34]. It is also quite prone to error as this analysis quickly falls into overthinking perspectives on what the other party thinks about *you* in turn, as it is very difficult for humans to visualise perspectives recursively without making any mistakes [35].

This leaves a major question, both if we want to understand humans, but also if we want to build systems that do: 'What are the origins and nature of all these scripts that humans use?' Clark argues that starting our reasoning from what he refers to as 'Common Ground' [36] already resolves a lot of complexity that other researchers have assumed or identified [37]. Quite similarly, it has been shown in AI research that one can reason effectively using experience-based patterns in a human-like fashion [38], and that in (at least) the medical domain, abstracting single beliefs into higher-level abstractions is an efficient method of dealing with complex domains [39]. In the next section, we explore additional behaviours that come across as ToM on a behavioural level, but do not make use of ToM on a mechanistic level (as explained in Section 2).

4.1 ToM: A Collaborative and Competitive Skill?

Successful cooperation and collaboration are often designated as resulting from the use of Human-ToM. This thought is especially prevalent in the field of evolutionary anthropology, where it is often assumed that ToM has played a major part in human evolution and its eventual development towards a highly social species [40, 41]. Much of this is expressed in the Social Brain Hypothesis, the notion that social complexity was an important driver behind the evolution of intelligence in primates and various species across other orders [42, 43]. It has been shown that various aspects of species' social life, such as group size or mating style, are reflected in their brain size and brain organisation, as well as in social-behavioural repertoire.

Indicators of this have also been shown through agent modelling. ToM has been shown to be a successful strategy when modelling evolution of social behaviour, resembling human behaviour better than making fully rational decisions in the Incremental Centipede Game [44], and research into the negotiation game Coloured Trails has shown that ToM is a skill that grows in benefit as both the environment and resource dilemmas become more complicated [45].

We do not wish to argue with or contest these findings: ToM is provably useful in both collaborative and competitive settings, and even more so in combinations of the two [46]. Yet we do argue that ToM is neither always required nor used by default. A first example of this comes from a model that plays the aforementioned Coloured Trails without making use of explicit ToM [47]. This model trained on human gameplay data is able to perform with similar effectiveness to the original ToM coloured trails model [48], using only statistical approximation. This performance holds up for multiple ToM levels (it was not trained beyond ToM level-2 due to the computational complexity). This is an initial indication that if a system is exposed to enough experience with a situation (e.g., the human gameplay data), it is able to perform similar to ToM.

Additional examples of this exist when we analyse more traditional boardgames. Both Chess [49] and Go [50, 51] have been solved through the use of statistics and emulated experience, even though both games have been argued to be heavily (ToM)-reasoning dependent. The models were purely trained by extensive self-play, turning the problem into one of search optimisation. What is more, similar to the non-ToM-human-play Coloured Trails model, models for Chess that are able to capture a human-style of play without using explicit ToM *do* exist (such as 'Maia') [38]. These models rely neither on ToM nor on pure probabilistic reasoning. The major benefit of this system is that it can play chess much more like a human would, at the specific chess ranking (ELO) the human corresponds to. Its abstractions are able to predict the human player's next move in approximately half of the cases for positions when there are multiple 'sensible' options. Effectively, the system is able to capture human-like competitive behaviour in Chess without the need for explicit reflection on the

competitor's mental state, instead relying on pre-trained human data applied to the state space (the game board).

Research has also shown that it is possible to emulate computational ToM by abstracting single beliefs into higher-level concepts [39]. Using the value of pre-established roles, social values, and social norms, can short-cut a decision that would otherwise require advanced reflection with high accuracy, at least in the medical domain. It has similarly been shown that in cooperative counting tasks, the use of a previously established strategies and synergy is sufficient to find good solutions without the need of continued explicit reflection on one's partner, provided the rules of the game or the collaboration partner do not change [52]. In this particular case, establishing this strategy did require a mental model and explicit reflection on one's partner to reach this point – the point is not that ToM is redundant, but that it is no longer necessary after a certain level of familiarity (and perhaps implied trust) has been established.

4.2 Revisiting the Human Perspective

Knowing ToM can occasionally be replaced by different strategies, that use experience or other reasoning short-cuts, gives us a perspective on how the cost [34] and inaccuracy [35] problems of ToM can be dealt with in practice. However, the knowledge that these alternatives exist, does not answer when humans do, and when they do not use ToM, especially not when they use the intensive 'higher-level' ToM so famous for this costliness.

So, how advanced does ToM need to be, both for human-human collaboration, and for human-AI collaboration? To answer this question, it is important to evaluate its value. Revisiting the Coloured Trails negotiation simulation, we are led to conclude that ToM's benefits drop off after three levels of recursion, both in experimental agent-agent settings, but also when an agent negotiates with a human [53][2]. The level that humans realistically take seems to be slightly higher (level-3, level-4, presumably even higher) in tasks that are perceived as less mentally taxing, as indicated by research that solves the Mod-game (a numerical variant of Rock-Paper-Scissors with proximity-based point-scaling) [54] using the same recursive ToM setup [55]. In other words: depending on the task, there is a benefit and realistic usage drop-off, even if this drop-off depends on the complexity of the task.

Looking at situations with more complex pay-off structures and higher mental effort, the same recursive reasoning setup reveals that humans sometimes already have trouble with using second-order reasoning [56]. They can, however, become better in ToM-reasoning through explicit step-wise training. This also holds for higher orders of ToM [57]. A key component to ToM reasoning

[2]In negotiation, the 3rd level of recursive ToM would be: Taking into account that your partner will account for you accounting for them.

seems to be the awareness that ToM-like reasoning is required to begin with, which is something humans still struggle with in adulthood [58]. Experience with the topic helps in these cases [20, 59]. How deeply human ToM reflections go, however, still differs per person [60], and per skill level [4], which is yet another indicator that even in adulthood, one's experience with the situation is key to their problem-solving aptness.

There is a crucial realisation here: Humans often do not apply ToM automatically unless prompted, and experience and familiarity with the topic weigh heavily into how well this ToM is applied once prompted. It would be unwise to overestimate the capacity of a human collaboration partner, both when we are a human, and when we are an AI system. The behaviour from the human is probably better explained using a short-cut strategy, or a relatively low level of ToM (level-one ToM, only reflecting on our partner's thoughts on the situation, meaning level-two would be sufficient from either side in hopes of this leading to a successful synergy).

4.3 Conclusion: Working Smarter

The use of higher ToM levels, going beyond a first-order reflection on one's partner's perspective, will generally require more effort, and is often unnecessary. If the situation is not very dynamic, using too high of a ToM level might result in a non-human-like algorithm, that overestimates a human partner's reasoning methods. In human-human interaction, many collaborations rely on pre-established or predictable patterns. We recommend a similar approach in AI systems: Rely on scripts and heuristics by default, and only defer to active ToM reasoning only when needed.

In short, it is not necessary for *every* ToM system to be 'as good as possible'. Only if we need an AI system that is an expert on the situation, or has to be in an overseer position that specifically works with (human) experts on the domain, we may need higher levels of advanced, explicit, ToM.

5 All ToM Is The Same

Humans have a tendency to somewhat personalise their use of ToM based on their experiences with a person: Knowing someone, or even having a superficial first impression, impacts how we, as humans, think about one another. If a person comes across as quite snobbish, we will have a different impression of them as when they come across as an enthused savant, and our reasoning about these two individuals will be different. In the case of the snob, we may feel like they have met people like this before: the know-it-all, who believes their tastes superior to someone else's (and looks up to very specific individuals). They will know a lot about a specific niche subject, be very expressive, sometimes even aggressive, about their knowledge on that topic, and might exaggerate

that knowledge to a related-but-slightly-broader domain. We might be wrong about these assumptions, and might even be aware that we are biased about this 'snob' based on previous experiences, but we will still rely on the bias quite often (referring back to Section 4.2: Biases and stereotypes are an example of a shortcut or heuristic in the social domain).

These stereotypes are useful and effective, but intuitively, they harm collaboration on a deeper level. We 'know' the snob thinks differently than us about some things (even if we may be wrong about some), and so far this is useful for collaboration. What we do not know, is how the snob really thinks about *us*, as we cannot draw an accurate picture of what biases the snob will have about us (only an approximation) unless they ask and get a truthful answer. Perhaps the snob thinks of us as very interested in what they have to say, estimating their trust in us to be at a far higher than it actually is (after all, the snob is presumably quite arrogant). What is an optimal choice for one individual, is not for the other, especially from a social rather than an economic point of view. Different people have different models of one another, and those models are sometimes biased to a degree that hampers collaboration. We need a 'snob' translation manual to fully do the 'correct' mentalisation. Knowing this, why try to capture ToM in a one-size-fits-all 'generalisable' model? In this section, we elaborate on the tested differences in ToM across entities.

5.1 Human-Human Interaction

Intuitively, it is easy to see why robots, all sorts of animals[3], and humans think differently. And that the snob's mentalisation functions differently than the non-snob. Maybe less intuitively, it needs to be said that ToM differs *between* 'categories' of humans as well, regardless of their specific personality (although this also matters). Across human populations, there are a few factors that greatly influence their 'use' of ToM.

A strong characteristic of the accuracy of one's use of ToM is how similar the other individual is. An example drawn directly from ToM literature in developmental psychology is research into autistic individuals, and the long discussions on the supposed lack of ToM in these individuals. The conclusion on their problems with ToM is heavily flawed, see for example [64]. In what is referred to as the 'Double Empathy problem' [65], it is argued instead that autistic individuals have less trouble understanding (and thus mentalising) other autistic individuals, and that the same holds between allistic/neurotypical individuals, but that both 'categories' sometimes have trouble making predictions about each other's behaviour and mental states and motivations[4].

[3]Whether many types of animal thought can be classified as ToM to begin with is very much an open empirical question, with many shades of doubt [61, 62, 63]

[4]This is in practice a fair bit more nuanced, since neurodivergence is a spectrum, but similarity on the spectrum is presumed to aid with mentalising the other

The 'individual similarity' principle also holds for other groups that are on some level generalisable. For example, culture is of significant influence when it comes to one's mentalisation tendencies. Seminal work by Wellman and Liu has categorised the development of ToM into an order by which children tend learn ToM-related skills [16]: they first learn to understand one may [i] desire different things, then that [ii] one may believe different things, [iii] one may have different knowledge, [iv] meaning one may have false beliefs, which can later be [v] explicitly represented false beliefs, and one can then [vi] have emotions to do with these beliefs after which they [vii] can lie about their emotions.

This order seemed to be pretty clear-cut. However, later research showed this order is different depending on what the local cultural values about some of these concepts are. The Wellman-Liu ordering more strongly applies to cultures that generally value beliefs over knowledge (which can probably be linked to individualism). In more colloquialism-based cultures, that generally value knowledge over beliefs, children instead learn to recognise different ToM-states on knowledge before they learn to recognise different ToM-states on beliefs, also influencing the timetable of false-belief recognition (e.g. Iranian [66, 67] and Chinese children [68, 69]).

The power that culture has over shaping one's mind is especially present during development [70, 71]. The differences in development 'priorities' between colloquialism- and individualism-driven cultures implies that the exposure children have to these ToM-aspects varies significantly. After all, more strongly weighing one aspect of ToM over the other, actively influences the priorities in perspective-taking the culture transfers to its next generation[5].

This is not the only aspect where the culture one has grown up in is influential: It also affects how humans mentalise others on a less ingrained scale. When stories use protagonists and objects familiar to a human's specific culture, that human will more often (and more speedily) make the correct ToM assertion, even if the situation described is not inherently cultural (f.e. a complex false belief task) [72]. Additionally, growing up with input from two or more cultures, and even self-evaluated openness to other cultures, both aid in mind reading people from other cultures in general [73]. Note that this study has also shown that, true to any discussion on specialisation versus generalisation, under the same culture, mono-culturalists seem to be better at mind-reading people from that specific culture than multiculturalists.

Knowing that *on average* similarity and openness to cross-cultural experiences are both beneficial to successful mind-reading efforts provides us with a solution on improving specific human-human interaction: awareness. Spending active

[5]Note that there is no overall difference in the 'ability to understand aspects of ToM to begin with': in the end, children who have enough social (and language) exposure develop each of the listed skills, although they may lean toward mind-reading based on one of the listed skills over the other [69]

and conscious effort into making the humans in a collaboration aware of the aspects that matter to their mind-reading, and asking them to reflect on these matters with the right materials, should nudge both parties in the interaction towards more productive mind-reading.

5.2 Human-AI Interaction

Of course, the concepts that apply to human-human interaction, presumably also apply to some degree in human-AI interaction. What we add on top of this, is that AI systems also lack a more general 'human' experience. So far, what we have seen, is that common ground between entities [36] is a major contributor to how well they understand each other. Humans share a same biology, have the same sensors they use to interact with the world[6], and know that all other humans have this as well. There may be many mental differences, but they are grounded by needing sustenance, and by their beliefs, desires and intentions. The divergence of these grounds starts at a different level (and actively happens through what we call 'grounding' [74, 75].

This is not the story of the AI system. The AI system does not have the same sensors as a human, has a very different 'brain', does not need food or water... and as such misses an intricate ground with humans [76]. It is the designer's task to allow a human and an AI system to bridge this gap. When talking about specific states of an object, of knowledge, of a belief, etc., the human and the AI need to ground whether they are on the same page (sadly more often than a human and a human, as the gap is larger). In specialistic tasks, this is doable, as the domain is limited and the possibility to add knowledge to a pre-defined domain can be added into a system: Things may not be truly human, but the human and AI can both, in their own way, reason about the knowledge/world states that they now both have. For more general AI, a solution lies further on the horizon.

It is very important for the human to realise the difference in ground that they and the system have. As we know from human-animal interaction, it is very easy for humans to attribute ToM to animals, as communicative intention with the entity inherently draws out attribution of an assignment of mental states [77]. This holds especially for pet owners [78], but is also true in general. Exposure to human-like animals in fiction strongly raises the human tendency to apply ToM to animals in real life, because of the 'blueprint' impressions these fictionalised animals leave that are then applied to real-world animals [79]. This is despite evidence to the contrary [7].

Returning to human-AI interaction: This problem is similar when we look at (especially) human-robot interaction. A robot using human speech patterns is

[6] How these sensors make sense of the world can still be very different!

[7] Note that this on its own is an indicator that humans prefer quick strategies over the use of rationalising their way through evidence, in reference to the previous misconception

perceived as 'more human' [80], looking like a human increases a robot's perceived intelligence [81], and a robot expressing human-like attitudes (i.e. express actions one would also evaluate during ToM-like reasoning) is considered more friendly [82]. In practice, all of these aspects are but part of a whole that combines into a concept of anthropomorphisation.

Having established that the differences between a human and an AI system are vastly significant for how they deal with beliefs, desires and intentions, this is a problem. Anthropomorphism can be a harmful bias for human-AI – one that can be worked on to overcome – but it becomes more problematic when it becomes a practice that is actively leveraged. If a system is anthropomorphised on purpose, without any active mechanisms that allow it to genuinely ground with its human interactants, it will result in a lot of ToM misconstrual.

5.3 Conclusion: Finding Common Ground

What do these insights mean for AI & CS developers and researchers? A preliminary advice: Do not fake it. It is not enough to *seem* human-like: The system should not trigger the user to make unfounded assumptions about the AI system's human-likeness, even if the aforementioned anthropomorphism raises trust in the system [83, 84]. Be open about what the system can and cannot do. Try to create common ground with the system where possible, so stay inquisitive: Collect information to adjust the mental map of both the human user and the AI system.

Ultimately, we suggest two solutions. The first solution is that humans need to invest in building up a 'theory of AI mind', i.e., learn how the specific AI system reasons, functions, and acts in the world. Developers need to explain to the user how the specific system maps its beliefs to behaviour, and how its sensors perceive the world. On the flip-side, the system needs to clearly communicate its intentions to the human user and explain how it interprets what the human is doing (again, trying to find common ground). Our shared responsibility as developers is making sure that these interpretations contain a conceptualisation that can be translated to correct actions, to position the AI system as Dennett refers to it, an intentional one [1]. Alternatively, the system needs to be designed in a human-like way, so that humans can apply what they know about interaction with humans to interacting with this AI system. This is, however, a far trickier solution, since it requires a thorough understanding of emulating the human brain.

6 Current AI Systems Already Have ToM

Computational models of ToM have a long tradition in agent-based modelling, including recursive, Bayesian, and neural frameworks [85, 86, 87, 88]. Each of these frameworks has their own individual claim to ToM, but there are differ-

ences how and for what purpose this capacity is modelled. In this section, we discuss these properties, and the misconceptions relating to these properties. Additionally, we elaborate on the perceived ToM behind a specific implementation of a neural framework: ToM with respect to large language models (LLMs).

Recursive ToM-based models are usually rooted in reasoning that is based on utility, epistemic logic, doxastic logic, and combinations thereof. Their deterministic nature makes them quite useful for modelling exact knowledge and belief states, in ways that also enable a model to rationally reason about both the knowledge it has and the knowledge it has yet to acquire. This makes these models perfect for ToM-scenarios that rely on the resolution of mental states to the nth level (ToM level-1, level-2, level-3, etc.) in an explainable, traceable, way, enabling them to take a clearly defined (counter)action based on their interaction partner's perspective (BDIs). These models do require a mechanism that determines the relevance and importance of these BDIs. This mechanism is often represented by economic rationality' the most optimal action from the perspective of their partner (depending on the ToM level), which often does not match with reality. We have already questioned this notion in "Every Social Interaction Requires (Advanced) ToM" (Section 4), but would like to reiterate that humans that use short-cuts and lazy reasoning are often not rational. This makes these systems likely to reason erroneously if applied in real-world social settings, even if it makes them perfect for resource negotiation and strategy simulations (and in a collaborative setting, a far better reasoner when dealing with deliberate experts than when dealing with novices!).

Bayesian models, in turn, are able to better incorporate non-rational actions as they can model hidden behavioural properties when enough data is available. They do only function realistically (with a high performance rate) when the priors are right, leveraging a lot of real-world information that may not be practically available, as information existing in the world, through statistical patterns, does not automatically imply this information is also retrievable by the designer. In terms of realism, although the underlying model might *map* human ToM in a seemingly realistic fashion, we know from experience that humans prefer shortcuts and are not very probabilistically driven, which one needs to take in constant account when modelling their agents in a Bayesian fashion.

We will focus on Machine Learning-based Systems in more detail, due to the attention Large Language Models have drawn with respect to specifically ToM, driving a need to analyse a potential misconception.

6.1 "Machine Learned Systems Have A Theory of Mind"

Recent developments in neural frameworks have seen rapid developments in the human-likeness of AI: Generative pre-trained transformer models, forming the basis for LLMs and large multimodal models (LMMs), have very convincing

linguistic capabilities, especially when looking at their most recent iterations[8]. Their success in several interactive contexts has sparked debate over whether a form of ToM may have emerged in such models. This is not an unreasonable claim, given the intricate ties between ToM and language in human development and evolution [5] – and thus warrants an in-depth look.

LLMs beyond a certain size and level of fine-tuning pass traditional false-belief tests [93, 94], i.e. tests that have been developed to assess ToM competence in human children and specific populations [3, 95]. However, such performance was mostly attributed to ample presence of the benchmarked ToM tests in the training data, meaning that superficial task recognition was sufficient to give the right answer. When tests were adapted to avoid this, LLM performance was shown to drop [96, 97] or in need of nuance [98].

Despite our own scepticism whether performance and standardised tests really indicate ToM abilities, we do feel that general rebuttals sell the story a bit short. LLMs were neither designed nor trained specifically to perform ToM tasks, and we have seen from studies in developmental psychology that some aspects of ToM *can* indeed be an emergent property of language acquisition [23, 99].

Following the initial results and debate, ToM benchmarks were introduced [100, 101, 102], comparisons were made against human (child) scores [98, 94], other modalities were integrated [103, 104], integrations with older model architectures were explored [87], and theoretical reflection was added [105]. Resulting from this literature, we briefly reflect on three additional aspects indicating that the abilities of LLMs should to be taken seriously, but are far from being there.

6.1.1 Correlation with Real-World Social Ability

The observed ability of LLMs to score well on standardised ToM tests may not correlate with real-world social abilities [98]. In principle this is an inherited flaw of how humans are often evaluated in their ability to solve ToM-related problems, as these evaluations, similar to LLMs, rely heavily on linguistical ability [95]. However, for humans, a large body of work associates performance on standardised ToM tests with various landmarks in children's socio-cognitive development [106]. These landmarks go beyond the linguistic abilities that these ToM tests are biased towards, meaning that ToM tests can reasonably be argued to be an effective means of evaluating the human ability to mentalise: No comparable landmarks exist for LLMs.

For *LLMs*, it is thus an open empirical question how well test scores generalise to their social competencies in *actual* interactions with humans. We have no such corresponding socio-cognitive data for LLMs, as LLMs have no 'lived' experience, only knowledge driven by their text- and static image-heavy train-

[8]Examples of commercial, closed-source models are OpenAI's ChatGPT [89, 90] and Google's Gemini [91]; open-source variants are e.g. the OLMo model family [92].

ing corpora [107]. As a consequence, while LLMs *may* be able to generalize beyond statistical pattern prediction, the knowledge of LLMs is very differently (and probably much less firmly) grounded compared to humans's (see also the discussion on Misconception 3, "All ToM is the Same").

6.1.2 Perspective Grounding

There is an important distinction between exposure-based *third-person* and experience-based *first-person* social reasoning. This is not properly captured by traditional benchmarks. Most benchmark ToM tests, also those in lab settings, take an 'observer perspective'. These tests may provide LLMs with an 'unfair' advantage given a training set with ample descriptions of social life from, e.g., literary fiction or online fora with people sharing their experiences in everyday life (all top-down provided information, exposure). However, once more: Answering questions about social situations from an outside perspective differs greatly from actually engaging *in* such situations.

Specific explorations into this topic have shown that, indeed, when an LLM is forced to take individual perspectives through dialogue (i.e., a first-person perspective), its performance sharply declines [100]. In the first-person-dialogue case, models were quite likely to incorporate characters that were unaware of the presented information into their chain of reasoning. Additionally, presenting more than just directly relevant information, involving characters who do not have mental states about the subject, with the systems often unable to identify relevant concepts from irrelevant ones. Needless to say, this differs significantly from the interaction-driven ToM that humans often experience, where keeping track of information relevance and information availability are key to human social skill.

Similarly, Hou et al. found that using first-person perspectives severely hampers LLM performance in even the currently most advanced LLMs, such as GPT 4, Claude 3.5 and Llama 3.8 [108]. In contrast, converting the same (textual) social situation to one presented in a 3rd person, more narrative, style, majorly bumped up the performance of the systems.

6.1.3 Scale over Substance

We know that model characteristics, test type, and test approach influence performance. It is no coincidence that more recent models, fed with more data than ever before, outperform the older models. The claims that ToM performance reach teen levels [109, 110] can sometimes even be refuted in-paper by controlling for the associative reasoning that LLMs have gotten quite good at due to amount of data they are trained on: Forcing a relational (logical) reasoning style instead of an associative one shows major differences in ToM performance (1-2 years vs early teens) [111].

Additionally, fine-tuning and prompting approaches boost scores on ToM-standardised tests [112, 113, 114]. The study by Moghaddam and Honey reveals that ToM tests that LLMs do still struggle with, are significantly better executed when human feedback is added to the system through reinforcement learning. This casts doubt on the idea that the exposure-based techniques that LLMs are driven by, is enough to capture human ToM. In fact, the additional training through human feedback was beneficial, but modern LLMs were still unable to solve the complex second-order ToM tasks at hand.

6.2 Conclusion: Nuance Needed in the Debate over ToM in LLMs

What we learn from LLMs struggling to successfully adopt a first-person ToM perspective is the importance of real-life experience within social settings. This is also our advice for future benchmarks: Evaluate your systems through social settings, *in practice*, rather than through benchmarks adapted from tests developed for humans. This is understandably a big ask, with a potentially high engineering effort, but to us, it does feel like the only way to evaluate the aspects of ToM that are considered to be informed attributes in humans. Expose the system to real-life human behaviour: ToM as a social skill should not be lab-bound, especially knowing how the framing of a setting can already determine whether even a human uses ToM or not, and whether or not this is successful. If we do not evaluate a system's ability to reason beyond limited benchmarks, we may not end up testing how truly socially dynamic such a system is. ToM is a skill that deals with the unexpected, applied in specific settings: Research into emulating ToM for our AI systems should ensure the situation is *actually* unexpected.

7 Discussion and Concluding Remarks

We have defined ToM as the ability to make sense of someone's behaviour and attitudes towards the world by reasoning from their perspective, in terms of their beliefs, desires, intentions, motivation, and so on. AI systems can make estimations about beliefs, desires, and intentions, in specific situations, using specific techniques. Calculated problems, such as a resource negotiation, a false belief puzzle, a prediction on what direction someone will walk into, all benefit from ToM, but are specialised instances of a far bigger whole. Generalisability beyond these instances is challenging for even the most capable AI systems. Yet we have discussed that ToM might not be as generalisable of a skill in humans either, often acting as a function of how experienced a human seems to be with the specifics of the situation, which is linked to individual and socio-cultural differences.

Humans tackle the situation head-on and learn on the go, modelling their collaborative partners and not-so-collaborative competitors in the new and specific

situation as they learn new things about them. Similarly, online and dynamic learning feels like a feasible approach for many AI systems as well, even if this is a demanding process. It is the human ability to learn that creates the apparent flexibility. After all, 'What would I do, if I were in their shoes?' is a valid question when the situation calls for a suspicion of one's motives, but is effortful and might require a manual to translate between cultures or different entities altogether. This manual needs to be either pre-delivered to aid with the translation, or the entities need to be similar enough to ourselves to not require such a translation to begin with. What often leads to misunderstandings or worse, is to simply assume this similarity.

We end by pointing to an alternative approach: the notion of Hybrid Intelligence [115, 116]. Humans and AI Systems each have their separate strengths. While keeping the discussed misconceptions in mind, the strengths of both can be leveraged. AI Systems have a good memory, fast retrieval, can quickly make difficult calculations, and are very good at finding patterns in data. Humans are better at evaluating the pragmatic value of these memories and patterns, resolving ambiguity in communication, and at dealing with social situations and complexity in general. These skills can be made complementary, as long as a cooperative setting is designed for. Revisiting our Healthcare Robot: The AI system driving the robot might be better at detecting when something is off, noticing a (minor) change in behavioural or other patterns. The human side of this duo would be able to collaborate with the robot and work out the nuanced social dynamics of the situation. For example, the robot could determine the relevant questions to ask a patient, whereas the human would be able to formulate them adequately and adapt to the direct responses they evoke. In the end, through such forms of collaboration, humans and AI might even learn something from each other and improve their ToM skills on both ends.

References

[1] D. C. Dennett, "Intentional systems," *The journal of philosophy*, vol. 68, no. 4, pp. 87–106, 1971.

[2] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," *Behavioral and brain sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[3] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind"?," *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

[4] I. A. Apperly and S. A. Butterfill, "Do humans have two systems to track beliefs and belief-like states?," *Psychological Review*, vol. 116, no. 4, p. 953, 2009.

[5] M. J. van Duijn, *The lazy mindreader: a humanities perspective on mindreading and multiple-order intentionality*. PhD thesis, Leiden University, 2016.

[6] D. D. Hutto, *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. The MIT Press, 2008.

[7] I. Apperly, *Mindreaders: the Cognitive Basis of "Theory of Mind"*. Psychology Press, 2010.

[8] A. S. Rao, M. P. Georgeff, *et al.*, "Bdi agents: from theory to practice.," in *Icmas*, vol. 95, pp. 312–319, 1995.

[9] M. Kirtay, E. Oztop, M. Asada, and V. V. Hafner, "Trust me! i am a robot: An affective computational account of scaffolding in robot-robot interaction," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 189–196, IEEE, 2021.

[10] S. Devin and R. Alami, "An implemented theory of mind to improve human-robot shared plans execution," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 319–326, IEEE, 2016.

[11] O. C. Görür, B. S. Rosman, G. Hoffman, and S. Albayrak, "Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention," in *Workshop on Intentions in HRI at ACM/IEEE International Conference on Human-Robot Interaction.*, 2017.

[12] A. M. Leslie, "Tomm, toby, and agency: Core architecture and domain specificity," *Mapping the mind: Domain specificity in cognition and culture*, vol. 29, pp. 119–48, 1994.

[13] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, pp. 13–24, 2002.

[14] M. Ruocco, W. Mou, A. Cangelosi, C. Jay, and D. Zanatto, "Theory of mind improves human's trust in an iterative human-robot game," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, pp. 227–234, 2021.

[15] H. Wimmer and J. Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, pp. 103–128, 1983.

[16] H. M. Wellman and D. Liu, "Scaling of theory-of-mind tasks," *Child development*, vol. 75, no. 2, pp. 523–541, 2004.

[17] J. I. Carpendale and C. Lewis, "Constructing an understanding of mind: The development of children's social understanding within social interaction," *Behavioral and brain sciences*, vol. 27, no. 1, pp. 79–96, 2004.

[18] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[19] D. Kahneman, *Thinking Fast and Slow*. MacMillan, 2011.

[20] D. Samson, I. A. Apperly, J. J. Braithwaite, B. J. Andrews, and S. E. Bodley Scott, "Seeing it their way: Evidence for rapid and involuntary computation of what other people see.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 36, no. 5, p. 1255, 2010.

[21] J. W. Astington and J. M. Jenkins, "A longitudinal study of the relation between language and theory-of-mind development.," *Developmental psychology*, vol. 35, no. 5, p. 1311, 1999.

[22] T. Ruffman, L. Slade, and E. Crowe, "The relation between children's and mothers' mental state language and theory-of-mind understanding," *Child development*, vol. 73, no. 3, pp. 734–751, 2002.

[23] J. G. De Villiers and P. A. de Villiers, "The role of language in theory of mind development," *Topics in Language Disorders*, vol. 34, no. 4, pp. 313–328, 2014.

[24] B. van Dijk and M. van Duijn, "Modelling characters' mental depth in stories told by children aged 4-10," in *Proceedings of the annual meeting of the cognitive science society*, pp. 2384–2391, The Cognitive Science Society, 2021.

[25] L. Slade and T. Ruffman, "How language does (and does not) relate to theory of mind: A longitudinal study of syntax, semantics, working memory and false belief," *British Journal of Developmental Psychology*, vol. 23, no. 1, pp. 117–141, 2005.

[26] W. C. Hall, L. L. Levin, and M. L. Anderson, "Language deprivation syndrome: A possible neurodevelopmental disorder with sociocultural origins," *Social psychiatry and psychiatric epidemiology*, vol. 52, pp. 761–776, 2017.

[27] S. J. Carrington and A. J. Bailey, "Are there theory of mind regions in the brain? a review of the neuroimaging literature," *Human Brain Mapping*, vol. 30, no. 8, pp. 2313–2335, 2009.

[28] M. Schurz, J. Radua, M. G. Tholen, L. Maliske, D. S. Margulies, R. B. Mars, J. Sallet, and P. Kanske, "Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind.," *Psychological Bulletin*, vol. 147, no. 3, p. 293, 2021.

[29] C. E. Mahy, L. J. Moses, and J. H. Pfeifer, "How and where: Theory-of-mind in the brain," *Developmental Cognitive Neuroscience*, vol. 9, pp. 68–81, 2014.

[30] R. C. Schank and R. P. Abelson, *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 2013.

[31] H. Meng, "Social script theory and cross-cultural communication," *Intercultural Communication Studies*, vol. 17, no. 1, pp. 132–138, 2008.

[32] D. Taylor, G. Gönül, C. Alexander, K. Züberbühler, F. Clément, and H.-J. Glock, "Reading minds or reading scripts? de-intellectualising theory of mind," *Biological Reviews*, vol. 98, no. 6, pp. 2028–2048, 2023.

[33] S. Lin, B. Keysar, and N. Epley, "Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention," *Journal of Experimental Social Psychology*, vol. 46, no. 3, pp. 551–556, 2010.

[34] P. A. Lewis, A. Birch, A. Hall, and R. I. M. Dunbar, "Higher order intentionality tasks are cognitively more demanding," *Social Cognitive and Affective Neuroscience*, vol. 12, pp. 1063–1071, 03 2017.

[35] R. Wilson, A. Hruby, D. Perez-Zapata, S. W. van der Kleij, and I. A. Apperly, "Is recursive "mindreading" really an exception to limitations on recursive thinking?," *Journal of Experimental Psychology: General*, 2023.

[36] H. Clark, "Context and common ground," in *Encyclopedia of Language & Linguistics. (Second Edition)* (K. Brown, ed.), vol. 3, pp. 105–108, Elsevier, 2006.

[37] C. O'Grady, C. Kliesch, K. Smith, and T. C. Scott-Phillips, "The ease and extent of recursive mindreading, across implicit and explicit tasks," *Evolution and Human Behavior*, vol. 36, no. 4, pp. 313–322, 2015.

[38] R. McIlroy-Young, S. Sen, J. Kleinberg, and A. Anderson, "Aligning superhuman AI with human behavior: Chess as a model system," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1677–1687, 2020.

[39] E. Erdogan, F. Dignum, R. Verbrugge, and P. Yolum, "Abstracting minds: Computational theory of mind for human-agent collaboration," in *HHAI2022: Augmenting Human Intellect*, pp. 199–211, IOS Press, 2022.

[40] V. E. Stone, "Theory of mind and the evolution of social intelligence," *Social neuroscience: People thinking about thinking people*, pp. 103–129, 2006.

[41] M. Tomasello and A. Vaish, "Origins of human cooperation and morality," *Annual review of psychology*, vol. 64, no. 1, pp. 231–255, 2013.

[42] R. W. Byrne, "Machiavellian intelligence," *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, vol. 5, no. 5, pp. 172–180, 1996.

[43] R. I. Dunbar and S. Shultz, "Understanding primate brain evolution," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1480, pp. 649–658, 2007.

[44] T. Lenaerts, M. Saponara, J. M. Pacheco, and F. C. Santos, "Evolution of a theory of mind," *Iscience*, vol. 27, no. 2, 2024.

[45] H. De Weerd, R. Verbrugge, and B. Verheij, "Higher-order theory of mind is especially useful in unpredictable negotiations," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 2, p. 30, 2022.

[46] R. Verbrugge, "Logic and social cognition: The facts matter, and so do computational models," *Journal of Philosophical Science*, vol. 38, no. 6, pp. 649–680, 2009.

[47] S. G. Ficici and A. Pfeffer, "Modeling how humans reason about others with partial information," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 315–322, 2008.

[48] H. de Weerd, R. Verbrugge, and B. Verheij, "The effectiveness of higher-order theory of mind in negotiations," in *Reasoning About Other Minds: Logical and Cognitive Perspectives: RAOM 2014*, pp. 35–39, Citeseer, 2014.

[49] M. Campbell, A. J. Hoane Jr, and F.-h. Hsu, "Deep blue," *Artificial intelligence*, vol. 134, no. 1-2, pp. 57–83, 2002.

[50] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[51] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[52] R. van der Meulen, R. Verbrugge, and M. van Duijn, "Common ground provides a mental shortcut in agent-agent interaction," in *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pp. 281–290, IOS Press, 2024.

[53] H. de Weerd, R. Verbrugge, and B. Verheij, "Agent-based models for higher-order theory of mind," in *Advances in Social Simulation: Proceedings of the 9th Conference of the European Social Simulation Association*, pp. 213–224, Springer, 2014.

[54] S. Frey and R. L. Goldstone, "Cyclic game dynamics driven by iterated reasoning," *PloS one*, vol. 8, no. 2, p. e56416, 2013.

[55] K. Veltman, H. de Weerd, and R. Verbrugge, "Training the use of theory of mind using artificial agents," *Journal on multimodal user interfaces*, vol. 13, pp. 3–18, 2019.

[56] R. Verbrugge, B. Meijering, S. Wierda, H. van Rijn, and N. Taatgen, "Stepwise training supports strategic second-order theory of mind in turn-taking games," *Judgment and Decision Making*, vol. 13, no. 1, pp. 79–98, 2018.

[57] A. Valle, D. Massaro, I. Castelli, and A. Marchetti, "Theory of mind development in adolescence and early adulthood: The growing complexity of recursive thinking ability," *Europe's Journal of Psychology*, vol. 11, no. 1, p. 112, 2015.

[58] B. Keysar, S. Lin, and D. J. Barr, "Limits on theory of mind use in adults," *Cognition*, vol. 89, no. 1, pp. 25–41, 2003.

[59] I. Apperly, "Can theory of mind grow up? mindreading in adults, and its implications for the development and neuroscience of mindreading," *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*, pp. 72–92, 2013.

[60] J. Stiller and R. I. Dunbar, "Perspective-taking and memory capacity predict social network size," *Social Networks*, vol. 29, no. 1, pp. 93–104, 2007.

[61] J. Call and M. Tomasello, "Does the chimpanzee have a theory of mind? 30 years later," *Trends in cognitive sciences*, vol. 12, no. 5, pp. 187–192, 2008.

[62] E. Van Der Vaart, R. Verbrugge, and C. K. Hemelrijk, "Corvid re-caching without 'theory of mind': A model," *PloS one*, vol. 7, no. 3, p. e32904, 2012.

[63] L. Barrett and B. Würsig, "Why dolphins are not aquatic apes," *Animal Behavior and Cognition*, vol. 1, no. 1, pp. 1–18, 2014.

[64] M. A. Gernsbacher and M. Yergeau, "Empirical failures of the claim that autistic people lack a theory of mind.," *Archives of scientific psychology*, vol. 7, no. 1, p. 102, 2019.

[65] D. E. Milton, "On the ontological status of autism: The 'double empathy problem'," *Disability & society*, vol. 27, no. 6, pp. 883–887, 2012.

[66] A. Shahaeian, C. C. Peterson, V. Slaughter, and H. M. Wellman, "Culture and the sequence of steps in theory of mind development.," *Developmental Psychology*, vol. 47, no. 5, p. 1239, 2011.

[67] A. Shahaeian, M. Nielsen, C. C. Peterson, and V. Slaughter, "Iranian mothers' disciplinary strategies and theory of mind in children: A focus on belief understanding," *Journal of Cross-Cultural Psychology*, vol. 45, no. 7, pp. 1110–1123, 2014.

[68] H. M. Wellman, F. Fang, D. Liu, L. Zhu, and G. Liu, "Scaling of theory-of-mind understandings in chinese children," *Psychological science*, vol. 17, no. 12, pp. 1075–1081, 2006.

[69] D. Liu, H. M. Wellman, T. Tardif, and M. A. Sabbagh, "Theory of mind development in chinese children: a meta-analysis of false-belief understanding across cultures and languages.," *Developmental psychology*, vol. 44, no. 2, p. 523, 2008.

[70] L. S. Vygotskij and V. John-Steiner, *Mind in society: The development of higher psychological processes*. Harvard University Press, 1979.

[71] M. Tomasello and H. Moll, "The gap is social: Human shared intentionality and culture," in *Mind the Gap: Tracing the Origins of Human Universals*, pp. 331–349, Springer, 2010.

[72] D. Perez-Zapata, V. Slaughter, and J. D. Henry, "Cultural effects on mindreading," *Cognition*, vol. 146, pp. 410–414, 2016.

[73] L. R. Kim, J. Jetten, A. Pekerti, and V. Slaughter, "Mindreading across cultural boundaries," *International Journal of Intercultural Relations*, vol. 93, p. 101775, 2023.

[74] E. V. Clark, "Common ground," *The handbook of language emergence*, pp. 328–353, 2015.

[75] B. Geurts, "Convention and common ground," *Mind & Language*, vol. 33, no. 2, pp. 115–129, 2018.

[76] 2018. Essay: What is it like to be a robot?

[77] G. Airenti, "The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind," *Frontiers in psychology*, vol. 9, p. 2136, 2018.

[78] T. J. Eddy, G. G. Gallup Jr, and D. J. Povinelli, "Attribution of cognitive states to animals: Anthropomorphism in comparative perspective," *Journal of Social issues*, vol. 49, no. 1, pp. 87–101, 1993.

[79] C. Grasso, C. Lenzi, S. Speiran, and F. Pirrone, "Anthropomorphized nonhuman animals in mass media and their influence on human attitudes toward wildlife," *society & animals*, vol. 31, no. 2, pp. 196–220, 2020.

[80] F. Eyssel, D. Kuchenbrandt, S. Bobinger, L. De Ruiter, and F. Hegel, "'if you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 125–126, 2012.

[81] J. Fink, "Anthropomorphism and human likeness in the design of robots and human-robot interaction," in *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings 4*, pp. 199–208, Springer, 2012.

[82] E. P. Bernier and B. Scassellati, "The similarity-attraction effect in human-robot interaction," in *2010 IEEE 9th International Conference on Development and Learning*, pp. 286–290, 2010.

[83] Q. Wang, K. Saha, E. Gregori, D. Joyner, and A. Goel, "Towards mutual theory of mind in human-AI interaction: How language reflects what students perceive about a virtual teaching assistant," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.

[84] A. Placani, "Anthropomorphism in ai: Hype and fallacy," *AI and Ethics*, vol. 4, pp. 1–8, 2024.

[85] H. De Weerd, R. Verbrugge, and B. Verheij, "Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information," *Autonomous Agents and Multi-Agent Systems*, vol. 31, pp. 250–287, 2017.

[86] C. Baker and R. Saxe, "Bayesian theory of mind: Modeling joint belief-desire attribution," *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*, 01 2011.

[87] C. Jin, Y. Wu, J. Cao, J. Xiang, Y.-L. Kuo, Z. Hu, T. Ullman, A. Torralba, J. B. Tenenbaum, and T. Shu, "Mmtom-qa: Multimodal theory of mind question answering," *arXiv preprint arXiv:2401.08743*, 2024.

[88] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine theory of mind," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227, PMLR, 10–15 Jul 2018.

[89] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[90] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[91] G. Team, "Gemini: A family of highly capable multimodal models," 2024.

[92] D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. H. Jha, H. Ivison, I. Magnusson, Y. Wang, S. Arora, D. Atkinson, R. Authur, K. R. Chandu, A. Cohan, J. Dumas, Y. Elazar, Y. Gu, J. Hessel, T. Khot, W. Merrill, J. Morrison, N. Muennighoff, A. Naik, C. Nam, M. E. Peters, V. Pyatkin, A. Ravichander, D. Schwenk, S. Shah, W. Smith, E. Strubell, N. Subramani, M. Wortsman, P. Dasigi, N. Lambert, K. Richardson, L. Zettlemoyer, J. Dodge, K. Lo, L. Soldaini, N. A. Smith, and H. Hajishirzi, "Olmo: Accelerating the science of language models," 2024.

[93] M. Kosinski, "Evaluating large language models in theory of mind tasks," *Proceedings of the National Academy of Sciences*, vol. 121, no. 45, p. e2405460121, 2024.

[94] J. W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi, M. S. Graziano, and C. Becchio, "Testing theory of mind in large language models and humans," *Nature Human Behaviour*, 2024.

[95] P. Barone, G. Corradi, and A. Gomila, "Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis," *Infant Behavior and Development*, vol. 57, p. 101350, 2019.

[96] T. Ullman, "Large language models fail on trivial alterations to theory-of-mind tasks," *arXiv preprint arXiv:2302.08399*, 2023.

[97] N. Shapira, M. Levy, S. H. Alavi, X. Zhou, Y. Choi, Y. Goldberg, M. Sap, and V. Shwartz, "Clever hans or neural theory of mind? stress testing social reasoning in large language models," *arXiv preprint arXiv:2305.14763*, 2023.

[98] M. van Duijn, B. van Dijk, T. Kouwenhoven, W. de Valk, M. Spruit, and P. van der Putten, "Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests," in *CoNLL* (J. Jiang, D. Reitter, and S. Deng, eds.), (Singapore), pp. 389–402, Association for Computational Linguistics, Dec. 2023.

[99] K. Milligan, J. W. Astington, and L. A. Dack, "Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding," *Child development*, vol. 78, no. 2, pp. 622–646, 2007.

[100] H. Kim, M. Sclar, X. Zhou, R. Bras, G. Kim, Y. Choi, and M. Sap, "FANToM: A benchmark for stress-testing machine theory of mind in interactions," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 14397–14413, Association for Computational Linguistics, Dec. 2023.

[101] Z. Chen, J. Wu, J. Zhou, B. Wen, G. Bi, G. Jiang, Y. Cao, M. Hu, Y. Lai, Z. Xiong, and M. Huang, "ToMBench: Benchmarking Theory of Mind in large language models," 2024.

[102] H. Wang, X. Feng, L. Li, Z. Qin, D. Sui, and L. Kong, "Tmgbench: A systematic game benchmark for evaluating strategic reasoning abilities of llms," 2024.

[103] R. van Berkel, "Large multimodal models and theory of mind." Bachelor's Thesis, LIACS, Leiden University, the Netherlands, 2024.

[104] J. W. A. Strachan, O. Pansardi, E. Scaliti, M. Celotto, K. Saxena, C. Yi, F. Manzi, A. Rufo, G. Manzi, M. S. A. Graziano, S. Panzeri, and C. Becchio, "Gpt-4o reads the mind in the eyes," 2024.

[105] S. Goldstein and B. A. Levinstein, "Does chatgpt have a mind?," 2024.

[106] C. Beaudoin, É. Leblanc, C. Gagner, and M. H. Beauchamp, "Systematic review and inventory of theory of mind measures for young children," *Frontiers in psychology*, vol. 10, p. 2905, 2020.

[107] B. M. A. van Dijk, T. Kouwenhoven, M. R. Spruit, and M. J. van Duijn, "Large language models: The need for nuance in current debates and a pragmatic perspective on understanding," 2023.

[108] G. Hou, W. Zhang, Y. Shen, Z. Tan, S. Shen, and W. Lu, "Entering real social world! benchmarking the theory of mind and socialization capabilities of llms from a first-person perspective," 2024.

[109] T. Webb, K. J. Holyoak, and H. Lu, "Emergent analogical reasoning in large language models," *Nature Human Behaviour*, vol. 7, no. 9, pp. 1526–1541, 2023.

[110] W. Street, J. O. Siy, G. Keeling, A. Baranes, B. Barnett, M. McKibben, T. Kanyere, A. Lentz, R. I. Dunbar, *et al.*, "Llms achieve adult human performance on higher-order theory of mind tasks," *arXiv preprint arXiv:2405.18870*, 2024.

[111] C. E. Stevenson, M. ter Veen, R. Choenni, H. L. van der Maas, and E. Shutova, "Do large language models solve verbal analogies like children do?," *arXiv preprint arXiv:2310.20384*, 2023.

[112] X. Ma, L. Gao, and Q. Xu, "ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind," in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (J. Jiang, D. Reitter, and S. Deng, eds.), (Singapore), pp. 15–26, Association for Computational Linguistics, Dec. 2023.

[113] S. R. Moghaddam and C. J. Honey, "Boosting theory-of-mind performance in large language models via prompting," 2023.

[114] X. A. Huang, E. L. Malfa, S. Marro, A. Asperti, A. Cohn, and M. Wooldridge, "A notion of complexity for theory of mind via discrete world models," 2024.

[115] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister, "Hybrid intelligence," *Business & Information Systems Engineering*, vol. 61, no. 5, pp. 637–643, 2019.

[116] Z. Akata, D. Balliet, M. De Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, *et al.*, "A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence," *Computer*, vol. 53, no. 8, pp. 18–28, 2020.