Mutual Theory of Mind for Human-AI Communication

Qiaosi Wang
qswang@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Ashok K. Goel
ashok.goel@cc.gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

ABSTRACT

New developments are enabling AI systems to perceive, recognize, and respond with social cues based on inferences made from humans' explicit or implicit behavioral and verbal cues. These AI systems, equipped with an equivalent of human's Theory of Mind (ToM) capability, are currently serving as matchmakers on dating platforms, assisting student learning as teaching assistants, and enhancing productivity as work partners. They mark a new era in human-AI interaction (HAI) that diverges from traditional human-computer interaction (HCI), where computers are commonly seen as tools instead of social actors. Designing and understanding the human perceptions and experiences in this emerging HAI era becomes an urgent and critical issue for AI systems to fulfill human needs and mitigate risks across social contexts. In this paper, we posit the Mutual Theory of Mind (MToM) framework, inspired by our capability of ToM in human-human communications, to guide this new generation of HAI research by highlighting the iterative and mutual shaping nature of human-AI communication. We discuss the motivation of the MToM framework and its three key components that iteratively shape the human-AI communication in three stages. We then describe two empirical studies inspired by the MToM framework to demonstrate the power of MToM in guiding the design and understanding of human-AI communication. Finally, we discuss future research opportunities in human-AI interaction through the lens of MToM.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; **Empirical studies in HCI**; *Natural language interfaces*; • **Computing methodologies** → *Artificial intelligence*.

KEYWORDS

theory of mind, human-AI interaction, social intelligence

ACM Reference Format:
Qiaosi Wang and Ashok K. Goel. 2024. Mutual Theory of Mind for Human-AI Communication. In *Proceedings of Workshop on Theory of Mind in Human-AI Interaction at CHI 2024 (ToMinHAI at CHI 2024).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With new technology advancements, AI systems are increasingly serving different social roles across contexts. For example, AI systems are acting as matchmakers to provide matches for our business or life partners, as personal assistants to manage our daily routines, as learning assistants to facilitate student learning, and more. These AI systems are often able to perceive, recognize, and react to human characteristics, needs, and perceptions embedded in our behavioral and verbal cues. This presents a new interaction paradigm in Human-AI Interaction (HAI) that diverges from the traditional Human-Computer Interaction (HCI)— people are expecting AI systems to possess social intelligence for diverse social functions, yet are often uncertain about the AI's capabilities and social roles during interactions. Oftentimes during HAI, people build different perceptions or mental models of the AI based on the AI outputs [13, 42, 44, 46]; at the same time, AI systems are also constantly building different interpretations of human characteristics, needs, and goals based on human's input [9, 12, 34, 39, 41]. In this emerging HAI paradigm, interpretations of each other, from both the human side and the AI side, are playing an increasingly crucial part in shaping human-AI interactions, but how should the HAI community study it in a systematic way to enhance human-AI interactions?

In this paper, we propose viewing this emerging HAI paradigm through the lens of Mutual Theory of Mind (MToM). We draw inspirations from the human-human communication process, largely enabled by our basic cognitive and social ability of Theory of Mind (ToM). ToM is our ability to make conjectures about ourselves and others' mental states (e.g., emotions, intentions) [2, 16], and it is the key to enabling many human social behaviors such as communication repair and making shared plans or goals [3]. Having the capability of ToM enables us to construe a mental model of others' minds, which includes their thoughts, preferences, goals, needs, plans, etc. [2, 33]. In typical human-human communications, having a **Mutual Theory of Mind (MToM), meaning all parties involved in the interaction possess the ToM**, enables us to continuously refine our interpretations of each others' minds through behavioral and verbal feedback, helping us to maintain constructive and coherent communications.

Drawing on the parallel between the MToM in human-human communications and the emerging HAI paradigm where both humans and AIs can construct representations of each other during communications, we propose the framework of Mutual Theory of Mind to guide the next generation of research in human-AI communications. We argue that *MToM as a framework provides a process and content account of human-AI communication* that emphasizes the iterative mutual shaping of each party's interpretations and feedback through different stages of the communication process. We will first review relevant literature on ToM and human-AI communication, then describe the MToM framework in details. We then summarize two of our empirical studies inspired by the MToM framework, focusing specifically on the second-level ToM— the idea of "I can think about what you think about my mind"— in human-AI communications. Finally, we discuss potential research opportunities in human-AI interaction through the lens of MToM.

ToMinHAI at CHI 2024, May 12th, Honolulu, Hawaii
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 RELATED WORK

2.1 Theoretical Perspectives of Communication

Communication is commonly defined as *"the process of transmitting information and common understanding from one person to another."* [32] Scholars across disciplines have offered different perspectives to study and enhance communication.

In communication studies, researchers have focused on the different components at play during the communication process. The classic Shannon-Weaver model of communication [38] outlines several key components during the communication process [32]: *sender* who initiates the communication process by sending messages *encoded* using symbols, gestures, words, or sentences through a chosen *channel* to the *receiver*. While the message is transmitting through the channel, there could be *noises* that could distort the message. After receiving the message from the sender, the receiver will *decode* the message into meaningful information, depending on how the receiver interprets the message. Finally, the receiver will provide *feedback* as a response to the sender. These key components determine the quality and effectiveness of the communication.

The Cognitive Science perspective of communication highlights the critical role of ToM [33]. ToM enables us to make suppositions of other's minds through verbal and behavioral cues, acting as the foundation of human-human communication [2, 3]. From this perspective, both interlocutors during communication can form interpretations of what's on the other interlocutor's mind based on the implicit and explicit communication cues. For example, we can often infer the interlocutors' goals, plans, or preferences based on what they said, their facial expressions, or their bodily expressions [2, 33]. Based on that interpretation we formed about the other's mind, we will act accordingly to correct, explain, or persuade. This cycle of building an interpretation of other's minds and then act upon that interpretation continues iteratively throughout the communication process. Inferring about each other's minds through behavioral cues, according to this perspective, is therefore crucial to a smooth and successful communication.

Communication process can also be interpreted from the social science perspective through impression management [14]. In his seminal work, Goffman describes social interaction as an information game between individuals and their audience to maintain the "veneer of consensus" to keep the conversation going and to avoid awkwardness. During social interactions, the audience usually try to gather as much information as they could about the individuals they interact with in order to elicit a desirable response from the individual; whereas individuals put up performances through two kinds of expressions—– expressions that are intentionally performed to leave a certain impression (expression given) or expressions that are unintentionally given off that could influence the audience's impressions of them (expression given off)—– to manage impressions [14]. Throughout interactions, each party conveys their definition of the situation through communications: individuals by expressions and audience by reactions to the individuals.

These three perspectives on communication emphasize different aspects of the communication process: the communication study perspective focuses on the encoding and decoding process of messages; the cognitive science perspective discusses how behavioral cues can inform our interpretations of interlocutor's minds; the social science perspective describes how interpretations of others' minds could predict our behaviors. Our Mutual Theory of Mind framework attempts to bring these different emphasis together into one coherent framework to understand the mutual shaping process of interpretations and feedback during communication.

2.2 Theory of Mind in Human-AI Communication

Over the years, many researchers have recognized the crucial role of ToM in HAI. In human-robot teaming research, ToM has been intentionally built in as part of the system architecture to help robots monitor world state as well as the human state [8], to construct simulation of hypothetical cognitive models of the human partner to account for human behaviors that deviate from original plans [34], and to help robots to build mental models about user beliefs, plans and goals [20, 24]. Robots built with ToM have demonstrated positive outcomes in team operations [8] and are perceived to be more natural and intelligent [29].

Other research in HCI and human-centered AI has also been exploring along the realm of ToM, focusing mostly on enhancing user's mental models and understanding of the AI systems. Prior research has explored people's mental model of AI systems— people's mental model of AI agents could include global behavior, knowledge distribution, and local behavior [13]. People's perception of AI systems is instrumental in guiding how they interact with AI systems [13] and thus serves as a precursor to their expectation of AI's behavior. Some recent research has also begun to examine how to automatically infer user's mental model of AI. Prior research suggests the potential of leveraging linguistic cues to indicate people's perception of AIs during human-AI interactions. Researchers have been able to infer users' emotions towards an AI agent [40] and signs of conversation breakdowns [26]from communication cues.

Given that AI's behavior and output could also influence user's mental model of the AI, and therefore how the user decides to interact with the AI, we want to highlight that the interpretation-feedback loop is mutual during the human-AI communication process— user's mental model of the AI can be informed by the AI's output, yet AI's interpretation of the user can also be informed by the user's output, which is determined by the user's mental model of the AI. We propose the Mutual Theory of Mind framework to capture this mutual shaping process of interpretation-feedback during human-AI communication.

3 MUTUAL THEORY OF MIND FRAMEWORK

Drawing from theoretical and empirical work, we posit the MToM framework to guide the understanding and design of communications between humans and AI systems that exhibit social behaviors enabled by ToM-like capability. The MToM framework provides both process and content account of human-AI communication by highlighting *three elements* that mutually shape the human-AI communication process in *three stages*.

3.1 Three Elements of the MToM Framework

In the MToM framework, three elements are critical for humans and AI to reach mutual understanding during the communication process: *interpretation, feedback, and mutuality*.

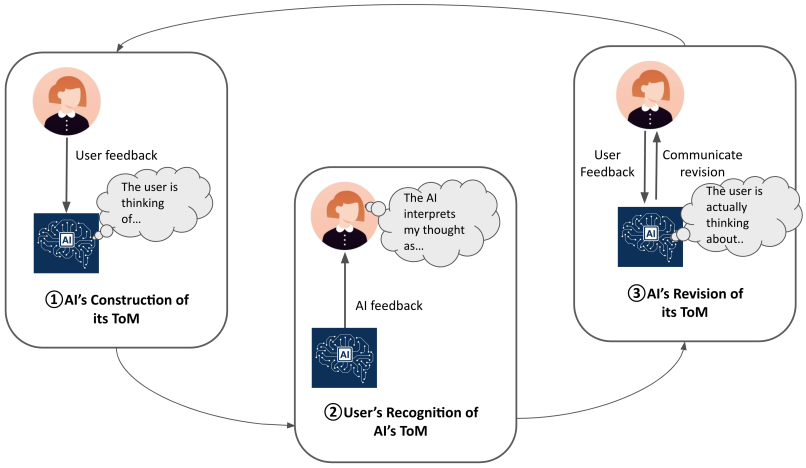

Figure 1: The Mutual Theory of Mind framework that illustrates the iterative process of human-AI communication. The figure also illustrates the three elements across the three stages in MToM.

In human-AI communication, humans and AI can each construct and revise their *interpretations* of each other based on feedback from the other party. These interpretations are their interpretations of what's on the other party's mind. This can include the other party's understanding or representations of the world, of the task, of the interlocutor, and of other things. For example, humans can build an interpretation of the AI's representations of the world, of the human, of the plans and goals of the current task during human-AI collaboration, and vice versa. It is important to note that these interpretations can be recursive given that ToM can be of higher levels, meaning that interpretations can not only refer to "my interpretation of your mind", but can also refer to "my interpretation of your interpretation of my mind." We further illustrate this recursive property of MToM through our empirical studies in Section 4, both focused on second-level ToM.

Feedback, often in the form of verbal or behavioral cues, is generated with different complexities based on the interpretations of each other. For instance, in a human-chatbot conversation, humans would generate simpler command when they believed the chatbot could not understand complex human language; the chatbot would generate simpler feedback when they interpreted the human needs were simple (e.g., asking about the weather).

While each party involved in the communication is capable of constructing interpretations and generating feedback on their own, communication is a two-way interaction, which means all parties involved in the communication process are *mutually shaping* each other's interpretations of each other's minds through feedback. Human's interpretations of the AI is constantly shaped by the AI's output, which is shaped by the AI's interpretations of the human's mind, and vice versa.

These three key elements play a critical part in determining the success and failure of a communication— inaccurate feedback could inform inaccurate interpretations, inaccurate interpretations could generate inaccurate feedback. Failure in any of them can undermine the mutual shaping process in human-AI communication.

3.2 Three Stages of MToM Framework

In the MToM framework, these three elements are constantly shaping the communication between humans and AI during three stages: *AI's construction of its ToM, user's recognition of AI's ToM, and AI's revision of its ToM.*

In the first stage, *AI's construction of its ToM*, the AI system takes in user feedback and tries to interpret what's on the user's mind. Depending on the specific communication context, this could be the user's goals, needs, preferences etc. Based on the interpretation, the AI *constructs* its theory of the user's mind, which helps the AI to generate responses accordingly to help the user fulfill their goals and needs in this instance of communication.

After the AI generates its response to the user, the user then *recognizes* the AI's interpretation of the user's mind based on AI's response. This recognition leads the user to construct their interpretation of the AI, which includes the AI's capability, working mechanism, and how they are interpreted by the AI.

The AI's interpretation of the user's mind might not always be accurate. Based on the user's interpretation of the AI, the user provides feedback to the AI, which the AI takes in to *revise* or update its interpretation of the user's mind based on the user's feedback. To make sure that the user's interpretations of the AI is accurate after such revisions, it is crucial for the AI to also communicate this revision of its ToM back to the user through feedback.

Throughout these three stages, the three elements of MToM (interpretation, feedback, and mutuality) interact with each other to shape the communications between the human and the AI. Outlining these elements and stages not only provide a content and process account of communications between humans and AI systems equipped with ToM-like capability, but also surface research opportunities to enhance such human-AI communication process to explore, understand, and examine these elements in specific stages. In the next section, we summarize two empirical studies inspired by the MToM framework to examine these elements during the construction and the recognition stages.

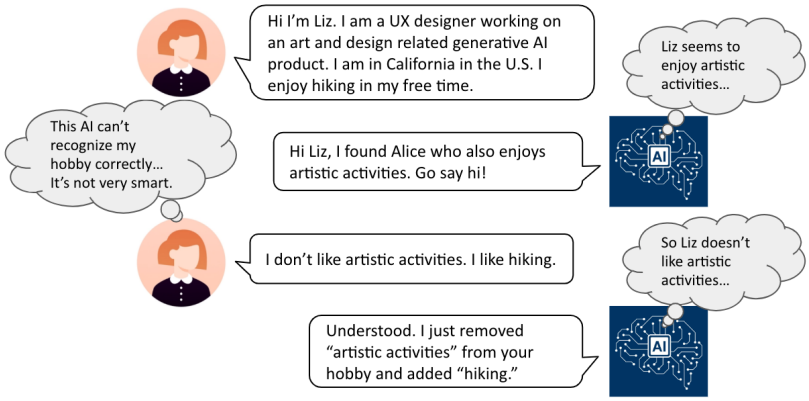

Figure 2: An example human-AI communication dialogue between a student Liz and an AI agent that can provide social recommendations based on Liz's self-introduction. This dialogue shows the recursive nature of the MToM in human-AI communication.

4 EMPIRICAL EXPLORATION OF MTOM IN HUMAN-AI COMMUNICATION

Guided by the MToM framework, our work aims to enhance and understand the human-AI communication process by examining the interplay between interpretation, feedback and mutuality during the communication between humans and AI systems that exhibit ToM-like capability. Our work so far has focused on the recursive nature of ToM, specifically, second-level ToM, in human-AI communication, which is the idea of "I can think about what you are thinking about me." So far we have examined second-level ToM in the first two stages of MToM: the construction stage and the recognition stage. At the construction stage, we examined the feasibility of how the AI can construct its interpretations of the user's interpretations of the AI; at the recognition stage, we explored the human's perceptions and reactions of the AI after recognizing the AI's interpretations of them, specifically when the AI's interpretations are wrong. We summarize our empirical work at these two stages in this section.

Most of our work took place in online learning context, where AI agents are increasingly deployed as teaching assistants or social facilitators to provide informational and social support to online learners. Figure. 2 shows a human-AI communication dialogue between an AI agent acting as a social facilitator to provide social recommendations to online learners based on the inferences made about the student's social preferences from their self-introduction.

4.1 Constructing User's Perceptions of AI through Linguistic Cues

Recent technical advancements have made it possible for AI systems to "read users' minds" by predicting our shopping preferences [30], emotional states [7, 37, 43], and personalities [15] with fairly high accuracy. However, user's perceptions of such advanced AI systems have been under explored. Understanding user perceptions of such AI systems is critical to the success of human-AI communication, especially given people's often unrealistically high expectation of AI system's capability. This is especially common during the communications between humans and Conversational Agents (CAs), which can present human-level natural language understanding and generation when powered by LLMs, yet often present inconsistent performance across various task-specific capabilities. This "gulf" between user expectation and experience with CAs [31] has led to constant user frustration, frequent conversation breakdowns, and eventual abandonment of CAs [31, 47].

To understand how we could potentially mitigate or bridge this "gulf" between user expectation and experience with the CAs, we looked to the MToM framework for inspiration. We took inspirations from the recursive interpretations and feedback in the MToM framework, and envisioned an adaptive CA that could automatically construct a representation of the user's perceptions of the CA through verbal or behavioral cues embedded in user feedback. An automatic construction of the user's perceptions of the CA would enable the CA to monitor users' changing perceptions and adapt their behaviors accordingly to cater to users' needs, or even provide nudges to help users build a better mental model of the CA.

To examine the feasibility of automatic construction of the user's perceptions of AI, we deployed an AI agent acting as a virtual teaching assistant to answer students' logistic questions about the class in an online class discussion forum for 10 weeks with about 376 students enrolled. We collected students' bi-weekly perceptions of the agent in terms of perceived anthropomorphism, intelligence, and likeability, as well as students' questions asked to the agent throughout. We then extracted various linguistic cues from students' questions asked to the agent, such as readability (the level of ease readers can comprehend a given text), sentiment (emotions conveyed through the language), linguistic diversity (diversity of the conversation topics or the richness of language used), and adaptability (how adaptable are students' questions to the agent's responses). We then built linear regression models using the linguistic characteristics as independent variables to predict each of the student community's perceptions of the AI agent (anthropomorphism, intelligence, likeability).

We found that verbosity negatively associates with student perceptions of the AI agent, whereas readability, sentiment, diversity, and adaptability positively associate with anthropomorphism, intelligence, and likeability. Our findings suggest that it is feasible to extract linguistic features to measure users' perceptions of CA during conversations, and thus enable the CA to constantly interpret and provide desirable responses that cater to user perceptions. More details about this study and the model results can be found in our CHI 2021 paper [44].

4.2 User Reactions and Perceptions of AI Misrepresentations

Many hyper-personalized AI systems that can profile users' characteristics and traits have been deployed in people's daily lives, with the ultimate goal of providing personalized recommendations in shopping, music, social media, etc. As these systems become more advanced in profiling people's most personal and complex traits such as personalities and emotions [17, 19, 28], they sometimes give the illusion of "machines can read our minds" [18]. This illusion has led to various—rather concerning—reactions and perceptions of AI with people attributing AI with beyond-human expertise at reading people's emotions and personalities [21, 45]. However, people's perceptions and reactions of AI when this illusion is broken in the face of *AI misrepresentations* have not yet been explored.

AI misrepresentations refer to when AI systems make the wrong inferences about people's implicit characteristics and traits, such as personality, based on user data. Even algorithms with supposedly high accuracy can make mistakes when powering hyper-personalized AI systems in-the-wild [35]. Guided by the MToM framework, we situate this problem in the recognition stage of the MToM framework to examine people's reactions and perceptions of AI after recognizing that AI has an inaccurate interpretations of the human. Understanding people's reactions and perceptions of the AI after encountering AI misrepresentations could offer valuable insights into whether and how people changed their intuitions, beliefs, and reactions of AI in the face of AI fallibilities. This could provide critical implications for the future design and development of responsible interventions, mitigation, and repair strategies to retain user trust, minimize harms, and prevent overreliance when such AI systems inevitably err.

To understand people's reactions and perceptions of AI misrepresentations on their personalities, we conducted semi-structured interviews with 20 college students and a large survey experiment with 198 students on the Prolific platform. In both studies, we took a Wizard-of-Oz approach to fabricate intentionally inaccurate/accurate personality inferences based on participants' personality ground truth. We showed participants in both studies their "AI-generated personality inferences" to elicit their perceptions and reactions of AI misrepresentations.

In both the interviews and survey experiment, we first familiarized participants with some sample student-AI dialogues where the AI agent provided a paragraph describing the student's personality based on a paragraph of the students' self-introduction. We measured participants' baseline perceptions of the agent after viewing the sample dialogues. In both studies, participants were randomly assigned to either receive accurate or inaccurate "AI-generated personality inferences" and their perceptions of the agent and reactions were measured after seeing their own inferences. We analyzed our interview data using reflexive thematic analysis, then built linear regression models and conducted moderation analysis to understand the changes in students' perceptions of AI before and after viewing AI misrepresentations.

Our results showed that people's existing and newly acquired knowledge of AI are highly connected to people's reactions and perceptions of AI after encountering AI misrepresentations. Specifically, we found that participants acquired new knowledge from AI (mis)representations. Such newly acquired knowledge prompted participants to adopt different rationales to interpret how AI worked: AI works like a machine, human, and/or magic. These rationales could co-exist at any given time, yet are bounded by participants' existing AI knowledge, tech proficiency, and how much they could make sense of AI's specific inferences. Through our linear regression models, we also established that people's existing AI knowledge, i.e., AI literacy, can significantly moderate changes in people's trust of the AI after encountering AI misrepresentations, highlighting the importance of taking into account of people's knowledge and characteristics when building trustworthy AI systems [4, 5, 11, 36].

Based on our interviews, we found that people's AI knowledge, especially the rationales participants adopted after acquiring new knowledge from AI misrepresentations, are highly connected to participants' reactions to AI misrepresentations. After being shown the AI-generated inaccurate personality inferences about them, participants displayed a range of reactions: some participants believed there was some truth to AI misrepresentation; some participants rationalized it and blamed themselves instead; some participants were forgiving of the AI's mistakes. Building on top of prior work that has suggested people's tendency of over-trusting and viewing AI as an authority [21, 22, 45], we highlighted that these reactions and perceptions still persisted, and even exacerbated, when people encountered AI misrepresentations.

This work provides important implications of how AI systems can and should be designed to be aware of people's ever-evolving AI knowledge, and provide customized repair strategies accordingly to mitigate potential user harms from AI misrepresentations. We provided a specific set of rationales and encourage future work to explore techniques that could allow automatic identifications of users' rationales adopted in real-time. One mitigation strategy could be to provide explanations tailored to the specific rationale that people adopted at the time. For instance, if a user adopted the magic rationale, the AI could provide explanations to nudge the user to adopt the machine rationale to reduce overreliance.

5 DISCUSSION

In the previous section, we presented two empirical studies inspired by the MToM framework to understand the construction and recognition stage of MToM in human-AI interaction. The first study shows that by leveraging the linguistic cues embedded in user feedback, it is feasible to equip the AI with a ToM to constantly model and interprets users' perceptions of the AI; The second study shows that by exploring users' reactions and perceptions of AI

after recognizing AI's inaccurate ToM through AI feedback, we could provide more personalized and adaptive AI repair strategies to mitigate potential harms. Both studies contributed unique implications for enhancing human-AI communication by focusing on the interplay between the three elements (interpretations, feedback, mutuality) in the first two stages of the MToM framework. Here, we discuss other research opportunities that can be further explored to enhance human-AI interaction by examining the interplay between the three elements at each stage in the MToM framework.

Research Opportunities in the Construction Stage. At the construction stage, the AI system constructs its interpretations of the user's mind based on user feedback. In this stage of human-AI interaction, the following research questions can be asked by examining the three elements of user feedback, user feedback shaping AI interpretation, and AI interpretation, as seen in Figure 1: (1) What kind of user feedback would help the AI construct accurate interpretations of the user's mind? (2) How can we construct AI's ToM through cues embedded in user feedback? (3) What should be constructed as part of the AI's interpretation of the user? Many existing research surrounding ToM in human-AI interaction have explored or begun to explore some of these questions. For example, Baker et al. and many others have been exploring different techniques such as Bayesian ToM for the AI to detect and analyze observable or unobservable user feedback to model the human minds. Other fields such as emotion detection, ubiquitous computing, physical sensing can also offer valuable implications to the construction of AI's ToM. An underlying issue that should be examined as the foundation of these research questions is the problem of operationalization— how can we operationalize ToM given the huge variations of human minds in different contexts across individuals?

Research Opportunities in the Recognition Stage. At the recognition stage, the user recognizes the AI's interpretations of the user's mind based on AI feedback. If we look at the three elements of AI feedback, AI feedback shaping user perceptions, and user perceptions at this stage (see Figure 1), several research questions can be examined: (1) What kind of AI feedback can convey its interpretations of the user's mind to the user? (2) How can design features of the AI feedback shape or trigger user's perceptions if the AI? (3) What dimensions of user perceptions could change after recognizing AI's interpretations of their mind? A fundamental research question and opportunity here is to explore and map out the design characteristics/features of the AI feedback that could lead to changes in certain dimensions of the user perceptions of the AI. For example, certain wording or phrasing of the AI feedback could lead to increased levels of anthropomorphism [25]. Understanding and mapping out design features of AI feedback and changes in user perceptions could offer valuable implications to mitigate and even prevent harmful perceptions of AI.

Research Opportunities in the Revision Stage. At the revision stage, the AI system conducts revision of its interpretations of the user's mind based on user feedback, and then communicates the revision back to the user to re-establish mutual understanding in human-AI communication. The following research questions can be explored at this stage by looking at the three elements of user feedback, user feedback shaping AI's interpretations of the user's mind, and AI feedback shaping user's perceptions of the AI (see Figure 1): (1) What kind of user feedback can help the AI revise its interpretations of the user's mind? (2) How can the AI take into account of user feedback to revise its interpretations of the user's mind? (3) What kind of AI feedback can help the user understand its revised interpretations of the user's mind? Existing work on human-centered explainable AI [10, 11, 27] have been trying to tackle question (3) by taking into account of user traits and characteristics when designing AI feedback that can explain AI's working mechanism to the user. Research efforts that aim at addressing question (1) and (2) are also underway, with some scholars seeking to mimic the human capability of metacognition and introspection [6, 23] when implementing AI systems.

6 CONCLUSION

This paper proposed the Mutual Theory of Mind (MToM) framework to understand and design communications between humans and AI systems that are performing increasingly diverse social functions in human society. The MToM framework highlighted three key elements (interpretations, feedback, mutuality) that continuously interact with each other throughout the three stages of the communication process (construction, recognition, and revision). The MToM framework thus provides a process and content account of human-AI communication as a iterative, mutual-shaping process of the human's and AI's interpretations of each other based on communication feedback. We summarized two empirical studies that were inspired by the MToM framework: the first study demonstrated an innovative use of user feedback to enhance AI's interpretation of user's perception of the AI during communication; the second study examined user perceptions and reactions after recognizing AI's incorrect interpretations of the user, highlighting the role of people's newly acquired and evolving knowledge in shaping human-AI communication. We concluded by discussing research opportunities in human-AI interaction through the lens of MToM at the construction, recognition, and revision stage by examining the interplay between the three key elements in the MToM framework.

ACKNOWLEDGMENTS

This research has been supported by US NSF Grant #2247790 to the National AI Institute for Adult Learning and Online Education (AI-ALOE; aialoe.org).

REFERENCES

[1] 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.
[2] Simon Baron-cohen. 1999. Evolution of a Theory of Mind? In *The Descent of Mind: Psychological Perspectives on Hominid Evolution*. Oxford University Press, 1–31.
[3] Simon Baron-Cohen, Alan M Leslie, Uta Frith, et al. 1985. Does the autistic child have a "theory of mind". *Cognition* 21, 1 (1985), 37–46.
[4] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2023. Machine Explanations and Human Understanding. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1–1.
[5] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *arXiv preprint arXiv:2301.07255* (2023).
[6] Michael T Cox. 2005. Metacognition in computation: A selected research review. *Artificial intelligence* 169, 2 (2005), 104–141.
[7] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

[8] Sandra Devin and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. *ACM/IEEE International Conference on Human-Robot Interaction* 2016-April (2016), 319–326. https://doi.org/10.1109/HRI.2016.7451768

[9] Rahul R Divekar, Jeffrey O Kephart, Xiangyang Mou, Lisha Chen, and Hui Su. 2019. You talkin'to me? A practical attention-aware embodied agent. In *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part III 17*. Springer, 760–780.

[10] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[11] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*. Springer, 449–466.

[12] Bobbie Eicher, Kathryn Cunningham, Sydni Peterson Marissa Gonzales, and Ashok Goel. 2017. Toward mutual theory of mind as a foundation for co-creation. In *International Conference on Computational Creativity, Co-Creation Workshop*.

[13] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376316

[14] Erving Goffman. 1978. *The Presentation of Self in Everyday Life*. London: Harmondsworth.

[15] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 149–156.

[16] Alison Gopnik and Henry M Wellman. 1992. Why the child's theory of mind really is a theory. (1992).

[17] Liang Gou, Michelle X Zhou, and Huahai Yang. 2014. KnowMe and ShareMe: understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 955–964.

[18] Andrea L Guzman. 2020. Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication* 1 (2020), 37–54.

[19] Margeret Hall and Simon Caton. 2017. Am I who I say I am? Unobtrusive self-representation and personality recognition on Facebook. *PloS one* 12, 9 (2017), e0184417.

[20] Maaike Harbers, Karel Van Den Bosch, and John Jules Meyer. 2009. Modeling agents with a theory of mind. *Proceedings - 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2009* 2 (2009), 217–224. https://doi.org/10.1109/WI-IAT.2009.153

[21] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On being told how we feel: how algorithmic sensor feedback influences emotion perception. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–31.

[22] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. " Because AI is 100% right and safe": User attitudes and sources of AI authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[23] Mitsuo Kawato and Aurelio Cortese. 2021. From internal models toward metacognitive AI. *Biological cybernetics* 115, 5 (2021), 415–430.

[24] Kyung-Joong Kim and Hod Lipson. 2009. Towards a simple robotic theory of mind. (2009), 131. https://doi.org/10.1145/1865909.1865937

[25] Mengjun Li and Ayoung Suh. 2021. Machinelike or humanlike? A literature review of anthropomorphism in AI-enabled technology. In *54th Hawaii International Conference on System Sciences (HICSS 2021)*. 4053–4062.

[26] Q. Vera Liao, Werner Geyer, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, and N. Sadat Shami. 2018. All Work and No Play? Conversations with a Question-and-Answer Chatbot in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Vol. 8. ACM Press, New York, New York, USA, 1–13. https://doi.org/10.1145/3173574.3173577

[27] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[28] Tony Liao and Olivia Tyson. 2021. "Crystal Is Creepy, but Cool": Mapping Folk Theories and Responses to Automated Personality Recognition Algorithms. *Social Media+ Society* 7, 2 (2021), 20563051211010170.

[29] Shuhong Lin, Boaz Keysar, and Nicholas Epley. 2010. Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology* 46, 3 (2010), 551–556. https://doi.org/10.1016/j.jesp.2009.12.019

[30] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, 1 (2003), 76–80.

[31] Ewa Luger and Abigail Sellen. 2016. "Like having a really bad PA": the gulf between user expectation and experience of conversational agents. *CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 5286–5297. https://doi.org/10.1145/2858036.2858288

[32] Fred C Lunenburg. 2010. Communication: The process, barriers, and improving effectiveness. *Schooling* 1, 1 (2010), 1–10.

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

[34] David V. Pynadath and Stacy C. Marsella. 2005. PsychSim: Modeling theory of mind with decision-theoretic agents. *IJCAI International Joint Conference on Artificial Intelligence* (2005), 1181–1186.

[35] Ashwini Rao, Florian Schaub, and Norman Sadeh. 2015. What do they know about me? Contents and concerns of online behavioral profiles. *arXiv preprint arXiv:1506.01675* (2015).

[36] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1616–1628.

[37] H Andrew Schwartz, Maarten Sap, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, et al. 2016. Predicting individual well-being through the language of social media. In *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, 516–527.

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

[39] Mei Si, Stacy C Marsella, and David V Pynadath. 2010. Modeling appraisal in theory of mind reasoning. *Autonomous Agents and Multi-Agent Systems* 20 (2010), 14–31.

[40] Marcin Skowron, Stefan Rank, Mathias Theunis, and Julian Sienkiewicz. 2011. The good, the bad and the neutral: affective profile in dialog system-user communication. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 337–346.

[41] Sarah E Walsh and Karen M Feigh. 2022. Understanding human decision processes: Inferring decision strategies from behavioral data. *Journal of cognitive engineering and decision making* 16, 4 (2022), 301–325.

[42] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–24.

[43] Qiaosi Wang, Shan Jing, David Joyner, Lauren Wilcox, Hong Li, Thomas Plötz, and Betsy Disalvo. 2020. Sensing Affect to Empower Students: Learner Perspectives on Affect-Sensitive Technology in Large Educational Contexts. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*. 63–76.

[44] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. 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*. 1–14.

[45] Jeffrey Warshaw, Tara Matthews, Steve Whittaker, Chris Kau, Mateo Bengualid, and Barton A Smith. 2015. Can an Algorithm Know the" Real You"? Understanding People's Reactions to Hyper-personal Analytics Systems. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 797–806.

[46] Justin D Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection not required? Human-AI partnerships in code translation. In *26th International Conference on Intelligent User Interfaces*. 402–412.

[47] Jennifer Zamora. 2017. I'm sorry, dave, i'm afraid i can't do that: Chatbot perception and expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 253–260.