The Need for a Common Ground: Ontological Guidelines for a Mutual Human-AI Theory of Mind

Marvin Pafla
mpafla@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Mark Hancock
mark.hancock@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Kate Larson
kate.larson@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Jesse Hoey
jhoey@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

ABSTRACT

While traditional accounts of Theory of Mind (ToM) like theory theory (TT) and simulation theory (ST) have highlighted the epistemological challenge in mindreading, perception theory (PT) has posited that we directly perceive the mental states of others. In this paper, we highlight the ontology of these ToM accounts and reject metaphysical realism (i.e., objects exist independent of thought), thereby resolving their apparent opposition. By relying on Smith's account of a participatory metaphysics [34], we argue for the need of a theory of common ground that pays deference to both realistic and constructionist elements of Gallagher's smart perception [11]. In doing so, we argue for a ToM that includes negotiation of reference between agents, emergence of ToM from social interactions rather than pre-definition, a focus on non-conceptual intentions, and flexible representations of objects, thereby questioning the notion of ground truth in AI. Finally, we propose a collaboration between AI and HCI to provide such a mutual human-AI ToM.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; **Collaborative interaction**; Social navigation; • **Computing methodologies → Multi-agent systems**; *Ontology engineering*; Simulation theory; • **Theory of computation** → *Semantics and reasoning*.

KEYWORDS

theory of mind, mutual theory of mind, ontology, metaphysics, registration of objects, perception theory, simulation theory, theory theory

ACM Reference Format:
Marvin Pafla, Mark Hancock, Kate Larson, and Jesse Hoey. 2024. The Need for a Common Ground: Ontological Guidelines for a Mutual Human-AI Theory of Mind. In *Proceedings of Workshop on Theory of Mind in Human-AI Interaction at CHI 2024 (ToMinHAI at CHI 2024).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

With the advent of deep learning and large language models (LLMs) [1, 21], many exciting language applications like text summarization are now feasible. Recent language models that consist of billions of parameters have even been attributed with reasoning and planning skills, including emergent Theory of Mind (ToM) abilities [5, 16, 35]. These abilities help us understand the mental states of others and their behavior [13]. However, studies have also questioned these capabilities of LLMs, stating that they might be an illusion [36]. For example, Verma et al. [36] perform variations of the false-belief test and find that trivial or irrelevant perturbations easily break the illusion of ToM for LLMs.

Given these models' brittleness, studying their contents and what they attempt to represent might be sensible. In the field of AI particularly, but also in the field of HCI, we have seen a strong commitment to metaphysical realism which states that objects exist independent of thought [34]. This philosophical approach has framed the question of a ToM as an epistemological one [18]: the mental states of others exist but are unknown to us which is why we need additional tools like a theory or a simulation to infer what someone else is feeling or thinking. However, perception theory (PT) has started to question this assumption by stating that we directly perceive the mental states of others [11], but has not provided a clear account on how this directness of perception is achieved. Overall, we might ask how the individual and the context she is embedded in impacts the construction of a ToM and its contents.

This paper attempts to explain how embodied agents seem to perceive and construct mental states through interactions with others and their environment, arguing that the contents of such states can only be found within these interactions, never outside them (e.g., somewhere 'in reality'). Hence, we make an ontological argument about a ToM. To illustrate this argument, imagine a smartphone and the fact that there are people on this planet who probably have never seen a smartphone. How would such people perceive such an object? Overgaard [24] argues that though this smartphone might be placed directly in front of them, and thus be visible, its "meaning" or "type" might not be visible in the same way.

Consider an even more basic object like a hammer. In fact, the word "hammer" might already trigger a plethora of associations and meanings such as its affordance, i.e., the activity this object is designed *for*, its principal components and how it feels like to use such a hammer. But let's imagine this hammer is right in front of us when we try to describe it. How would we do this? We might try to describe its components such as its wooden handle or its shining metallic head. In doing so, one might realize that such descriptions can feel arbitrary; for example, even if it had a handle made of hardened plastic, we would still call the object a hammer. Furthermore, we might leave out certain details of the hammer such as the scratch along its head that was caused when impatiently missing the last nail necessary to finish the wooden deck last spring.

With these examples we attempt to show that the act of registering simple objects like a hammer is already an *abstraction* from reality that requires a subject (in this case, you) that does the abstracting. Hence, when wondering about the origin of the object [34] the interaction between subject and object, between perceiver and perceived, is important. That is not to say that the hammer described above is not real; it is very much "made out of stuff", and we can touch it, and so on. However, this perspective highlights the human element in constructing the object in the first place. We can develop this even further: the hammer in front of you is not only achieved in the abstraction of what appears to be a hammer in reality but also is the result of the social interactions that designed and produced the hammer in the first place. For example, the accidental use of a hammer-like object long ago could have sparked the idea of a hammer, leading to the profession of blacksmiths making hammers, and so on.

In this paper, we draw from Smith's account of metaphysics [34] that reconciles both realism and constructionism and reject metaphysical realism by showing the ontological challenges inherent in not only the perception of the mental states of others but also the registration of everyday objects. We highlight the fact that the abstractions we make are mediated by other people and the social context we are embedded in. Thus, we align with perception theory (PT) [11] in that we directly perceive others' mental states, particularly emotions, which we 'see' in their faces. However, while PT moves away from the epistemological question of what the mental states of others are, and posits a ToM in the enacted, embodied interaction with others, it does not fully embrace the ontological groundwork our brain has already done when we simply perceive what is given, describing first and foremost a phenomenological account.

The framing question, i.e., the problem of how the brain constructs the subjective direct experience of mental states given context, then becomes the holy grail of ToM. In the words of Gallagher [11], we might ask what actually makes our perception "smart". While we attempt to lay the ontological groundwork for it, we do not answer this question in this paper. Interested practitioners might refer to active inference approaches that minimize free energy [25] or prediction errors [15] as a starting point, though these approaches might have to be supplemented with an ontological account of emergence.

We first present the three main accounts of ToM in our related work section, before describing their ontological challenges. Given these challenges, we make four guidelines for a ToM based in a participatory metaphysics: we (1) argue against an individualistic ToM and advocate for a negotiation of reference between multiple agents, (2) warn against ontologically defining a ToM a priori instead of letting it emerge, (3) encourage to focus on non-conceptual intentions embedded in a common ground before building mental models of language, and (4) question the assumption of ground truth in AI and argue for flexible representations of objects. Finally, with these guidelines, we argue that the idea of a mutual Theory of Mind between humans and AI is not only a perspective on or a field of ToM, but is, given the socio-contextual interactions we are embedded in, the *only* place to look for a ToM of AI (and humans). We present the philosophical assumption of agent agency for both AI and HCI and argue that a collaboration of the two fields can develop a theory of common ground underlying ToM that includes shared and negotiated references and meanings within a multi-agent context.

2 RELATED RESEARCH

Theory theory (TT) states that humans develop a theory of mind to understand the mental states of others [13]. This theory is developed in a science-like fashion in which lawlike generalizations are produced to link observable inputs, mental states, and output behaviors. In developmental psychology, children are seen as "little scientists", who form assumptions, collect evidence, and revise their theories not just about physical phenomena but also about unobservable mental states [13]. When they start to pass the false belief test of Wimmer and Perner [37] at around the age of four, TT posits the overcoming of a "conceptual deficit" [26], in which children learn that beliefs can be false, gradually developing a more sophisticated theory of mind. After gaining popularity in the 1990s, TT has come under pressure for evidence that showed inhibitory control as a confounding variable in the false-belief test [6], highlighting children's ability to understand false beliefs as early as 15 months, and thus questioning the need to possess a theory to understand the mental states of others.

Simulation theory (ST) states that humans imagine themselves to be in the situation of the other when understanding their mental states [13]. The core idea behind this form of mindreading is the attempt to create proxy or surrogate mental states that can be projected onto the other [13]. ST is built on the ideas of European hermeneutic tradition that highlight the process of "feeling with" others, "reexperiencing", and "putting oneself into their shoes" when trying to make sense of others [13]. The discovery of mirror neurons, that fire not only when an action is performed but also when this action is performed by another person, has been connected to ST as a neural basis for simulation [12]. However, it is unclear whether the re-experiencing of an intention equals the attribution of an intention and whether mindreading occurs as an upshot of mirroring [13]. ST has been criticized for collapsing into TT because knowledge and theory are needed to simulate (e.g., a basic understanding of the context in which we simluate) [10].

Both standard TT and ST frame the problem of mindreading as a problem of access to mental states [11]. People have only access to their own mental states via introspection, but the mental states of others remain hidden [11]. To unlock these "hidden minds" [24], extra-perceptual cognitive processes are necessary that involve theoretical inferences (TT) or simulations (ST). This Cartesian perspective reduces perception to third-person observations where we stand at the margins of a situation, disembodied, without the ability to interpret behavior unless we call forth some theory or run a simulation routine [11]. As this perception is "not-so-smart" and must be supplemented [11], both TT and ST do not develop a proper theory of perception.

Perception theory (PT), or direct perception, in contrast to TT and ST, is built on a "smart" perception that does not need to be supplemented by inferential mechanisms, at least on the personal level [11]. Therefore, perception is direct, e.g., objects are recognized as objects without the need to infer what those objects are given the image of them [11]. This directness of perception, which delivers sufficient information to understand others, makes PT a phenomenological account that is supported by the writings of Wittgenstein and Merleau Ponty: "Grief, one would like to say, is personified in the face" [38, § 570, cf. Z, § 225]. Gallagher [11] further explains that enabling such direct experience of mental states involves complex sub-personal processes, including mirror resonance mechanisms that identify others' intentions. Despite acknowledging these underlying processes, PT faces criticism for promoting the "myth of the given mind" [24], challenging the notion that inferential (TT) or simulationist (ST) models are irrelevant due to the direct nature of perception.

On the other hand, defenders of a traditional ToM have been accused of "promiscuity", indiscriminately applying to term "theory" to sub-personal processes when trying to explain the phenomenological account of direct perception [32]. At the core of this accusation lies the question of what constitutes a theory: does it require a reflective consciousness, or is it enough to describe the structural, functional, and dynamic aspects of a theory of mind [32]? While Slors [32] argues that sub-personal processes can themselves describe a ToM, he also maintains that the frame problem — the challenge of determining which cues in social interactions are relevant for understanding others — remains an unresolved issue. To explain why ToM is ubiquitous despite phenomenological arguments, Slors [32] advances the "model-model" of ToM, proposing that ToM is not only a social-cognitive mechanism, but also a model to explain and make sense of the social cognitive processes we engage in, even though this model might not reflect the actual underlying cognitive processes engaged during social interaction.

In the following section, we argue for a different ontological account of a ToM that rejects metaphysical realism. In doing so, we build on PT and argue for the need of a theory of common ground on which a ToM can rest, including higher-functioning capabilities like theoretical inferences (TT) and simulations (ST), thereby mitigating the inherent tension between these different accounts of ToM.

3 THE ONTOLOGY OF A THEORY OF MIND

Regardless of Slors [32]'s application as a model to explain, the ubiquity of ToM might also be result of the philosophical practices that surround it. Slors [32] finds a Kuhnian paradigm in which it is very hard to think outside of a ToM and its assumptions (e.g., third-person observation) because of perpetual reinforcement of philosophical practices and discussions. Nonetheless, it might be worth questioning the ontological foundations of such a theory, especially when ToM is mistaken for "the real thing" when it is used as a model to explain our assessment of behavior rather than directly describing underlying socio-cognitive processes [32]. Otherwise, we might run into what Smith [34] calls the ontological wall, i.e., the limitations faced by one's own parsing of a theoretical situation into objects, properties, relationships, etc. While Smith [34] admits that some of this parsing is always necessary in advance, the danger of inscription errors is high, where a set of ontological assumptions are inscribed or imposed onto a system, and then read back off the system "as if that constituted and independent empirical discovery or theoretical result" [p.50].

With regards to a theory of mind, Slors [32] points out exactly this danger when the ubiquity of ToM is mistaken for the ubiquity of the talk around ToM and its application as a model to explain behavior. However, the ontological challenges might not stop there. Citing Heidegger's term "being-with", Kiverstein [18] makes the argument that through feelings like empathy we experience ourselves in relationship to others and the world. Taking this perspective of "being-with" shifts the focus onto the ontological question of who and what we are in the first place and the interaction and embodiment through which we attempt to answer these questions [18]. Thus, it stands in contrast to the traditional ToM view which posits the challenge of mindreading to be an epistemological one: observers not knowing what the mental states of others are.

Through this shift in focus the importance of an appropriate metaphysics is highlighted. In line with the epistemological argument above, in the field of ToM and AI in general, we find a persistent metaphysical realism, stating that objects exist independent of thought with only one "correct" representation [34]. This assumptions presents a world with objects "ready for us to be perceived" [22]. Consequently, we have built "perception modules" whose job it is to find the right descriptions of real objects given some input (e.g., images), which in turn are provided to "reasoning modules" that take these descriptions and transform them to solve certain problems [22]. Because representations can be directly learned from whatever data is available within a certain benchmark task, there is no need to embody agents in an environment.

In the Origin of Objects, Smith [34] challenges the assumption of metaphysical realism and argues that humans *register* objects in a participatory process in which reality, that is assumed to be whole and entire, is "gradually but only partially broken down or separated or articulated into objects, through complex and partially disconnected practices of registration" [34, p.269]. He thinks that the world does not arise out of objects, but rather that objects arise out of (One) world *in relationship to and under the effort* of the subject [34]. To then achieve the registration of objects by participatory subjects requires said subjects to be embodied in the very world they try to register [34]. This embodied and enactive metaphysical account is very much in line with, or even might form the basis of, the phenomenological account of direct perception. Yet despite this closeness, the field of ToM implicitly subscribes to a form of metaphysical realism as highlighted by the "objective thought" [23] that is approximated, but never reached, by the pre-objective maps of meaning of the phenomenological account [24].

By adopting a different ontological account that disposes of metaphysical realism and highlights the human element in the construction of everyday objects, PT has not much to lose. Its phenomenological account still allows us to investigate the enactive and emotional content of perceptual experience [11]. It is also correct in so far that the perception of mental states of others is usually direct. However, by adopting Smith's metaphysical account [34], the dichotomy between TT & ST and PT becomes resolved: TT and ST describe post-perceptional accounts of a ToM. Gallagher [11] is correct that we might rely on theoretical inferences and simulations to make sense of things we can not directly perceive. However, this sense-making already happens in an post-perceptional space consisting of mental states and their contents, won by perception, that are necessary for inferences and simulations. PT, in contrast, applies earlier in the perception process and argues for a smart perception that allows us to directly perceive mental states. In other words, TT and ST rely on an ontological parsing of environments, mental states, and so forth, whereas PT describes the phenomenological aspects of this parsing. With this ontological differentiation between perception and post-perception, the tension between the different theories of ToM gets resolved.

In this paper, we attempt to add to PT by arguing for a need of a theory of common ground that can describe how the mental states that we directly perceive come into being. In the following section, we provide guidelines for such a theory.

4 ONTOLOGICAL GUIDELINES FOR A THEORY OF MIND

The exploration of pre-objective maps of meaning emerges as a fascinating area, particularly when considering the consensus within perception theory. Gallagher [11] posits that our ability to directly perceive mental states is rooted in our embodied nature, shaped through interactions with others, past situational experiences, and the assimilation of cultural norms and practices. This notion suggests that our perceptual processes are refined through experience. Such a perspective aligns with the principles of enactivism, central to the phenomenological approach to direct perception, and finds support in Smith's metaphysical framework [34]. Smith [34] attempts to bridge the gap between realism and constructionism, positing that agents are engaged in an ongoing endeavor to align their conceptual system to the "in-part differently conceptualized, and in-part unconceptualized, world" [p.110]. This effort underscores not only our attempts to comprehend the world and other agents within a ToM framework but also illuminates the fact that the fundamental elements of this framework — its concepts, representations of objects, mental states, behaviors, etc. — are shaped by the social contexts in which we are enmeshed.

From this participatory metaphysics, four things follow. First, the location of a ToM must include the population level. This conclusion already follows from perception theory: the directness of perception is enabled through socio-cultural processes that allow the individual to experience herself in the context of social interactions [11]. The social impact on our perception of things, objects, places, etc., requires an abandoning of ToM as an explanation for individual behavior. Rather, direct perception could be supplemented with a proper account of a theory of common ground (see Dafoe et al. [8]) in which we find, create and share references to the world we live in, and weigh the relevance of references when determining the essence of objects and mental states, addressing the frame problem.

This theory could follow Smith's pluralism that posits that "there is no core of stable opinions and meaning but an active, political, violent, feisty negotiation of reference" [34, p.108]. Similarly, we might have to see everyday-objects and their affordances, at least partly, as the social processes that designed and produced them in the first place [34]. To advance the investigation of a mutual ToM between AI and humans, a framing of mutual ToM as a multi-agent problem, in which groups of agents (both AI and humans) try to establish reference, could be a starting point to establish and investigate the diversity of individual agents and the synchronization effects of socio-cultural processes. An excellent playground for such an investigation could be cooperative games [9] like Hanabi [3] that requires agents to manipulate physical artefacts, coordinate play, establish roles, and negotiate rules [31].

Second, as a consequence of the shift in focus on the population level, the environment in which AI is embedded in becomes as important as the AI's architecture. Traditionally, game-theoretic and reinforcement-learning approaches have modelled behavior by mapping pre-defined signals (e.g., game states) with agent actions. Much of the success of second-wave AI and the proliferation of large neural nets can be seen as an attempt to find these associations between outcomes and rewards, and signals and behavior [33]. However, in pre-defining the space of possible behaviors, signals, and outcomes we encounter the same ontological issues described above. With regards to communication, Scott-Phillips [27] argues that we need to account for the fact that communication might *not* occur instead of prescribing it. He posits communicative behavior to be an emergent property that separates itself from non-communicative behavior and that communication is the pragmatic expression and recognition of intentions. Hence, the field of AI and HCI should be careful to not ontologically pre-define a ToM but think about environments, simulations, and games in which it can emerge in social interactions with humans.

The field of AI has worked on emergent communication, particularly in the context of multi-agent reinforcement learning (e.g., [14, 19, 20]). For example, these works have found that language with compositional elements can emerge in the interaction between agents [19, 20]. Nonetheless, this line of work, and the field of AI in general, has been criticized for following what Scott-Phillips [27] calls the code model of communication: an information-theoretic approach to communication [29] that *associates* signals with game states and posits that signals contain encoded messages that are sent along some pre-determined channel to be received and decoded. However, following this approach, Scott-Phillips [27] points out the underdeterminancy of signals in which literal utterances cannot ever fully convey speaker meaning and instead argues for a pragmatic account of communication including the expression and recognition of *intentions*. Thus, Scott-Phillips [27] makes a strong argument for the embodiment of agents in an environment and questions the emergence of communication in AI research.

To study this embodiment, Scott-Phillips et al. [28] designed the embodied cognition game in which an agent had to communicate with another agent by appropriating non-communicative behavior. For example, agents that were able to use movement on a grid as a way to communicate intentions were able to coordinate their behavior and succeed at the game. This form of communication was not encoded into the rules and actions of the game, it emerged in the interaction of different agents. Scott-Phillips et al. [28] studied humans play the game, but it might be interesting to further learn which communication strategies humans use and allow AI agents to learn them without pre-scribing them or encoding them into the architecture of agents. This effort would require AI and HCI to work together, as sketched out in the last section of this paper, by building architectures that can recognize intentions in interactive games with humans.

Third, intentions then seem to be at the core of a theory of mind. While in the field of psychology we refer to the goal-directedness of intentions (e.g., a signaler's informative intention that the receiver change their representation of the world), intentions are also mental states that preclude a semantic directness, a pointing of representations, experiences, and thoughts of subjects towards the world that surrounds them [34]. When investigating the social cognition of infants, they seem to possess a non-conceptual, non-mentalistic embodied understanding of intentions, especially that of caretakers, that seem to precede a conceptual interpretation of mental states as well as language adaptations necessary to pass the classical false belief test [11]. And yet, "no one has successfully formulated what the mind computes", let alone discovered what the primitives of intentions are [17]. Hence, the field of AI and HCI should focus on the development of intentions as the basis for a ToM that are shaped by the common ground we share with others. This focus would be in line with Scott-Phillips' account of communication, arguing that communication is ostensive-inferential (i.e., providing and recognizing evidence for intentions) *before* being made powerful by language [27].

Fourth, the notion of ground truth in AI must be weakened (but never fully abolished). The process of registration that Smith [34] describes is a process of abstraction from reality in which subjects interact with the world and "achieve" objects under the expenditure of effort. For example, object representations must be constantly updated, such as their location in space. Given that this process is mediated by the social context in which the agent is embedded in, a ground truth can hardly be formulated given the human involvement in constructing object representations. Smith warns of the dangers of ideological reductionism and the dangers of imposing a ground truth on others [34]. Similarly, the registration of objects must also be based on reality itself, highlighting the embodiment of agents and an epistemic deference to realism. Hence, the field of AI and HCI should be interested in developing computer architectures that allow the creation of flexible representations that react to physical input and resonate with the conceptualizations of other agents, including humans.

Before concluding this section, we demonstrate how these guidelines can influence the practice of designing, training, and deploying AI systems in the present, attempting to make our ontological guidelines more practical. While we hope that the focus on multi-agent systems, environments, and intentions will bring about novel AI architectures that are embedded in human systems and follow human norms, it is our last point on ground truth that has immediate applicability to practice. If our ontological argument holds weight and indeed questions the universal ground truth of labels, then we might have to ask about the origin and purpose of labels we use in current AI practice. Otherwise, we might construct what the author Kate Crawford [7] calls an "epistemic machinery" that includes "systems of circular logic" based on the labeling of data, the training of AI systems with these labels, real changes depending on the predictions of these AI systems, and the repeated labeling of data after these changes were made. These ontological inscription errors might not only seriously question our ability to "debias" our AI systems as they currently stand, but also create serious harm in their construction of realities for the people that are impacted by them [2].

But then, how do we move away from this "recipe for cultural imperialism and blindness to alternative voices" [34, p.108]? If there is no universal ground truth, how can we ever know anything for sure? At this point, Smith [34] warns against "laissez-faire 'I'm OK; you're OK' pluralism" as reaction to the rejection of a monist ground truth. We should not soften the gravity of any discussion, no matter the subject, but choose a life a in the middle ground: finding some negotiated truth with the humility that comes from being a participant in the world that we try to make sense of. We argue that this humility, this deference for individual perspectives, this participation in the real world, this impact of social context we are embedded in, has immediate applicability to the AI systems that we create and deploy today.

So far, we have referenced AI and HCI in the same way, but there are philosophical differences between them. In this last section, we attempt to bridge these differences and suggest a collaboration of AI and HCI to study a mutual ToM of AI and humans.

5 MUTUAL HUMAN-AI THEORY OF MIND

The field of AI takes a more individualistic approach to agent agency, in which architectures and learning methods are centered on solving problems, rather than on interactions and environments with other agents. While the field of multi-agent systems has studied those, its origin lies in game-theoretic thinking that sees agent behavior as utility maximization, again highlighting its individualistic nature [9, 30, 34]. In contrast to AI, the field of HCI has highlighted human agency and autonomy [4]. Though these word have been giving a wide range of meanings — including subjective efficacy of our own actions, material opportunities, and value alignment between morals and behavior — HCI focuses on the interdependence of humans and how they are impacted by the social context they are embedded in [4]. While their particular philosophies and focus on agency can restrict both AI and HCI in defining an appropriate ToM, the collaboration of them provides an opportunity to define a potent mutual ToM.

Following from the ontological guidelines in the previous section, a ToM should be mediated by other agents, non-prescribed, intentional, and flexible. It should describe the embodiment of agents and their struggle to fit their intentions with the non-conceptualized world (i.e., their attempt to make sense of the world). Similarly, the impact of others should be highlighted and how they shape the concepts, ideas, and objects that we use to understand the world. In the words of Smith [34] and Gallagher [11], deference should be paid to both realistic and constructionist elements of a smart perception. In doing so, a dual level account of ToM is revealed: (1) the individualistic level in which agents come into the world with their own mental apparatus and (2) the social level in which agents are embedded in social contexts.

Given these two levels, both AI and HCI provide the tools to study both levels, respectively. AI, with its focus on agents and architectures, can develop architectures that allow agents to produce flexible representations that are mediated and shaped by other agents. HCI, on the other hand, can study the interactions and social milieu in which agents are embedded in and how they impact the understanding of agents. A ToM that is then build on both AI and HCI provides a potent framework to overcome the ontological challenge of a participatory metaphysics described in this paper.

To make one last point obvious, for AI agents to develop a ToM of other agents, they must be embedded in a *human* social context if that ToM is supposed to be sensible to humans. Hence, an AI-ToM must always be a mutual human-AI ToM. Without the social context we are embedded in, we (and AI) might not be able to make sense of the world we live in, particularly of the mental states of others. It might be as if living in a world that is full objects that we have never seen — like the people in the introduction who have never seen a smartphone — feeling and sensing their physical manifestations without being able to register them as objects. It might be as if having a not-so-smart perception [11] in a world full of human objects. While it should be obvious to the reader that this has to be true for the description of mental states (and an appropriate ToM), in this paper, we argued, given Smith's metaphysical account [34], that it needs to be extended to all conceptualizations and objects.

6 CONCLUSION

This paper underscores the need to consider the ontological assumptions inherent in our theories. By rejecting a metaphysical realism, which states that objects and mental states exist independent of interactions and perceptions, apparent conflicts between traditional accounts of ToM (TT and ST) and perception theory can be transcended. Our exploration into the participatory metaphysics suggested by Smith [34], combined with Gallagher's smart perception [11], highlights the post-perceptional space of TT and ST that both require an ontological parsing provided by perception. By building on PT, we can ask what is required of perception to become smart, paving the way for a novel theory of common ground that highlights the interplay between individual perception and social mediation of the references of everyday objects and mental states. By advocating for a theory of common ground, we propose a shift towards recognizing the mutual construction of reality through negotiation, social interaction, and shared intentions that is rooted in a participatory metaphysics. Specifically, we propose guidelines for a ToM that includes negotiation of reference between agents, emergence of ToM from social interactions rather than pre-definition, a focus on non-conceptual intentions, and flexible representations of objects. We explicitly question the notion of ground truth in AI in this ontological context. Finally, we argue that the references and mental states we share with others include both individual and social elements, which positions the fields of AI and HCI to collaboratively develop a mutual human-AI ToM.

REFERENCES

[1] Open AI. 2022. Introducing ChatGPT. Retrieved 08/23/2023 from https://openai.com/blog/chatgpt.

[2] Ali Alkhatib. 2021. To live in their utopia: Why algorithmic systems create absurd outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21) Article 95. Association for Computing Machinery, Yokohama, Japan, 9 pages. https://doi.org/10.1145/3411764.3445740.

[3] Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. 2020. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, 280, (March 2020), 103216. http://dx.doi.org/10.1016/j.artint.2019.103216.

[4] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D. Mekler. 2023. How does HCI understand human agency and autonomy? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23) Article 375. Association for Computing Machinery, Hamburg, Germany, 18 pages. https://doi.org/10.1145/3544548.3580651.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. (2023). arXiv: 2303.12712 [cs.CL].

[6] S M Carlson and L J Moses. 2001. Individual differences in inhibitory control and children's theory of mind. en. *Child Dev*, 72, 4, (July 2001), 1032–1053.

[7] Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

[8] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: Machines must learn to find common ground. *Nature*, 593, (May 2021), 33–36.

[9] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative AI. (2020). arXiv: 2012.08630 [cs.AI].

[10] Daniel Clement Dennett. 1981. *The Intentional Stance.* MIT Press.

[11] Shaun Gallagher. 2008. Direct perception in the intersubjective context. *Consciousness and Cognition*, 17, 2, 535–543. Social Cognition, Emotion, and Self-Consciousness. https://www.sciencedirect.com/science/article/pii/S1053810008000342.

[12] Vittorio Gallese and Alvin Goldman. 1998. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2, 12, 493–501. https://www.sciencedirect.com/science/article/pii/S1364661398012625.

[13] Arvin Goldman. 2012. Theory of mind. In *Oxford Handbook of Philosophy and Cognitive Science.* Eric Margolis, Richard Samuels, and Stephen Stich, editors. Oxford University Press.

[14] Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems.* I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors. Volume 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/70222949cc0db89ab32c9969754d4758-Paper.pdf.

[15] Jakob Hohwy. 2013. *The Predictive Mind.* Oxford University Press, (November 2013). https://doi.org/10.1093/acprof:oso/9780199682737.001.0001.

[16] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner monologue: Embodied reasoning through planning with language models. (2022). arXiv: 2207.05608 [cs.RO].

[17] P. N. Johnson-Laird. 1986. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Harvard University Press, USA.

[18] Julian Kiverstein. 2011. Social understanding without mentalizing. *Philosophical Topics*, 39, 1, 41–65. Retrieved 02/20/2024 from http://www.jstor.org/stable/43154592.

[19] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations.* https://openreview.net/forum?id=HJGv1Z-AW.

[20] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations.* https://openreview.net/forum?id=Hk8N3Sclg.

[21] Yann LeCun, Y. Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521, (May 2015), 436–44.

[22] Alexandre Linhares. 2000. A glimpse at the metaphysics of Bongard problems. *Artificial Intelligence*, 121, 1, 251–270. https://www.sciencedirect.com/science/article/pii/S0004370200000424.

[23] Maurice Merleau-Ponty. 1962. Phenomenology of perception (c. smith, trans.) (1962).

[24] Søren Overgaard. 2017. Merleau-Ponty and Wittgenstein on mindreading: Exposing the myth of the given mind. *Wittgenstein and Merleau-Ponty*, 49–65.

[25] Thomas Parr, Giovanni Pezzulo, and Karl J Friston. 2022. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior.* MIT Press.

[26] Josef Perner. 1991. *Understanding the Representational Mind.* MIT Press, Cambridge.

[27] Thomas C. Scott-Phillips. 2014. *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Bloomsbury Publishing.

[28] Thomas C. Scott-Phillips, Simon Kirby, and Graham R.S. Ritchie. 2009. Signalling signalhood and the emergence of communication. *Cognition*, 113, 2, 226–233. https://www.sciencedirect.com/science/article/pii/S0010027709001930.

[29] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27, 3, 379–423.

[30] Yoav Shoham and Kevin Leyton-Brown. 2008. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.

[31] Matthew Sidji, Wally Smith, and Melissa J. Rogerson. 2023. The hidden rules of Hanabi: How humans outperform AI agents. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23) Article 757. Association for Computing Machinery, Hamburg, Germany, 16 pages. https://doi.org/10.1145/3544548.3581550.

[32] Marc Slors. 2012. The model-model of the theory-theory. *Inquiry*, 55, 5, 521–542. https://doi.org/10.1080/0020174X.2012.716205.

[33] Brian Cantwell Smith. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. The MIT Press, (October 2019). https://doi.org/10.7551/mitpress/12385.001.0001.

[34] Brian Cantwell. Smith. 1996. *On the Origin of Objects*. eng. MIT Press, Cambridge, Mass.

[35] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. LLM-Planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. (October 2023), 2998–3009.

[36] Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. Theory of mind abilities of large language models in human-robot interaction: An illusion? *arXiv preprint arXiv:2401.05302*.

[37] Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 1, 103–128. https://www.sciencedirect.com/science/article/pii/0010027783900045.

[38] Ludwig Wittgenstein. 1980. Remarks on the philosophy of psychology, volume 2.