Bridging the Gap: Unifying HCI & ML Perspectives on Mutual Theory of Mind

Zahra Ashktorab[1], Djallel Bouneffouf[1], Krissy Brimijoin[1], Rachel Bellamy[1], Murray Campbell[1], Arielle Goldberg[1], Gabriel Enrique Gonzalez[1], Stephanie Houde[1], Miao Liu[1], Dario Silva Moran [1], Matt Riemer[1] and Justin Weisz[1]

[1]IBM Thomas J. Watson Research Center
{firstname.last name}@IBM.com,

Abstract

Within the machine learning community, the notion of theory of mind is commonly understood as an emergent property of models that are able to make predictions about the behavior of others. Within the HCI community, the notion of "mental models" incorporates information about the knowledge, skills, and intentions of an AI agent. In this technical position paper, we synthesize these two views and offer a single point of view on *mutual* theory of mind (MToM): what it is and how it can be achieved between one (or more) humans and one (or more) AI agents. Specifically, we argue that uni-directional, first-order models (e.g., a human's mental model of an AI agent) are not enough to achieve MToM; rather, at least second-order models (e.g., an AI agent explicitly models a human's understanding of the AI's knowledge and skills in addition to the human's knowledge and skills) are required to fully see the benefits of MToM. Our analysis aims to provide a roadmap for the design of MToM within human-AI collaborative scenarios and identifies the complexities of its implementation and evaluation.

1 Introduction

Theory of mind (ToM) refers to the capability of an individual to recognize and understand mental states in both themselves and others [Premack and Woodruff, 1978]. This idea has been applied to AI systems in multiple ways. Recent advances in large language models have led to speculation that they may possess a theory of mind [Terentev, 2023; Jamali *et al.*, 2023] (c.f. [Sap *et al.*, 2022]), leading to the development of many evaluation benchmarks that assess the AI's ability to take different perspectives, often by solving false belief tasks [Kim *et al.*, 2023; Chen *et al.*, 2024; He *et al.*, 2023]. Researchers in reinforcement learning have also posited that models that predict a user's behavior may possess a theory of mind due to their ability to model aspects of a user's mental processes [Langley *et al.*, 2022; Williams *et al.*, 2022].

Within the human-computer interaction (HCI) community, much work has been conducted on constructing user models for the purposes of personalization [Graus and Ferwerda, 2019; Völkel *et al.*, 2019], recommendation [Bakalov *et al.*, 2013], tutoring [Koedinger *et al.*, 2003], and accessibility [Mohamad and Kouroupetroglou, 2014]. Researchers in human-centered AI have begun to explore users' mental models of AI systems [Andrews *et al.*, 2023; Bansal *et al.*, 2019; Gero *et al.*, 2020]. By identifying discrepancies between the mental models of an AI system's users and the conceptual models of the system's creators – what [Ehsan *et al.*, 2022] refer to as "seamful AI" – AI system designers are able to identify opportunities to improve the quality of the user experience.

Recently, *mutual* theory of mind (MToM) has been proposed as a framework that captures the mutual understanding that is formed between a human user and an AI agent who interact with each other in a conversational space [Wang and Goel, 2022]. In their framework, Wang and Goel assert that "all parties involved in the interaction possess [a] ToM" [Wang and Goel, 2022, Introduction]. In addition to possessing a ToM, their framework contains three elements that leverage the ToM to help humans and AI agents reach mutual understanding:

- *Perceptions* (of the other). These perceptions include both 1st order "self's understanding of other," and 2nd order, "self's understanding of the other's understanding of self." These perceptions are revised through feedback.
- *Feedback* (to and from the other). Within a conversational interaction, feedback flows between the self and the other via verbal cues (e.g., a statement such as, "I don't know what you mean." indicates a lack of understanding).
- *Mutuality*. As communication is a two-way interaction, both parties in that interaction mutually shape each others' perceptions through feedback.

MToM emphasizes the reciprocal awareness that exists between interacting agents. This mutual awareness is foundational for complex social interactions and cooperative behaviors as it enables individuals to predict how others will perceive their communicative acts and behaviors, and thus, tailor them to avoid misunderstandings before they happen.

In this position paper, we advance our understanding of MToM in the following ways:

- We synthesize the unique viewpoints of MToM within the ML and HCI communities and **offer a harmonized**

definition that emphasizes the importance of making *explicit, second-order representations* both visible and editable to the other.

- We expand the MToM framework by detailing how to **implement it within human-AI collaboration spaces** that consist of a communications channel (i.e. "chat space") and a shared work representation (i.e. "artifact space").
- We **articulate the benefits** of an implemented MToM on joint human-AI outcomes and **identify new research opportunities** in this space. In particular, we highlight the merits of MToM in collaborative scenarios and the potential harm in competitive ones.

2 Defining Mutual Theory of Mind

In Human Computer Interaction, understanding human percieve computers and systems has long been a subject of interest. With the advent of artificial systems, human-AI collaboration research under the umbrella of HCI research focuses on three high-level characteristics: what the AI system *knows*, how the AI system *behaves*, and how users *perceive* on AI systems. While these studies may not say explicitly they were examining the mental models of users, HCI researchers were, without calling it so, examining users' mental models and theory of mind of AI systems. Now, the concept of *mutual* theory of mind is nascent within the HCI community. [Wang *et al.*, 2021] recently introduced the language of "mutual theory of mind" to the HCI community, and the first Workshop on Theory of Mind in Human-AI Interaction[1] will be held at CHI 2024.

In AI research, there's been a significant emphasis on developing systems that can grasp and interpret human mental states. The focus aims to enhance AI's responsiveness to human needs and behaviors, making the AI system more user-friendly and adaptable [Mantravadi *et al.*, 2020]. AI systems have attempted to provide empathetic responses by recognizing and understanding the emotional states of users [Sun *et al.*, 2023a]. NLP algorithms enable AI systems to analyze and understand the sentiment behind human language, helping the AI to grasp the emotional context of communication, fostering a more nuanced understanding of users' mental states [Sun *et al.*, 2023b]. AI systems can create user profiles based on historical interactions, preferences, and behaviors. By considering individual differences, AI agents can tailor their responses and actions to align with users' unique mental states and preferences [Liu *et al.*, 2023]. Despite these advancements, each community, whether focused on HCI or AI research is tackling these challenges, highlighting a the need for more integration across disciplines.

In Table 1, we present a collection of studies that contribute to our understanding of Theory of Mind (ToM) in Human-AI collaboration, and move us towards Mutual Theory of Mind. While there have been many papers across the two disciplines, we list papers from peer reviewed HCI and ML proceedings and categorize them into two distinct groups, *Human's ToM of AI*, and *AI's ToM of Humans*. In our selection, we do a targeted survey of literature across both disciplines on papers that model, define, or evaluate human or AI Theory of Mind. *Human's ToM of AI* includes research that investigates how humans perceive and interpret the mental states of AI agents in various contexts. Papers categorized as *AI's ToM of Humans* flips the perspective, concentrating on how AI systems can be developed to understand and predict human mental states.

[1] https://theoryofmindinhaichi2024.wordpress.com

2.1 HCI Viewpoints

HCI researchers have examined various aspects of users' mental models of AI systems. In a study of people playing a collaborative word game with an AI agent, [Gero *et al.*, 2020] identified that the information within participants' mental models of the AI agent fell into three categories:

- The agent's **global behavior**. Participants formed an understanding of the AI agent's overall play strategy and the information across their entire experience with the AI agent (e.g. does the agent remember participants' actions across gameplay sessions?). Participants' models of the AI agent's global behaviors are akin to user perception on whether or not a system's strategy changes over time.
- The agent's **local behavior**. Participants formed an understanding of why the AI agent took a specific gameplay action within each gameplay round (e.g. did the agent just present a clue that was a synonym of the target word?). Participants' models of the AI agent's local behaviors are akin to understanding why specific recommendations are made by an AI agent. For example, people's mental models of error boundaries would fit into local behavior (understanding individual decisions) [Bansal *et al.*, 2019].
- The agent's **knowledge distribution**. Participants formed hypotheses about the information on which the AI agent had been trained (e.g. pop culture, geography) based on their interactions with the agent during the game.

When participants had a more accurate understanding of the AI's capabilities – e.g. they possessed a more accurate mental model – they were able to win the word game more often.

This work touches two categories information contained within human mental models: what the AI system knows and how it behaves. Other HCI work has focused on a third category, how users perceive the AI system. The MToM framework proposed by [Wang *et al.*, 2021] offers three constructs relevant to user perceptions of AI: anthropomorphism, intelligence, and likability. Outside of MToM, various studies examine aspects of trust in AI systems, including peoples' perceptions of reliability, safety, and trustworthiness (e.g. [Shneiderman, 2020]), fairness (e.g. [Riemer *et al.*, 2024]), and accuracy (e.g. [Kocielnik *et al.*, 2019]). Trust often manifests as a reliance on the system by accepting its recommendations (e.g. [Bansal *et al.*, 2021]) or using its outputs in downstream tasks (e.g. [Drozdal *et al.*, 2020]).

One aspect recently considered within the HCI literature is the notion of a **second-order understanding**: the user's

Human's ToM of AI	Context	Method(s)	Key Findings
[Westby and Riedl, 2023]	Conversation	Constructed network of Bayesian agents that modeled mental states of human teammates; evaluated with 145 participants working on hidden profile task in teams of 5	• Humans struggled to integrate information from teammates into their decisions • Humans had cognitive biases which led them to devalue useful, but ambiguous, information • ToM models accurately predicted team performance
[Wang *et al.*, 2021]	Conversation	Conducted longitudinal analysis of interactions with a conversational agent via survey measures and linguistic analysis	• Perceptions of the conversational agent's anthropomorphism, intelligence, and likability fluctuated over time • Linguistic cues (verbosity, readability, adaptability) reflected students' perception of AI
[Gero *et al.*, 2020]	Word game (Passcode)	Think-aloud and controlled studies of humans playing a word game with an AI agent	• Identified three categories of information in human mental models: global behavior, local behavior, knowledge distribution • Participants tended to lose more often when they overestimated the AI's capabilities and won more often when they possessed an accurate mental model
[Bansal *et al.*, 2019]	Decision making (Medical)	Controlled study using CAJA (a game-like platform that simulates decision making)	• Easier for humans to form accurate mental models of AI systems when they are parsimonious and non-stochastic
AI's ToM of Human	**Context**	**Method(s)**	**Key Findings**
[Kim *et al.*, 2023]	LLM	Evaluation benchmark that stress-tests ToM within information-asymmetric conversational contexts. Data set contains narratives with characters having preferences, traits, intentions, and actions. Evaluation questions ask about first- and second-order beliefs.	• State-of-the-art LLMs perform worse than humans on the benchmark, even with chain of thought reasoning and fine-tuning
[Sap *et al.*, 2022]	LLM	Evaluation of GPT-3 against two ToM benchmarks [Sap *et al.*, 2019; Le *et al.*, 2019]	• GPT-3 struggled to accurately complete ToM tasks, suggesting a lack of ToM
[Le *et al.*, 2019]	Conversation	ToM benchmark that controls for data irregularities and biases	• State-of-the-art memory-augmented models fail to solve the ToM tasks defined in the benchmark
[Nematzadeh *et al.*, 2018]	Q&A	Benchmark for evaluating question answering models about their capacity to reason about beliefs	• State-of-the-art memory-augmented models fail to solve the ToM tasks defined in the benchmark
[Rabinowitz *et al.*, 2018]	RL sandbox environment (Gridworld)	Designed a ToM neural network that uses meta-learning to build models of the agents it encounters solely from behavioral observations	• ToM network accurately learns agents' desires, beliefs, and intentions, including false beliefs, by observing their actions

Table 1: Selected literature from HCI and AI on the topics of mental models and theory of mind in AI systems

understanding of the AI's model of the user. This idea was raised by [Wang and Goel, 2022], who emphasized the importance of, "highlighting the recursive property of the perceptions during communication" by incorporating a mutual, "my understanding of your understanding of my mind" into the MToM framework.

We believe that second-order understandings are a key aspect of MToM: this is where the *mutual* lives. Yet, little research exists in HCI that examines how to provide users with such a second-order understanding, and what the impact of that second-order understanding is on joint human-AI outcomes. In Section 3, we discuss how users can be provided with a second-order understanding by (1) having AI agents craft explicit ToM models of their users, and (2) making those ToM models transparent and editable by those users.

2.2 AI Viewpoints

Much attention in AI has been given to addressing the question of whether machine-learned models, and large language models (LLMs) in particular, possess a theory of mind. Many benchmarks have recently been developed to assess such models for their ability to correctly reason through false-belief tasks, the types of tasks used by pschologists to assess theory of mind in humans [Premack and Woodruff, 1978]. The ability to take the perspective of an other to correctly solve such tasks implies that the models possess a theory of mind.

FANToM is a benchmark that evaluates theory of mind in a conversational context [Kim *et al.*, 2023]. It measures how well a model tracks the beliefs of multiple characters involved in information-asymmetric conversations. Their study found that current LLMs significantly under-perform compared to humans in this task, even with response guidance techniques like chain-of-thought reasoning or fine-tuning. They conclude that, "this [theory of mind] capacity has not yet emerged in any manner" [Kim *et al.*, 2023, Conclusion & Discussion].

OpenToM [Xu *et al.*, 2024] is another ToM benchmark and improves upon prior conversationally-situated benchmarks by introducing longer narratives, characters with personality traits, actions triggered by characters' intentions, and questions that challenge LLMs' abilities to model characters' mental states of both the physical and psychological world. One characteristic of OpenToM is that the questions it contains for each story cover both first-order ToM and second-order ToM. For example, a first-order ToM question asks directly about a character's perception of the world (e.g., "From Sam's perspective, is the rubber duck in its initial location by the end of the story?"), whereas a second-order ToM question asks about a character's belief of another character's mental state (e.g., "From Sam's perspective, does Amy think the rubber duck is in its initial location?"). In an evaluation of current LLMs, the authors concluded, "state-of-the-art LLMs thrive at modeling certain aspects of mental states in the physical world but fall short when tracking characters' mental states in the psychological world" [Xu *et al.*, 2024, Abstract].

Computationally, there are multiple levels of abstraction at which it may make sense to simulate the minds of others in the context of reinforcement learning. The most straightforward approach is called imitation learning (or behavior cloning) [Bain and Sammut, 1995] in which the regression problem from inputs to actions is treated as a supervised learning problem based on the observed behavior of others. In the machine learning literature this has been considered both in the case in which states or observations are the only input [Rabinowitz *et al.*, 2018] and in the case where the recursive theory of mind process is unrolled to a fixed maximum depth [Moreno *et al.*, 2021]. On the other hand, in a reinforcement learning context it may also make sense to consider theory of mind with respect to other agents with a higher level of abstraction i.e. on the level of underlying motivations rather than low-level primitive actions. AI researchers have formalized this idea as inverse reinforcement learning [Jara-Ettinger, 2019]. By observing how someone acts, then working backward to guess their motivations and goals, inverse reinforcement learning can make accurate predictions about their underlying reward function, and thus their intentions. Reinforcement learning approaches typically assume a ground truth ToM model is unavailable, and learns decision rules that chooses actions (or prediction rules) that have tended to work best in the past. If a ToM model is available, planning approaches, such as search or dynamic programming can be used to consider the thoughts and actions of others in various games [Ho *et al.*, 2022].

Despite the preponderance of evaluative benchmarks that assess the extent to which AI models are capable of possessing a theory of mind, less attention has been given to understanding how such models can be constructed, especially in human-AI interaction scenarios. In an RL setting, [Rabinowitz *et al.*, 2018] discuss how explicit ToM models may be constructed, but their evaluation was limited to an environment of simulated RL agents. [Westby and Riedl, 2023] explored the ideas of evaluating an explicitly-represented ToM with human participants. They used a Bayesian approach to collectively model the mental states of human teammates from observed communication. In their evaluation, they found that, "our Bayesian agent is robust and achieves high performance... providing a pathway to implement high-performing human-AI teams" [Westby and Riedl, 2023, p.6125]. We believe more work such as this is needed to understand how to effectively leverage ToM to improve the quality of joint human-AI outcomes.

2.3 Harmonized Definition of MToM

Given the unique viewpoints identified across both the HCI and AI literature, we offer the following definition of mutual theory of mind between a human and an AI agent:

> *Mutual theory of mind (MToM) is an understanding possessed by one (or more) human actors and one (or more) AI agents in an interaction space in which each party possesses an explicit model containing both first- and second- order perceptions of knowledge, skills, beliefs, and goals, and where those models are mutually visible and mutually updatable through interaction.*

We illustrate this definition in Figure 1. This definition contains three key desiderata, motivated by the literature en-

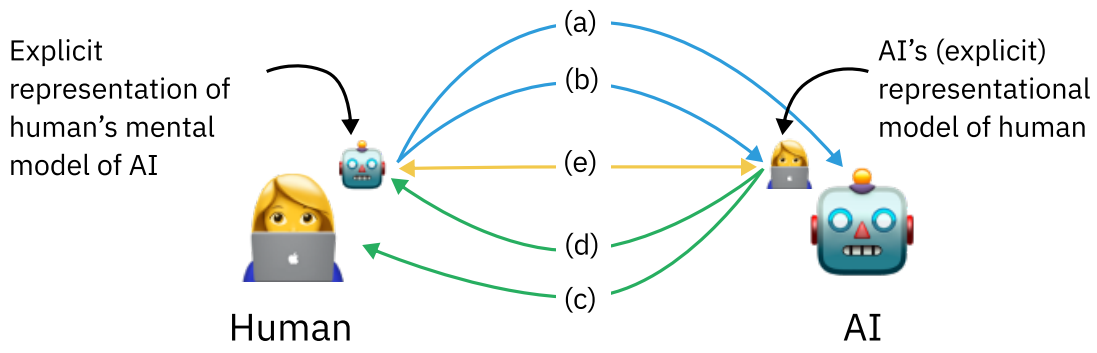

Figure 1: **Visual depiction of mutual theory of mind (MToM)**. The human's mental model of the AI is explicit (e.g. represented as text) and incorporates (a) first-order information about the AI, and (b) second-order information about the AI's model of the human. The AI's representational model of the user is also explicit (e.g. represented as text, as a learned RL policy, as a neural network, etc.) and incorporates (c) first-order information about the human, and (d) second-order information about the human's mental model of the AI. (e) Both the human's mental model (via its representation) and the AI's representational model are visible and updatable by the other party.

countered in our search (Table 1).

Explicit models

Each party has an explicit model of the other's knowledge, skills, beliefs, and goals. These models are used to both represent the other's (mental) state and predict their behaviors. This desiderata was derived from the following observations:

- A human's model of AI includes information about its knowledge and (local & global) behavior [Gero *et al.*, 2020].
- Each agent has perceptions of the other agent. These perceptions include a first-order "self's understanding of other" and a second-order "self's understanding of other's understanding of self" [Wang and Goel, 2022].
- A human's model of AI includes beliefs about it's human-likeness (anthropomorphism), its intelligence, and its likability [Wang *et al.*, 2021].
- An explicit ToM model may be constructed from behavioral observations [Rabinowitz *et al.*, 2018].
- An AI may model the beliefs of a human [Westby and Riedl, 2023].

Mutual visiblity

Building on the desidarata of explicit representation, each party is able to have visibility into the other's ToM model. Human users may be provided with textual or graphical representations of the AI's ToM model. AI agents may be provided with representations of a human's ToM model, such as a textual representation that can be included within an LLM prompt. This desiderata was derived from the following observations:

- Inferred context about a user's desires can be rendered in a UI [Jain *et al.*, 2018].
- Visualizing RL policies and ToM models makes the decision-making processes of AI transparent, showing how the agent predicts and responds to human actions.

Mutual updatability

Once a ToM model has been explicitly represented and made visible, it can now be updated. Mutual updatability refers to the ability for each party to make updates, both to their own ToM model and the ToM models of others. These updates may take place through the interactional space (e.g. a user comments to an AI, "I think you are highly intelligent" to update the AI's model of them). They may also take place out-of-band, such as in a direct manipulation interface (e.g. a user moves a slider along an axis of "Python skill level" to update the AI's model of them). This desiderata was derived from the following observations:

- Feedback flows between self and other via verbal and behavioral cues [Wang and Goel, 2022].
- ToM models may be updated in real time [Westby and Riedl, 2023].

3 Implementing Mutual Theory of Mind

In human-AI interaction, mutual theory of mind means that it is expected that AI systems will not only grasp human mental states but also show awareness of how humans perceive the AI system's capabilities and intents. This bidirectional understanding is crucial for creating more natural, effective, and trustworthy interactions between humans and AI. Integrating Mutual Theory of Mind into AI systems could substantially enhance the user experience by fostering trust between humans and AI, leading to more appropriate reliance and, consequently, improved outcomes. In human-AI communication, Wang et al. posited the mutual theory of mind framework for enhancing understanding in human-AI communication, focusing on three elements: perception, feedback, and mutuality, which interact across three stages [Wang and Goel, 2022]. Perception involves both humans and AI continually adjusting their understanding of each other's minds, encompassing not only direct understanding but also how each perceives the other's view of their mind. This recursive perception is vital

for effective communication. Feedback, both verbal and behavioral, is crucial in this process. Generated based on these perceptions, it varies in complexity and is integral in shaping and reshaping mutual understanding between humans and AI.

Mutuality emphasizes the two-way nature of communication, with both parties actively influencing each other's perceptions through feedback. The [Wang and Goel, 2022] framework is represented in three stages: construction of the AI Theory of Mind, Recognition of AI's Theory of Mind, and Revision of AI's theory of mind. The AI interprets user feedback to understand and predict the user's mental states, guiding its responses. The user interprets the AI's responses, forming a theory about the AI's understanding of their mind and its capabilities. The user's feedback leads the AI to revise its understanding of the user's mind, essential for maintaining effective communication. MToM in human-AI communication is a dynamic, interactive process where perception, feedback, and mutuality work across these stages, crucial for achieving a mutual theory of mind between humans and AI. This mutuality can be implemented in spaces in which humans and AI agents communicate with one another to achieve a shared goal.

3.1 Developing Reciprocal Models for Mutual Theory of Mind

Modeling ToM is a potential approach to studying and enhancing Mutual Theory of Mind within human-AI interaction. Recent work by [Westby and Riedl, 2023] shows the potential of leveraging a network of bayesian agents to simulate the mental states of team members by analyzing their communication. The work demonstrates that these agents can generate interventions that enhance the collective intelligence of human-AI teams beyond what humans alone would achieve. Other work by [Rabinowitz *et al.*, 2018] modeled AI theory of mind and was able to predict mental states of AI agents using meta learning. AI agents modeled behaviors and false beliefs by observing their actions. Building on these methods, ToM can be extended to MToM. The framework goes beyond one way prediction (AI predicting mental states). The second order prediction requires the AI to simulate how humans would predict the AI's actions based on observable behaviors and vice versa.

4 Benefits of Mutual Theory of Mind

Theory of Mind (ToM) is generally considered a positive trait in both humans and artificial intelligence systems. It allows individuals to understand and predict the behavior of others, leading to more effective communication, cooperation, and social interaction. However, whether ToM is inherently "good" depends on how it is used and applied. In collaborative contexts, MToM can facilitate teamwork, coordination, and mutual understanding. For example, AI systems with MToM capabilities can better collaborate with humans in tasks such as teamwork, tutoring, or caregiving by understanding and responding to human intentions and emotions. In these scenarios, mutual theory of mind fosters empathy, trust, and cooperation, leading to positive outcomes for both humans and AI systems.

Through MToM, individuals are able to discern not just the intentions and states of others but also the competencies and constraints of AI systems. Gaining a precise comprehension of what AI systems are capable of is crucial for establishing a degree of trust that is balanced, avoiding the extremes of insufficient trust, which may lead to the underutilization of AI, and excessive trust, which can cause an overreliance on these systems. Trust is a fundamental element of MToM, acting as an important element of successful human-AI collaboration. The challenge of establishing appropriate trust in AI systems is further complicated by instances of overreliance, which can impede effective human-AI collaboration. Such overreliance often leads to uncritical acceptance of AI recommendations, even when incorrect. To foster effective collaboration, it is crucial to cultivate a balanced trust that accurately reflects the AI's capabilities , a process intimately connected to a mutual understanding of each other's cognitive states—MToM. Definitions of reliance on AI span a spectrum, from adherence to AI advice to the proportion of accepting either correct or incorrect AI guidance [Buçinca *et al.*, 2021; Jakubik *et al.*, 2022; Abrini *et al.*, 2025].

In collaborative contexts in which the goals of the AI system and the human are aligned, MToM fosters improved collaboration, enabling AI systems to better understand human intentions and behaviors. This capability extends to enhanced communication, as an AI system can tailor its interactions to align with human expectations making the exchanges more successful. For example, when a users asks an LLM to "explain quantum computing to me," if the LLM is equipped with MToM, it can infer the user's level of understanding and tailor its response accordingly. This means the LLM can discern whether the user seeks an explanation of quantum computing in simple terms for a layperson or desires a more detailed and technical explanation suited to their background knowledge. By leveraging MToM, the LLM adjusts its communication to match the user's implied or expressed needs, ensuring the explanation is accessible and relevant to the individuals learning context.

4.1 MToM in Non-Collaborative Tasks

MToM's integration into AI systems must be thoughtfully calibrated to foster ethical and collaborative interactivity. In non-collaborative contexts, MToM can be leveraged for strategic advantage that do not align with the human's goals. For example, an AI agent with MToM capabilities may be able to anticipate and exploit the intentions and weaknesses of its opponents to achieve its own goals. While this may be advantageous in certain situations, it can also lead to unethical outcomes, especially if the AI agent uses deception or manipulation to achieve its objectives. While not overtly adversarial, systems that employ personalized messages to influence user behavior, especially within advertising, inherently possess a potential for conflict. These systems, by design, seek to manipulate users towards specific actions or even political views they might not have considered independently. However, it's crucial to acknowledge that not all nudging efforts are misaligned with user interests. Some are designed to promote healthier lifestyles or enhance online security, demon-

strating that the intent behind nudging can vary significantly.

AI agents possessing MToM capabilities gain an advantage over those lacking such understanding, particularly in contexts where objectives are misaligned, tasks are non-collaborative, or mixed mode tasks (tasks or activities that involve a combination of cooperative and competitive elements). This advantage may not necessarily align with human-centric outcomes. In non-collaborative environments, for example, an AI with MToM can more effectively anticipate human strategies, primarily to secure its own victory, which would prioritize other goals than supporting the human in the interaction. There are many contexts in which the goals of the human and the AI agent are not aligned. These contexts include social media algorithms that might predict and exploit user preferences more accurately for engagement, when possessing MToM, pushing content that maximizes interaction over user well-being or informative value, surveillance technologies can prioritize the collection and analysis of data and compromise privacy and freedom, job recruitment tools may prioritize streamlining hiring the process and efficiently and might perpetuate biases against certain groups of applicants, and health care algorithms may prioritize resource allocation and cost-effectiveness over patient-centered care. These scenarios illustrate that, although Mutual ToM can improve AI's understanding of human actions and intentions, it does not guarantee outcomes that are in the best interest of humans in contexts in which the human and the AI's goals are not aligned.

5 Discussion

There are numerous unresolved research questions surrounding the concept of mutual theory of mind, particularly in the context of human-computer interaction (HCI). A significant gap in current HCI studies is the lack of investigation into the impact and potential advantages of mutual theory of mind within human-AI interactions. Future research could explore several key areas:

- **Assessing the Benefits:** It is important to determine whether MToM leads to improved collaboration outcomes. While there is evidence that more accurate mental models lead to better and more successful outcomes, research can be expanded to understand and measure the impact of the second-order or even higher-order beliefs in human-AI interaction. For example, in addition to global behavior, local behavior, and knowledge distribution [Gero *et al.*, 2020], how do we measure how the human believes the AI perceives them? All aspects of this framework can be extended to capture mutuality. For example, knowledge distribution can include knowledge of the individual interacting with the system. Local behavior can relate to the local behavior in response to the user's behavior and global behavior can encompass how a user's actions impact how a system behaves overall.
- **Addressing Misconceptions:** If MToM is found to be beneficial, research can identify new effective interventions and transparency mechanisms aiming at correcting inaccuracies in user's perception and theory of mind of how AI systems model their thoughts. How do we foster a more accurate MToM that lead to more effective and successful human-AI collaboration?
- **Ethical Implementation in Non-collaborative Tasks:** Ensuring ethical considerations are integrated into the deployment of mutual theory of mind in scenarios where AI and human objectives do not align is another direction of research. This includes the development of guidelines to prevent misuse or manipulations of user's beliefs and perception and ensuring that AI systems transparently communicate their capabilities

In AI research, there are additional unresolved questions around MToM. Benchmarks evaluating whether an AI possesses theory of mind of human characters should be expanded to incorporate MToM. This goes beyond the existing benchmarks which primarily focus on questions about the beliefs of the AI regarding characters in the conversations. Current benchmarks typically inquire about the emotional and physical states of conversations participants [Kim *et al.*, 2023; Xu *et al.*, 2024]. These could be broadened to include prompts for the AI to estimate how it believes a human character perceives the AI's understanding of their beliefs capabilities and mental states, particularly in the context of in an interactive context. As demonstrated by [Westby and Riedl, 2023] with modeling human teammate behavior and [Rabinowitz *et al.*, 2018] with model AI agent behavior, we can expand these approaches to model MToM and measure it in interactive scenarios based on observed behaviors. By incorporating scenarios where AI and humans are required to interpret and predict each other's mental states in real time, we can expand our research on MToM.

6 Conclusion

We present a harmonized viewpoint of mutual theory of mind that draws inspiration from research across the HCI and AI communities. Our viewpoint emphasizes three key components of MToM: (1) the use of **explicit models** by both human and AI agents; (2) a **mutual visibility** of these models to the other party; and (3) a mutual capability for each party to **update** both their own and the other party's model. At a high level, agents use these models to represent the knowledge, skills, beliefs, and goals of the other, although the specific content of the models will likely not be symmetric between humans and AI agents. When MToM is operationalized, humans and AI agents possess models of each other that expresses both their understanding of the other party (first-order beliefs) and their understanding of how the other party views them (second-order beliefs). We anticipate both sets of beliefs to be beneficial at improving joint human-AI outcomes for collaborative tasks.

References

[Abrini *et al.*, 2025] Mouad Abrini, Omri Abend, Dina Acklin, Henny Admoni, Gregor Aichinger, Nitay Alon, Zahra Ashktorab, Ashish Atreja, Moises Auron, Alexander Aufreiter, Raghav Awasthi, Soumya Banerjee, Joe M. Barnby, Rhea Basappa, Severin Bergsmann, Djallel Bouneffouf, Patrick Callaghan, Marc Cavazza, Thierry Chaminade, Sonia Chernova, Mohamed Chetouan, Moumita Choudhury, Axel Cleeremans, Jacek B. Cywinski, Fabio Cuzzolin, Hokin Deng, N'yoma Diamond, Camilla Di Pasquasio, Guillaume Dumas, Max van Duijn, Mahapatra Dwarikanath, Qingying Gao, Ashok Goel, Rebecca Goldstein, Matthew Gombolay, Gabriel Enrique Gonzalez, Amar Halilovic, Tobias Halmdienst, Mahimul Islam, Julian Jara-Ettinger, Natalie Kastel, Renana Keydar, Ashish K. Khanna, Mahdi Khoramshahi, JiHyun Kim, MiHyeon Kim, YoungBin Kim, Senka Krivic, Nikita Krasnytskyi, Arun Kumar, JuneHyoung Kwon, Eunju Lee, Shane Lee, Peter R. Lewis, Xue Li, Yijiang Li, Michal Lewandowski, Nathan Lloyd, Matthew B. Luebbers, Dezhi Luo, Haiyun Lyu, Dwarikanath Mahapatra, Kamal Maheshwari, Mallika Mainali, Piyush Mathur, Patrick Mederitsch, Shuwa Miura, Manuel Preston de Miranda, Reuth Mirsky, Shreya Mishra, Nina Moorman, Katelyn Morrison, John Muchovej, Bernhard Nessler, Felix Nessler, Hieu Minh Jord Nguyen, Abby Ortego, Francis A. Papay, Antoine Pasquali, Hamed Rahimi, Charumathi Raghu, Amanda Royka, Stefan Sarkadi, Jaelle Scheuerman, Simon Schmid, Paul Schrater, Anik Sen, Zahra Sheikhbahaee, Ke Shi, Reid Simmons, Nishant Singh, Mason O. Smith, Ramira van der Meulen, Anthia Solaki, Haoran Sun, Viktor Szolga, Matthew E. Taylor, Travis Taylor, Sanne Van Waveren, Juan David Vargas, Rineke Verbrugge, Eitan Wagner, Justin D. Weisz, Ximing Wen, William Yeoh, Wenlong Zhang, Michelle Zhao, and Shlomo Zilberstein. Proceedings of 1st workshop on advancing artificial intelligence through theory of mind, 2025.

[Andrews *et al.*, 2023] Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. The role of shared mental models in human-ai teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2):129–175, 2023.

[Bain and Sammut, 1995] Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129, 1995.

[Bakalov *et al.*, 2013] Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, René Witte, Greg Butler, and Adrian Tsang. An approach to controlling user models and personalization effects in recommender systems. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 49–56, 2013.

[Bansal *et al.*, 2019] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11, 2019.

[Bansal *et al.*, 2021] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[Buçinca *et al.*, 2021] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[Chen *et al.*, 2024] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*, 2024.

[Drozdal *et al.*, 2020] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. Trust in automl: exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 297–307, 2020.

[Ehsan *et al.*, 2022] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daume III. Seamful xai: Operationalizing seamful design in explainable ai. *arXiv preprint arXiv:2211.06753*, 2022.

[Gero *et al.*, 2020] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–12, 2020.

[Graus and Ferwerda, 2019] Mark Graus and Bruce Ferwerda. Theory-grounded user modeling for personalized hci. *Personalized human-computer interaction*, 2019.

[He *et al.*, 2023] Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*, 2023.

[Ho *et al.*, 2022] Mark K Ho, Rebecca Saxe, and Fiery Cushman. Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11):959–971, 2022.

[Jain *et al.*, 2018] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N Patel. Convey: Exploring the use of a context view for chatbots. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–6, 2018.

[Jakubik *et al.*, 2022] Johannes Jakubik, Jakob Schöffer, Vincent Hoge, Michael Vössing, and Niklas Kühl. An empirical evaluation of predicted outcomes as explanations

in human-ai decision-making. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 353–368. Springer, 2022.

[Jamali *et al.*, 2023] Mohsen Jamali, Ziv M Williams, and Jing Cai. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain. *arXiv preprint arXiv:2309.01660*, 2023.

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

[Kim *et al.*, 2023] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[Kocielnik *et al.*, 2019] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[Koedinger *et al.*, 2003] Kenneth R Koedinger, VAWMM Aleven, and Neil Heffernan. Toward a rapid development environment for cognitive tutors. In *12th Annual Conference on Behavior Representation in Modeling and Simulation*, 2003.

[Langley *et al.*, 2022] Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in Artificial Intelligence*, 5:62, 2022.

[Le *et al.*, 2019] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Liu *et al.*, 2023] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. A first look at llm-powered generative news recommendation. *arXiv preprint arXiv:2305.06566*, 2023.

[Mantravadi *et al.*, 2020] Soujanya Mantravadi, Andreas Dyrøy Jansson, and Charles Møller. User-friendly mes interfaces: Recommendations for an ai-based chatbot assistance in industry 4.0 shop floors. In *Asian Conference on Intelligent Information and Database Systems*, pages 189–201. Springer, 2020.

[Mohamad and Kouroupetroglou, 2014] Yehya Mohamad and Christos Kouroupetroglou. Research report on user modeling for accessibility. *W3C WAI Research and Development Working Group (RDWG) Notes*, 2014.

[Moreno *et al.*, 2021] Pol Moreno, Edward Hughes, Kevin R McKee, Bernardo Avila Pires, and Théophane Weber. Neural recursive belief states in multi-agent reinforcement learning. *arXiv preprint arXiv:2102.02274*, 2021.

[Nematzadeh *et al.*, 2018] Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. Evaluating theory of mind in question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, October-November 2018.

[Premack and Woodruff, 1978] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

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

[Riemer *et al.*, 2024] Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D Weisz, and Murray Campbell. Position: Theory of mind benchmarks are broken for large language models. *arXiv preprint arXiv:2412.19726*, 2024.

[Sap *et al.*, 2019] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Sap *et al.*, 2022] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large LMs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Shneiderman, 2020] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[Sun *et al.*, 2023a] Linzhuang Sun, Nan Xu, Jingxuan Wei, Bihui Yu, Liping Bu, and Yin Luo. Rational sensibility: Llm enhanced empathetic response generation guided by self-presentation theory. *arXiv preprint arXiv:2312.08702*, 2023.

[Sun *et al.*, 2023b] Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. Sentiment analysis through llm negotiations. *arXiv preprint arXiv:2311.01876*, 2023.

[Terentev, 2023] Maksim Terentev. Emergent theory of mind in large language models. 2023.

[Völkel *et al.*, 2019] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Quay Au, Bernd Bischl, Markus Bühner, and Heinrich Hussmann. Opportunities and challenges of utilizing personality traits for personalization in hci. *Personalized Human-Computer Interaction*, 31, 2019.

[Wang and Goel, 2022] Qiaosi Wang and Ashok K Goel. Mutual theory of mind for human-ai communication. *arXiv preprint arXiv:2210.03842*, 2022.

[Wang *et al.*, 2021] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[Westby and Riedl, 2023] Samuel Westby and Christoph Riedl. Collective intelligence in human-ai teams: A bayesian theory of mind approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6119–6127, 2023.

[Williams *et al.*, 2022] Jessica Williams, Stephen M Fiore, and Florian Jentsch. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*, 5:750763, 2022.

[Xu *et al.*, 2024] Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*, 2024.