Theory of Whose Mind? Exposing the Shortcomings of One of HRI's Core Concepts

Cansu Elmadagli*
cansu.elmadagli@oru.se
Department of Media and Communication Studies, Örebro University
Sweden

Jennifer Renoux*
jennifer.renoux@oru.se
Center for Applied Autonomous Sensor Systems, Örebro Unversity
Sweden

Abstract

The concept of Theory of Mind (ToM) is central to many social robotic studies. It may be invoked during the design of social robots as a way to improve collaboration or create a form of "social intelligence". It is also considered as an established fact and never put in question. However, many scholars have analysed the concept of ToM from a critical perspective and argued that it is, in fact, a theory with significant shortcomings, that rests on neuronormative and neuroprivileged grounds. In this paper, we explore these arguments and what this change of perspective means for the field of Social Robotics. We argue that the field should abandon the concept of ToM and move forward to more appropriate and inclusive models and research questions, and propose some directions to do so.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; **Interaction paradigms**; **Interaction design theory, concepts and paradigms**.

Keywords

Theory of Mind, Critical Research, Social Robotics, Autism, Mindreading, Mentalizing

ACM Reference Format:
Cansu Elmadagli and Jennifer Renoux. 2026. Theory of Whose Mind? Exposing the Shortcomings of One of HRI's Core Concepts. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3757279.3788818

1 Introduction

Robotic technologies are rapidly advancing and are progressively integrating into public and private aspects of everyday life [18]. According to official estimations, robotic technology will become pervasive in various contexts of social interaction in the following two decades [41]. Scholars have pointed out how robotic technologies might reflect and influence society as they are coded with implicit norms, values, and ideologies, which result in reproducing and reinforcing power inequalities [41, 63]. Until now, social robots and artificial intelligence (AI) have been critically examined through multiple lenses, such as race, gender, disability, and sexuality [77]. In line with the said critiques, we look at Theory of Mind (ToM) as a commonly used framework to define human cognition and sociality in Human-Robot Interaction (HRI).

ToM is a central concept to many robotic studies [34], invoked during the design of social robots as a way to improve collaboration or create a form of "social intelligence". HRI research often refers to the concept of ToM as a human ability [20], "central to human behavior" [5], a "major component of human cognition" [16]. Its place as an accepted, proven component of the human mind is never questioned. However, critical analysis of the concept has highlighted many shortcomings of the ToM narrative and criticized it for being a neuronormative and neuroprivileged model of cognition. This analysis should at least prompt scholars in HRI to be more nuanced in their adoption, and maybe even warrant the total abandonment of the concept.

This paper aims to bring these critiques to the HRI field and explore what this change of perspective means for our research. We start our article by providing the relevant contextual knowledge about ToM (Section 2) and its current use in HRI (Section 3. Then we review the existing critiques against the concept of ToM (Section 4) and explore how these critiques impact the HRI field (Section 5). Finally, we look towards the future and highlight opportunities for the field to go forward with more inclusive approaches and methods (Section 6).

Before we embark on this journey, we want to state that we intend not to reproduce ableist and normative discourse in this paper. However, given the nature of our contribution, we are going to use terms that are considered ableist and normative, as we are placing our work in the historical context. Throughout the paper, we note when such terms are being used to raise awareness of their connotations.

2 What is Theory of Mind

Theory of Mind (ToM) is a multi-faceted concept, and it would be impossible to give a complete overview of it with all the nuances different approaches hold. Thus, in this section, we will emphasize salient aspects and important distinctions between approaches.

Before diving into these approaches, it is important to point out that since its inception, the field of ToM research has been mainly centered on animals and children. Thus, ToM as a term mostly and specifically refers to children's understanding of other minds. However, research on adults has been emerging for the past decade [2, 37, 87]. As Apperly [3] argues, although there seems to be a more or less consensus of what ToM is on the surface, this

*Both authors contributed equally to this research.

This work is licensed under a Creative Commons Attribution 4.0 International License.
HRI '26, Edinburgh, Scotland, UK
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3788818

is not necessarily the case since scholars approach it with "rather different interpretations". Inherently, ToM is "a loosely defined construct with diverse operationalizations" [87], and there is no clear "taxonomy of abilities that make up the construct of ToM" [87].

Today, ToM is seen as an "an important ability that facilitates social interaction" [87] as well as "essential to both self-reflection and coordinated social action" [44]. Over time, what is considered as a part of ToM has broadened and now includes abilities such as joint attention, perspective taking, empathizing, recognition of social-emotional cues, understanding of shared context, and joint action [10, 13, 37, 58].

The term Theory of Mind (ToM) was first coined in the 1970s by Premack and Woodruff [57] in their seminal work titled "Does the Chimpanzee Have a Theory of Mind?". This study aimed to understand intentionality in non-human primates, such as chimpanzees. From this perspective, ToM as a term was used to indicate the ability to impute mental states to oneself and to others[44, 57]. Extended to human beings, this ability demonstrates itself as mentalizing or 'mindreading' in 'normally developed' human beings. Therefore, ToM capabilities allow humans to take another person's perspective or be in each other's shoes. It allows humans to 'intuit' others' beliefs, intentions, and desires based on behaviors such as body language. This capability is also used by humans to not only understand and explain others' behaviors but also predict future ones [64].

2.1 Two opposing approaches

In its original and early understandings, the aim of ToM as a theory was to find empirical evidence for the "philosophical claim that our ordinary, common-sense understandings of mental life" were the result of a theorization process [44]. From this perspective, understanding others' minds was seen as a process in which one forms and applies a theory "about the nature of minds" or a folk psychology[1] [56, 64]. Thus, scholars argued that application of this theory as a "system of inferences" enabled one to interpret the unobservable mental states underpinning observable behavior [57]. In this vein, social actors are viewed as rational subjects that intentionally act based on their individual and subjective perspective [58]. This early conception of ToM is called *theory-theory* in literature. Within the theory-theory approach, scholars were divided into two in their understanding of ToM and the theorization processes. On the one hand, scholars such as Meltzoff and Gopnik [49] argued that theory of mind is formulated similarly to scientific theories through evidence and causal modeling. From this perspective, children are treated as 'little scientist(s)' who create and modify theories of other minds according to incoming data. On the other hand, scholars such as Baron-Cohen [7, 8] and Leslie [42] popularized the modularity approach that argued for the existence of built-in, innate modules or cognitive mechanisms of ToM which enable human beings to intuitively and automatically mindread [89].

In addition to the theory-theory perspectives, an opposing *simulation approach* was developed by scholars in the 1980s. According to this approach, understanding other minds depends on the imagination or simulation of what someone else desires, feels, and thinks by applying subjective and "first-hand experience" [37]. Thus, instead of theorization of observable behaviors and mental states, humans simulate others' mental states within themselves; in other words, step into their shoes.

Although the scholarly discussions on ToM have mostly revolved around the opposition of the theory-theory and the simulation approaches, "it has recently become more common to argue for some kind of mixed approach" [89]. Indeed, the discrepancies of empirical data, experimental results, and interpretations in the field led some scholars to believe that ToM might demonstrate itself as two systems. According to this view, regardless of their age, all human beings have ToM capacities that are "cognitively efficient but limited to simple problems" [2]. It is only adults and older children who can use ToM in a "much more flexible, but also more cognitively demanding" manner [2].

2.2 The False Belief Task

Regardless of how differences in how the ToM is developed in an individual, the False Belief Task was designed by Wimmer and Perner [84] as a measurement to assess if human or non-human primates have an 'intact' ToM [1]. The task aims to understand if they can predict someone else's "behavior on the basis of their false belief" [10]. To this day, false belief tasks are still traditionally and most commonly used for ToM measurement. Yet, as Apperly [3] emphasizes, scholars' interpretations of the same false-belief tasks "assume very different interpretations of children's performance." This speaks to the inconsistencies among the conclusions scholars draw based on ToM across the field.

2.3 A development in three waves

The conception of ToM started from a cognitive perspective and over time "has subsequently been hybridized to include sociocultural and/or psychoanalytic perspectives" [4]. The progression of ToM research can be seen in three parts. The early research, which some scholars call 'the first wave,' concentrated on the 'normal' development of ToM in early childhood, more specifically preschool children, and the so-called 'abnormal' development of it in specific populations such as autistic children [3, 37, 87].

The 'second wave' focused on diversifying the research by expanding small-scale studies of middle-class children to larger population groups. Additionally, it expanded the batteries of tasks that are used to measure ToM beyond binary categorization of children as "passers" or "failers" of false-belief tasks [37]. This approach allowed more sensitivity to variation in children's performance on tasks and aggregation of "measures that are more reliable than individual task scores" [37].

The 'third wave' of ToM scholarship can be seen as the current wave, which expanded "the developmental scope both downward into infancy and upward into adulthood" [37]. Although recent ToM research expanded its narrowly focused participant groups and experimental tasks, "the theoretical work has not kept pace with these changes" [3]. This lack of progress then results in adopting "assumptions about the nature of theory of mind from other research, with little regard for whether these assumptions are appropriate" [3]. This can also be said about the ToM approaches in

[1] A folk psychology is a non-expert, ordinary people way of attributing mental states to others and understanding behaviors. It is sometimes referred to as naive psychology or common-sense psychology.

social robotics, due to their utilization of not only outdated but also inaccurate understandings of ToM.

3 Theory of Mind in Social Robotics

The differences in interpretation of what ToM is in the Cognitive Sciences are reflected in the field of Social Robotics, and have been for more than two decades. In 2002, Scassellati [67] discussed the notion of implementing a ToM for humanoid robots, and the different models existing in the cognitive science literature. This work can probably be considered as one of the first to bring the notion of ToM into robotics. In a recent survey focusing on the implementation of an Artificial ToM, Gurney and Pynadath [34] discussed different theories on how ToM is considered in cognitive science (similarly to what we did in Section 2), and reviewed existing implementations of Artificial ToM. The authors concluded that the lack of consensus around what it means for a robot to have ToM capabilities, as different works implement different constructs, such as false-beliefs, recursive reasoning, prescriptive and descriptive ToM, etc. However, this review focused on the implementation aspects, but it is crucial to look at the reasons that motivates researchers to implement an Artificial ToM, as they are the basis for choosing which concept related to ToM will be implemented and how the corresponding assumptions shape the implementation or study.

Table 1 shows a summary of representative papers that explicitly address the question of an Artificial ToM. We categorized these papers by their expressed motivation behind the need for an Artificial ToM, with a representative quote from the paper, and indicated for each paper which background paper from the Cognitive Science they used regarding the definition of ToM. We observe that the motivations for implementing an Artificial ToM are mostly twofold: (1) Supporting Human-AI Collaboration, (2) Implementing a form of Social Intelligence. A lot of the studies that aim at supporting Human-AI Collaboration motivate themselves through the need of humans and robots to adapt to each other, and therefore be able to reason about what the other is doing. These works usually tackle the problem of false beliefs and perspective-taking (usually through reasoning over the human's plans or intentions). In this case, ToM is seen as a general "concept" rather than an actual theory in the strict sense, encompassing different capabilities, and researchers attempt to replicate one or several of these capabilities. On the Social Intelligence side, work usually starts with the assumption that the robots need to be endowed with basic social skills to be accepted by humans or to be efficient in their interaction. In this case, ToM is correctly considered as a theory to predict how the human mind works, and researchers are concerned with replicating this assumed mind's inner working in robots.

Note that though we limited ourselves to the field of Social Robotics in this paper, other fields of research are explicitly engaging with the concept of ToM, for instance in multi-agent systems [62], to model deception [65], support collaboration and coordination [55], or resolve conflict [26].

Another observation that can be made from the current state of the art in ToM for Social Robotics is that an overwhelming majority of work, even very recent, does not engage with the nuances we explained in the previous sections. Some studies have been built on top of the survey made Gurney and Pynadath [34], and therefore address the differences between the "theory-theory" and the "simulation" approaches. This is the case, for instance, of [51] and [35], even though this latter one refers to the false-belief task as "The quintessential ToM study, the Sally-Anne Test", without discussing the differences in interpretations that Apperly [3] highlighted. Most studies are still anchored in the "first wave" of ToM research and fail to engage with more recent and nuanced approaches.

4 The problem with the Theory of Mind

4.1 A field based on Autism

The concept of a Theory of Mind cannot be detangled from its links to Autism and, specifically, its medical model. According to the deficit-based medical model, Autism is caused by either the absence or malfunction of the biological cognitive mechanisms that generate ToM [7, 17]. Baron-Cohen and his colleagues went "so far as to suggest that a faulty ToM is the primary deficit in Autism and applies to all individuals on the autistic spectrum" [25]. Autistic individuals are therefore sometimes referred to as 'mindblind'[2].

As it has been historically common with various medical models, the creation of a medical model of Autism finds its infancy in studying what is considered "broken" in order to understand what it considered "normal." Prolific ToM scholars such as Meltzoff [48] even argue that it is autistic children who helped to "mold our current understanding of social cognition," thus highlighting that a considerable amount of the foundation of ToM framework relies on the assumption that autistic children are deficient in ToM capabilities. Autism is therefore seen as the 'abnormal', helping understand the 'normal'[3]. While this approach can be defended when dealing with diseases, where the 'broken part' can be clearly and factually observed (e.g. a liver that does not function, leading to a deteriorated health, leading to life-threatening issues), it is much more difficult to factually validate such interpretations when the object of study is human psychology. Not only does this approach have validity concerns, but it is also a harmful one that pathologizes human differences. In the case of Autism, autistic people are categorized as 'lacking' due to their so-called deficits in demonstrating 'correct' sociality in accordance with neurotypical norms. Through this lack, the Theory of Mind is defined and tested. Thus, resulting in a circular definition where the existence of something is proven through the supposed lack of it in certain children.

4.2 The pitfalls of the false-belief task

Measurements such as false-belief tasks are utilized to prove that autistic children lack ToM or have deficits of it. The Sally-Anne task, the poster child of false belief tasks, was developed further by Baron-Cohen and colleagues to prove that autistic children and children with Down syndrome are deficient in their ToM capabilities compared to their 'typically' developing peers. This was seen as a "landmark study" in the field and supported the validity of the false-belief task [37].

[2]The term 'mindblind' is considered ableist, and we are only going to use it in this paper to oppose its original conception.
[3]The terms 'abnormal' and 'normal' are here used in direct reference to the literature and not as a representation of our position.

Table 1: A summary of representative papers about artificial theory of mind, grouped by motivations

Motivation	Reference	Representative Quote	ToM background
Supporting Collaboration	Devin and Alami [20]	"we present [...] a framework that allows the estimation by the robot of its human partner's mental state related to collaborative task achievement."	Baron-Cohen [7] and Premack and Woodruff [57]
	Buehler and Weisswange [12]	"An autonomous agent that wants to assist a human partner effectively, will need a similar understanding of human mental reasoning to improve the cooperative performance while avoiding information overload and annoyance."	
	Shvo et al. [73]	"For example, a robot could communicate to its human teammate that the conditions necessary to the success of her plan do not hold or, alternatively, the robot could act in the world to ensure that those conditions hold."	Premack and Woodruff [57]
	Sarthou [66]	"To facilitate the interaction with humans, making it smoother, more natural, and more efficient, a key capability of the robot at the situation assessment (SA) level is Visual Perspective Taking (VPT)."	Baron-Cohen et al. [9]
	Yang et al. [85]	"With the rise of artificial intelligence (AI) and the desire to ensure that such machines work well with humans, it is essential for AI systems to actively model their human teammates, a capability referred to as Machine Theory of Mind (MToM)."	Frith and Frith [28]
	Yu et al. [88]	"By reasoning about human mental states and behaviors, robot ToM can improve the communicability and trust of robots in human-robot collaborative settings"	Fodor [27] and Baker et al. [6]
Social Intelligence	Cuzzolin et al. [16]	"This flags the need for AI to tackle 'hot' cognition[...]. Hot cognition refers to emotional and social cognition, including Theory of Mind (ToM)."	Baron-Cohen [7]
	Scassellati [67]	"If we are to build human-like robots that can interact naturally with people, our robots must know not only about the properties of objects but also the properties of animate agents in the world."	many sources, mostly explore Baron-Cohen [7] and [43]
	Dissing and Bolander [23]	"For these robots to be accepted by the users, they will need to possess basic social skills and behave in a socially acceptable manner"	Premack and Woodruff [57]
	Gurney et al. [35]	"With much of human social cognition hinging on ToM, it is unsurprising that AI researchers see it as a possible solution to many hard problems in artificial social intelligence (ASI)."	Premack and Woodruff [57] and Gurney and Pynadath [34]
	Patricio and Jamshidnejad [52]	"In order to interact as humanly as possible, SARs should exhibit similar understanding of rational agents"	Premack and Woodruff [57]
	Tavella et al. [78]	"artificial tools, to be effective in these societies, should somehow adopt a social interaction perspective that is closer to the human one"	Wimmer and Perner [84]

It is important to note that, in the field of Cognitive Science, ToM is treated as "a simple quantitative entity" [3], and a limited number of false-belief tasks are considered "as the gold-standard way " to measure it [3]. As Apperly [3] points out, though, it is "highly unlikely" that a single false-belief score can meaningfully capture various aspects of ToM, regardless of the task used to calculate it. Furthermore, traditional false belief or (visual) perspective-taking tasks are not useful to assess typical adult ToM capabilities, since they measure "basic mindreading concepts" [2]. In this vein, no individual task or combination of tasks can act as the most accurate measure of ToM across all age groups [3]. In addition, studies have shown that an autistic individual's improved performance in

false-belief tasks does not translate into "parent- and/or teacher-rated social adjustment" [37]. According to the ToM framework, the false-belief task is supposed to measure ToM capabilities, and failing the task demonstrates social impairments, as in the case of Autism. However, as evidence suggests, an improved score in a false-belief task does not correlate with the perceived social capability of autistic children. In line with said critiques, the validity of false-belief tasks in the evaluation of the ToM capabilities comes under scrutiny.

4.3 Theory of Mind is not a good theory

As scholars pointed out, the innate existence of the cognitive mechanisms of ToM is an a priori [70]. Leudar et al. [45] state that framing human beings "as theoreticians[4] is not actually an empirical discovery at all, but a background assumption". In studies that measure ToM, its existence is presented as factual and then tested through experiments that will validate this claim. As Gernsbacher and Yergeau [31] emphasized in their detailed study on ToM, "the claim that autistic people lack a theory of mind fails empirically; it fails in its specificity, universality, replicability, convergent validity, and predictive validity". These failures have been observed on more than one occasion, yet autistic people are still deemed as lacking a ToM within different disciplines.

It is not only autistics who cannot 'mindread', nobody can. Our inferences of others' mental states are partly based on our personal experiences. Moreover, if there are considerable differences between two people's life experiences, one's empathic accuracy reduces [39, 47]. Human beings are subject to implicit and egocentric biases when inferring others' mental states and interpreting others' actions. These biases result in inaccuracies and manifest individual variations [2, 44, 87]. Similarly, Gallagher [29] points out that ToM is not a "good explanation of non-autistic intersubjective experience"[5], let alone explain autistic intersubjectivity. This is because an understanding of human cognition through ToM is reductionist in its essence. ToM is overall a narrow, decontextualized, disembodied, and intellectualized understanding of human minds [45, 54]. ToM is not only reductionist but also 'inadequate' when it comes to defining and representing Autism because autistic mental states are not considered. Thus, "the theory is, in effect, mindblind with regard to autistic perspectives" [74].

In line with this, we would like to further argue that ToM is not only mindblind to autistic experiences and perspectives, but mindblind on a larger scale regarding human diversity and difference. ToM scholars claim that it is not only autistic people who have ToM deficits but also "deaf people, blind people, indigenous peoples, poor people" [86], people with conduct disorder and Down syndrome [37], schizophrenia [37, 44], and traumatic brain injury [13]. ToM is therefore not only ableist due to its ideologically neuronormative and neuroprivileged position [14] but also possesses classist and racist undertones. Against ToM theorists' claims, not only are there many autistic people who pass false belief tasks, but also a lot of non-autistic people who are more likely to fail them [31]. Things that affect ToM positively or negatively are multi-faceted and various. Socioeconomic status, number of siblings in the house, cultural background, executive function, memory, language capabilities, one's personality and social motivation, having 'mind-minded' parents, bilingualism, race, sex and even one's mood in the moment [2, 3, 36, 37, 58].

The problems with the ToM approach are manifold, yet, as emphasized by some scholars, critical engagement with it is scarce. This is surprising considering how conceptualization of ToM "has been associated with probably the fastest growing body of empirical research in psychology" [45]. As various scholars also pointed out, the problematics of the ToM framework are "fairly representative of contemporary psychology [in terms of] philosophical misconceptions and inappropriate experiments" [70]. ToM is a framework that has been debated in terms of its "unethics and its methodology" [82]. Through an examination of existing research publications, scholars highlight the tendency of researchers to portray autistic individuals as lacking agency and "full humanity" [40]. Moreover, autistic individuals are often portrayed as objects of study rather than active participants, resulting in the neglect of their lived experiences [75].

4.4 The double-empathy problem

To challenge the deficiency model of Autism, both autistic researchers and the Autism community have been involved in changing the narrative around Autism and autistic people. One of the most important critiques of ToM has been put forth by Milton [50], who is an autistic scholar himself. He stated that it is not only autistic people who have difficulties with interactions and communication, but there is what we call a 'double-empathy problem' between autistic and non-autistic individuals. According to this 'double-empathy problem' approach, both groups display equal levels of difficulty in interacting with and understanding each other. This claim is supported by further research, which demonstrates that non-autistic people do "have difficulty interpreting the mental states" of autistic people based on their body language [71]. Sheppard et al. [71] also notes that if this difficulty goes both ways, as their results demonstrate, autistic people could be facing "a barrier to social interaction due to their mental states being misinterpreted by others" which could prompt "confusing social interaction(s) and consequent negative experiences" as a neurominority group.

We would like to argue that this double-empathy problem, especially in relation to ToM, extends to other marginalized groups that fail false-belief tasks, which we previously mentioned. Scholars argue that ToM tests most likely do not measure ToM capabilities. They are heavily reliant on (spoken) language that is complex [31, 76]. It is documented in the literature that language "seems to be necessary to acquire fully fledged meta-representational ToM" [58]. Language acts as the foundation to develop understanding of false-beliefs, although the precise nature of this foundation has not yet been established [37]. Deaf people who grow up with deaf parents who are native sign language users demonstrate 'typical' language and ToM development. On the contrary, deaf children "who grow up in hearing families without native signers" show delayed language and ToM development [58]. We believe that this example demonstrates how language can act as both a barrier and a facilitator of establishing common ground among individuals. Additionally,

[4] In reference to the theory-theory approach discussed in Section 2.1

[5] Here, intersubjective experience, or intersubjectivity, refers to the shared understanding that emerges from interpersonal interactions.

we do not only mean language in its literal sense per se, but also include inherent communicative differences and preferences here.

5 What does it mean for Social Robotics?

Scholars frequently use studies on Autism and ToM as their starting points to develop artificial intelligence and robots that match human capabilities (as already explored in Section 3). In Picard [53]'s words, computers "will be like autistics" who are "not good at understanding emotional significance" (79) or, in Kaminka [38] 's words, "robots and other synthetic agents are generally autistic" unless they possess the ability to "behave correctly towards others". Thus, scholars should aim for "curing robot Autism." Both claims are formulated on the uncontested presupposition that ToM is a fact instead of a hypothesis [9], and the uncontroversial, so-called scientific arguments of autistic people lacking ToM [74]. As Williams [81] points out, the cycle of autistics being a model of "how not to build machines" and machines being a model for how autistic behaviors can be corrected is "is sustained in large part by interpretations of Simon Baron-Cohen's Theory of Mind." This is also reflected by the fact that, within the domain of human-computer and human-robot interaction studies, autistic individuals are frequently depicted as 'computers', 'robots', or machine-like entities [81].

In this context, what do the critiques of the ToM paradigm and the double-empathy problem explored in Section 4 entail for research in Social Robotics? We believe that the answer to that question lies in the motivations that scholars follow to perform their research. As we saw in Section 3, the question of implementing an Artificial ToM mainly follows two types of motivations: supporting collaboration, and endowing robots with a form of social intelligence.

For many scholars, an Artificial ToM is a "concept," a way for a robot to reason upon the beliefs, intentions, and plans of a human teammate, thus allowing better collaboration. We fully agree that skills such as false-belief management and perspective-taking are necessary for collaboration. In fact, research works have been addressing these questions for a while without explicitly referring to the concept of ToM, both in robot-robot [32, 60] and human-robot interaction [24, 79]. However, these concepts have been born and are still anchored in the normative framework of ToM, and researchers should make a deliberate effort to break free of this link. This does not mean that we must discount the important body of research that has been performed in making robots manage false-beliefs or perspective taking, but that such capabilities need not be gathered under the normative and refuted concept of "Theory of Mind." Such works can live "on their own," free from the harmful narrative of invoking the creation of an artificial ToM.

Regarding the works that aim at giving robots a form of social intelligence, the issue is more pervasive as it becomes a matter of questioning our own internalized biases and biased representations. What is "social intelligence?" What are the capabilities that make a robot "social" and which humans have been used to represent the norm? In fact, recent research has shown that what is represented as the norm in HRI is heavily biased towards a W.E.I.R.D. (Western, Educated, Industrialized, Rich, Democratic) representation [46, 68]. In addition, Seaborn et al. [68] also noted an able-bodied and neurotypicality sampling bias in the HRI field. Generalization of such subpopulations as representative of the entire human race without any empirical evidence is problematic [36]. What we currently associate with 'social intelligence' is quite likely 'W.E.I.R.D., able-bodied, and neurotypical social intelligence'.

Researchers should also be aware that their choices in design and implementation risk marginalizing populations already marginalized even more. Indeed, the implementation of Artificial ToMs as it is currently done represents the status quo and legitimizes a neuronormative and neuropriviliged sociality in general. Additionally, the deficit-based, medical model of Autism results in discriminatory social robotics designs, which in turn result in technoableism [59, 72].

The critiques of the ToM concept must bring researchers to reconsider the skills and functionalities they want to implement in robots to make them social. If we move away from the "Autism-as-a-problem" formulation, as we should, then we must ask ourselves: if robots are truly autistic, why do we want to fix them?

6 What's Next?

Without challenging normative theoretical approaches upon which we build technologies, we cannot ensure technological progress that does not discriminate against, alienate, or exclude certain groups, specifically marginalized ones. Theoretical assumptions reflect biases within research communities. In addition to our standpoints as researchers, it is essential for us to also critically reflect on the theories that we make use of [33]. Both research traditions and research communities possess implicit and explicit biases, which "have a profound impact on the shape of research processes and technological artefacts" [30]. Theories such as ToM cannot be taken for granted and treated as facts. If we are to develop inclusive technologies that "do no harm", an epistemological and ontological investigation of the theoretical concepts we use is necessary. As Leudar and Costall [44] point out, more ecologically sound observational and analytic methods, conversation, and (critical) discourse analyses are dismissed by ToM scholars/cognitive psychologists as only contributing subjective evidence. However, it is through these subjective experiences and context that we can fully understand human minds. For this reason, we propose an alternative way of understanding others' minds as interactional, intersubjective, situational, and context dependent [19, 22, 80].

We need to shift the paradigm from (neuro)-normativity to (neuro)-diversity, from privilege to inclusion. As scientists, we all know that paradigm shifts are possible, as challenging as they might be. The scientific history is full of examples in terms of paradigm shifts, although the pushback for this type of shift might be strong. When it comes to such a paradigm shift, what is at stake are scientific careers, reputations, funding opportunities, and positions of power built upon the Theory of Mind. Yet, the question is, what do we want our future technologies to look like? One that reproduces (neuro)normativity, (neuro)privilege, and inequality, or one that fosters (neuro)diversity and inclusion?

From a practical standpoint, several frameworks have been developed and can be of use to scholars who want to engage with these questions. The "Countervention" framework, from Williams et al. [83], challenges us to reflect on our research questions and processes in a way that is "reparative" for populations harmed by

previous technological research and development. The "Celebratory Technology" approach [11] aims at highlighting the positive aspects of neurodiversity, instead of what society considers as 'undesirable traits', to empower stigmatized populations. The "Data Feminism" framework proposed by D'ignazio and Klein [21] allows us to critically examine technologies in terms of power, oppression, and structural inequalities. Additionally, the "Design Justice" framework developed by [15] helps researchers interrogate design practices, values, narratives, sites, and pedagogies in relation to matrices of domination and intersectionality.

All of the suggested frameworks aim to critically reflect upon design practices from the ground up. They are concerned with issues of inequality, power relations, oppressive ideologies, and exclusion. Thus, they make sure to not only include but also center marginalized groups' perspectives and lived experiences. By de-centering W.E.I.R.D., able-bodied, and heteronormative populations, these frameworks create a deliberate inversion of focus as a mode of critique of and resistance against the status quo. This allows researchers to not only critically examine exclusionary frameworks such as ToM, but also imagine alternative ways of understanding human sociality and create artificial sociality. Such alternative ways will be difficult to reach and will require a lot of research. But these questions start to arise in the field, for instance with discussions of the differences between "sociomorphing" and "anthropomorphizing" [69].

As a starting point for researchers reading this article and willing to start shifting their practices, we provide some examples in relation to ToM capabilities that the aforementioned frameworks would ask us to consider:

(1) Question what ToM capabilities really are, based on whose and what kinds of norms they get defined as 'normal' and measured as 'acceptable'.
(2) Examine gendered, racial, cultural, and disability aspects of turn-taking, perspective-taking, empathizing, (joint) attention, and recognizing socio-emotional cues. Reflect on them in relation to power inequalities, subject positions, and situated knowledges. [6]
(3) Consider whose lived experiences and perspectives would be considered as legitimate knowledge due to their privileges, while which groups' would be discredited due to their 'less favorable' societal positioning. [7]
(4) Consider the issue of agency when pathologizing vulnerable population groups - such as autistic and/or intellectually disabled children - when it comes to research tasks. Is what we observe actually an inability or ToM deficiency, or can it be disinterest, resistance, and exercise of one's agency?

In alignment with their socio-political foci, these suggested frameworks would most likely prompt one to reject the current ToM framework due to its ableist, classist, and racist views of the world. They would instead ask us to elevate non-normative understandings of human minds and interactions in a context-dependent and community-informed manner. A more radical alternative would be an outright refusal to design technologies that possess exclusionary approaches such as ToM. [8]

As scholars in HRI, we are already used to conducting interdisciplinary studies and engaging with other fields than our own, primarily psychology and cognitive sciences. For this to happen, we need a "critical, inclusive, and future-oriented dialogue" [61] across disciplines and sub-disciplines. Now, we advocate for an expansion of our interdisciplinary practices to include critical research. Indeed, scholars in these fields can bring new methodologies to the table to help us question our practices and ultimately create better human-robot interactions.

7 Conclusion

In this paper, we presented the state of research in the critical analysis of the concept of Theory of Mind. Specifically, we explained that its conception is highly dependent on the premise that autistic individuals lack a ToM, and that this is actually challenged by many findings in the field. Moreover, we explored what these challenges mean for the field of HRI, in which ToM is used as a core concept. We conclude that scholars in HRI should move past the concept of ToM and cease to invoke it as a justification for studying false-beliefs, perspective-taking, or other necessary capabilities of a robot. These research works can live "on their own," without the need for an Artificial ToM. Above all, scholars should abandon the idea that a ToM is a necessary component of "social intelligence" and engage with critical analyses of their research questions and practices.

The goal of the current paper is not to condemn scholars who may have implemented some Artificial Theory of Mind, or used it as a justification for their work. In fact, one author of the current paper is guilty of doing so themselves. Our goal is to help the HRI field as a whole move past dated constructs and towards reparative reflections [83]. Quoting Gurney and Pynadath [34], "roboticists, and AI researchers more generally, are in an enviable position of being able to take the best from psychology, philosophy, and other disciplines working on human social cognition, implement it, and improve upon it without the need to justify their design choices." We hope that the next phase of Human-Robot Interaction research will be based on better representations of human cognition, in all its diversity.

Acknowledgments

This research is partially funded by FORTE (Forskningsrådet för hälsa, arbetsliv och välfärd, gr no. 2024-01349),Vetenskapsrådet (gr no. 2022-04676), and the European Commission via HORIZON project EuRobin (gr no. 101070596).

References

[1] Gabriella Airenti. 2015. Theory of mind: a new perspective on the puzzle of belief ascription. *Frontiers in psychology* 6 (2015), 1184.
[2] Ian Apperly. 2013. 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* (2013), 72–92.
[3] Ian A Apperly. 2012. What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly journal of experimental psychology* 65, 5 (2012), 825–839.

[6] Situated knowledges is an approach that highlights the situatedness, subjectivity, and partiality of knowledge (see Haraway 1988).
[7] This point is directly related to the concept of epistemic injustice (see Fricker 2007).
[8] Such a alternative could be seen as a part of the TechWontBuildIt Movement as addressed in Constanza-Chock's (2020) work.

[4] F Baglio and A Marchetti. 2016. Editorial: when (and how) is theory of mind useful. *Evidence from Life-Span Research. Front. Psychol. 7: 1425. doi: 10.3389/fpsyg* (2016).

[5] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 33.

[6] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. 2017. Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human mentalizing. *Nature Human Behaviour* 1, 4 (mar 2017), 0064. doi:10.1038/s41562-017-0064

[7] Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.

[8] Simon Baron-Cohen. 2008. *Autism and Asperger syndrome*. Oxford university press.

[9] Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition* 21, 1 (1985), 37–46.

[10] Jill Boucher. 2012. Putting theory of mind in its place: psychological explanations of the socio-emotional-communicative impairments in autistic spectrum disorder. *Autism* 16, 3 (2012), 226–246.

[11] LouAnne Boyd. 2023. Conceptualizing celebratory technologies for neurodiversity to reduce social stigma. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.

[12] Moritz C. Buehler and Thomas H. Weisswange. [n. d.]. Theory of Mind based Communication for Human Agent Cooperation. In *IEEE International Conference on Human-Machine Systems*.

[13] Lindsey J Byom and Bilge Mutlu. 2013. Theory of mind: Mechanisms, methods, and new directions. *Frontiers in human neuroscience* 7 (2013), 413.

[14] Amandine Catala, Luc Faucher, and Pierre Poirier. 2021. Autism, epistemic injustice, and epistemic disablement: A relational account of epistemic agency. *Synthese* 199, 3 (2021), 9013–9039.

[15] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

[16] F. Cuzzolin, A. Morelli, B. Cîrstea, and B. J. Sahakian. 2020. Knowing me, knowing you: theory of mind in AI. *Psychological Medicine* 50, 7 (2020), 1057–1061.

[17] Tim Dant. 2015. In two minds: Theory of Mind, intersubjectivity, and autism. *Theory & Psychology* 25, 1 (2015), 45–62.

[18] Maartje MA De Graaf. 2016. An ethical evaluation of human–robot relationships. *International journal of social robotics* 8, 4 (2016), 589–598.

[19] Hanne De Jaegher. 2023. Seeing and inviting participation in autistic interactions. *Transcultural Psychiatry* 60, 5 (2023), 852–865.

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

[21] Catherine D'ignazio and Lauren F Klein. 2023. *Data feminism*. MIT press.

[22] Mark Dingemanse, Andreas Liesenfeld, Marlou Rasenberg, Saul Albert, Felix K Ameka, Abeba Birhane, Dimitris Bolis, Justine Cassell, Rebecca Clift, Elena Cuffari, et al. 2023. Beyond single-mindedness: A figure-ground reversal for the cognitive sciences. *Cognitive science* 47, 1 (2023), e13230.

[23] Lasse Dissing and Thomas Bolander. 2020. Implementing Theory of Mind on a Robot Using Dynamic Epistemic Logic. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1615–1621. doi:10.24963/ijcai.2020/224 Main track.

[24] Fethiye Irmak Doğan, Sarah Gillet, Elizabeth J. Carter, and Iolanda Leite. 2020. The impact of adding perspective-taking to spatial referencing during human–robot interaction. *Robotics and Autonomous Systems* 134 (2020), 103654. doi:10.1016/j.robot.2020.103654

[25] John Duffy and Rebecca Dorner. 2011. The pathos of" mindblindness": Autism, science, and sadness in" theory of mind" narratives. *Journal of Literary & Cultural Disability Studies* 5, 2 (2011), 201–215.

[26] Emre Erdogan, Frank Dignum, Rineke Verbrugge, and Pınar Yolum. 2024. Resolving Multi-user Privacy Conflicts with Computational Theory of Mind. In *Proceedings of the Second International Workshop on Citizen-Centric Multiagent Systems*. figshare, 22–28.

[27] JA Fodor. 1995. „A Theory of the Child's Theory of Mind", w: Davies, M. i Stone, T, red. *Mental Simulation* (1995), 109–122.

[28] Chris Frith and Uta Frith. 2005. Theory of mind. *Current biology* 15, 17 (2005), R644–R645.

[29] Shaun Gallagher. 2004. Understanding interpersonal problems in autism: Interaction theory as an alternative to theory of mind. *Philosophy, Psychiatry, & Psychology* 11, 3 (2004), 199–217.

[30] Kathrin Gerling, Kay Kender, Katta Spiel, Saskia Van der Oord, Dieter Baeyens, Arno Depoortere, and Maria Aufheimer. 2022. Reflections on Ableism in Participatory Technology Design. In *Mensch und Computer 2022-Workshopband*. Gesellschaft für Informatik eV, 10–18420.

[31] Morton Ann Gernsbacher and Melanie Yergeau. 2019. Empirical failures of the claim that autistic people lack a theory of mind. *Archives of scientific psychology* 7, 1 (2019), 102.

[32] Piotr J. Gmytrasiewicz and Prashant Doshi. 2005. A Framework for Sequential Planning in Multi-Agent Settings. *J. Artif. Intell. Res.* 24 (2005), 49–79. doi:10.1613/jair.1579

[33] David Gollasch, Meinhardt Branig, Kathrin Gerling, Jan Gulliksen, Oussama Metatla, Katta Spiel, and Gerhard Weber. 2023. Designing technology for neurodivergent self-determination: challenges and opportunities. In *IFIP Conference on Human-Computer Interaction*. Springer, 621–626.

[34] Nikolos Gurney and David V. Pynadath. 2022. Robots with Theory of Mind for Humans: A Survey*. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 993–1000. doi:10.1109/RO-MAN53752.2022.9900662

[35] Nikolos Gurney, David V Pynadath, and Volkan Ustun. 2024. Spontaneous theory of mind for artificial intelligence. In *International Conference on Human-Computer Interaction*. Springer, 60–75.

[36] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences* 33, 2-3 (2010), 61–83.

[37] Claire Hughes and Rory T. Devine. 2015. A social perspective on theory of mind. In *Handbook of child psychology and developmental science: Socioemotional processes (7th ed.)*, M. E. Lamb & R. M. Lerner (Ed.). John Wiley & Sons, Inc., 564–609.

[38] Gal A Kaminka. 2013. Curing robot autism: A challenge. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. 801–804.

[39] Steven Kapp. 2019. How social deficit models exacerbate the medical model: Autism as case in point. *Autism Policy & Practice* 2, 1 (2019), 3–28.

[40] Os Keyes. 2020. Automating autism: Disability, discourse, and artificial intelligence. *The Journal of Sociotechnical Critique* 1, 1 (2020), 8.

[41] Stefan Larsson, Mia Liinason, Laetitia Tanqueray, and Ginevra Castellano. 2023. Towards a socio-legal robotics: a theoretical framework on norms and adaptive technologies. *International Journal of Social Robotics* 15, 11 (2023), 1755–1768.

[42] Alan M Leslie. 1991. *The theory of mind impairment in autism: Evidence for a modular mechanism of development?* Basil Blackwell, 63–78.

[43] Alan M Leslie. 1994. ToMM, ToBy, and Agency: Core architecture and domain specificity. *Mapping the mind: Domain specificity in cognition and culture* 29 (1994), 119–48.

[44] Ivan Leudar and Alan Costall. 2009. *Against theory of mind*. Palgrave Macmillan/Springer Nature.

[45] Ivan Leudar, Alan Costall, and Dave Francis. 2004. Theory of mind: a critical assessment. *Theory & Psychology* 14, 5 (2004), 571–578.

[46] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. doi:10.1145/3411764.3445488

[47] Anne E McGuire and Rod Michalko. 2011. Minds between us: Autism, mindblindness and the uncertainty of communication. *Educational Philosophy and Theory* 43, 2 (2011), 162–177.

[48] Andrew N Meltzoff. 1999. Origins of theory of mind, cognition and communication. *Journal of communication disorders* 32, 4 (1999), 251–269.

[49] Andrew N Meltzoff and Alison Gopnik. 2013. *Learning about the mind from evidence: Children's development of intuitive theories of perception and personality*. Vol. 3. Oxford University Press Oxford, UK, 19–34.

[50] Damian EM Milton. 2012. On the ontological status of autism: The 'double empathy problem'. *Disability & society* 27, 6 (2012), 883–887.

[51] Monica Nicolescu, Janelle Blankenburg, Bashira Akter Anima, Mariya Zagainova, Pourya Hoseini, Mircea Nicolescu, and David Feil-Seifer. 2025. Simulation theory of mind for heterogeneous human-robot teams. *Frontiers in Robotics and AI* 12 (2025), 1533054.

[52] Maria LM Patricio and Anahita Jamshidnejad. 2023. Dynamic mathematical models of theory of mind for socially assistive robots. *IEEE Access* 11 (2023), 103956–103975.

[53] Rosalind W. Picard. 1997. *Affective computing*. MIT Press, Cambridge, MA, USA.

[54] Michael Plastow. 2012. 'Theory of mind'II: difficulties and critiques. *Australasian Psychiatry* 20, 4 (2012), 291–294.

[55] Jan Pöppel, Sebastian Kahl, and Stefan Kopp. 2022. Resonating minds—emergent collaboration through hierarchical active inference. *Cognitive Computation* 14, 2 (2022), 581–601.

[56] Daniel J Povinelli and Todd M Preuss. 1995. Theory of mind: evolutionary history of a cognitive specialization. *Trends in neurosciences* 18, 9 (1995), 418–424.

[57] David Premack and Guy Woodruff. 1978. Does the Chimpanzee Have a Theory of Mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[58] Hannes Rakoczy. 2022. Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology* 1, 4 (2022), 223–235.

[59] Jessica Sage Rauchberg. 2022. Imagining a neuroqueer technoscience. *Studies in Social Justice* 16, 2 (2022), 370–388.

[60] Jennifer Renoux, Abdel-Illah Mouaddib, and Simon Le Gloannec. 2015. A decision-theoretic planning approach for multi-robot exploration and event search. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015*. IEEE, 5287–5293. doi:10.

1109/IROS.2015.7354123
[61] Jacquie Ripat and Roberta Woodgate. 2011. The intersection of culture, disability and assistive technology. *Disability and Rehabilitation: Assistive Technology* 6, 2 (2011), 87–96.
[62] Michele Rocha, Heitor Henrique da Silva, Analúcia Schiaffino Morales, Stefan Sarkadi, and Alison R Panisson. 2023. Applying theory of mind to multi-agent systems: A systematic review. In *Brazilian Conference on Intelligent Systems*. Springer, 367–381.
[63] Selma Šabanović. 2010. Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. *International Journal of Social Robotics* 2, 4 (2010), 439–450.
[64] Dana Samson and Ian A Apperly. 2010. There is more to mind reading than having theory of mind concepts: New directions in theory of mind research. *Infant and Child Development* 19, 5 (2010), 443–454.
[65] Ştefan Sarkadi, Alison R Panisson, Rafael H Bordini, Peter McBurney, Simon Parsons, and Martin Chapman. 2019. Modelling deception using theory of mind in multi-agent systems. *AI Communications* 32, 4 (2019), 287–302.
[66] Guillaume Sarthou. 2023. Overworld: Assessing the geometry of the world for human-robot interaction. *IEEE Robotics and Automation Letters* 8, 3 (2023), 1874–1880.
[67] Brian Scassellati. 2002. Theory of mind for a humanoid robot. *Autonomous Robots* 12, 1 (2002), 13–24.
[68] Katie Seaborn, Giulia Barbareschi, and Shruti Chandra. 2023. Not only WEIRD but "uncanny"? A systematic review of diversity in human–robot interaction research. *International Journal of Social Robotics* 15, 11 (2023), 1841–1870.
[69] Johanna Seibt, Christina Vestergaard, and Malene F Damholdt. 2020. Sociomorphing, not anthropomorphizing: towards a typology of experienced sociality. In *Culturally sustainable social robotics*. ioS Press, 51–67.
[70] Wes Sharrock and Jeff Coulter. 2004. ToM: A critical commentary. *Theory & Psychology* 14, 5 (2004), 579–600.
[71] Elizabeth Sheppard, Dhanya Pillai, Genevieve Tze-Lynn Wong, Danielle Ropar, and Peter Mitchell. 2016. How easy is it to read the minds of people with autism spectrum disorder? *Journal of autism and developmental disorders* 46, 4 (2016), 1247–1254.
[72] Ashley Shew. 2023. *Against technoableism: rethinking who needs improvement*. WW Norton & Company.
[73] Maayan Shvo, Toryn Q. Klassen, and Sheila A McIlraith. 2022. Resolving Misconceptions about the Plans of Agents via Theory of Mind. In *Proceedings of the 32nd International Conference on Automated Planning and Scheduling (ICAPS)*.
[74] David Smukler. 2005. Unauthorized minds: how "theory of mind" theory misrepresents autism. *Mental Retardation* 43, 1 (2005), 11–24.
[75] Katta Spiel, Christopher Frauenberger, Os Keyes, and Geraldine Fitzpatrick. 2019. Agency of Autistic Children in Technology Research—A Critical Literature Review. *ACM Trans. Comput.-Hum. Interact.* 26, 6, Article 38 (nov 2019), 40 pages. doi:10.1145/3344919
[76] Anna Stubblefield. 2012. Knowing Other Minds: Ethics and Autism. *The Philosophy of Autism* (2012), 143.
[77] Laetitia Tanqueray and Stefan Larsson. 2023. What norms are social robots reflecting? a socio-legal exploration on hri developers. In *Social Robots in Social Institutions*. IOS Press, 305–314.
[78] Federico Tavella, Federico Manzi, Samuele Vinanzi, Cinzia Di Dio, Davide Massaro, Angelo Cangelosi, and Antonella Marchetti. 2024. Towards a computational model for higher orders of Theory of Mind in social agents. *Frontiers in Robotics and AI* 11 (2024), 1468756.
[79] J Gregory Trafton, Nicholas L Cassimatis, Magdalena D Bugajska, Derek P Brock, Farilee E Mintz, and Alan C Schultz. 2005. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 35, 4 (2005), 460–470.
[80] Thalia Wheatley, Mark A Thornton, Arjen Stolk, and Luke J Chang. 2024. The emerging science of interacting minds. *Perspectives on Psychological Science* 19, 2 (2024), 355–373.
[81] Rua Mae Williams. 2019. Metaeugenics and metaresistance: From manufacturing the 'includeable body'to walking away from the broom closet. *Canadian Journal of Children's Rights/Revue canadienne des droits des enfants* 6, 1 (2019), 60–77.
[82] Rua M Williams. 2021. I, Misfit: Empty Fortresses, Social Robots, and Peculiar Relations in Autism Research. *Techne: Research in Philosophy & Technology* 25, 3 (2021).
[83] Rua Mae Williams, Louanne E. Boyd, and Juan E. Gilbert. 2023. Counterventions: a reparative reflection on interventionist HCI. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://api.semanticscholar.org/CorpusID:258217100
[84] 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 (1983), 103–128.
[85] Scott Cheng-Hsin Yang, Tomas Folke, and Patrick Shafto. 2025. The inner loop of collective human–machine intelligence. *Topics in cognitive science* 17, 2 (2025), 248–267.
[86] Melanie Yergeau and Bryce Huebner. 2017. Minding theory of mind. *Journal of Social Philosophy* 48, 3 (2017).
[87] Elaine Kit Ling Yeung, Ian A Apperly, and Rory T Devine. 2024. Measures of individual differences in adult theory of mind: A systematic review. *Neuroscience & Biobehavioral Reviews* 157 (2024), 105481.
[88] Chuang Yu, Baris Serhan, and Angelo Cangelosi. 2024. Top-tom: Trust-aware robot policy with theory of mind. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7888–7894.
[89] Dan Zahavi and Josef Parnas. 2003. Conceptual problems in infantile autism research: Why cognitive science needs phenomenology. *Journal of consciousness studies* 10, 9-10 (2003), 53–71.

Received 2025-10-17; accepted 2025-11-21