ToMinHAI at CUI'2025: Theory of Mind in Human-CUI Interaction

Qiaosi Wang
Human Computer Interaction
Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
qiaosiw@andrew.cmu.edu

Joel Wester
Department of Computer Science
Aalborg University
Aalborg, Denmark
joelw@cs.aau.dk

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

Minha Lee
Industrial Design
Eindhoven University of Technology
Eindhoven, Netherlands
m.lee@tue.nl

Justin D. Weisz
IBM Research
Yorktown Heights, New York, USA
jweisz@us.ibm.com

Mei Si
Cognitive Science
Rensselaer Polytechnic Institute
Troy, New York, USA
SIM@rpi.edu

Abstract

New AI developments are enabling CUIs to take on diverse social roles to facilitate interactions with humans. To support such increasingly complex and social interactions, researchers draw from Theory of Mind (ToM)—our ability to attribute mental states like intentions, goals, and emotions to ourselves and others for seamless communication. Given ToM's importance in human interaction, AI and HCI researchers explore both building ToM-like capabilities in CUIs and understanding how humans attribute mental states to CUIs. These perspectives form the emerging paradigm of Mutual Theory of Mind (MToM) in human-CUI interaction, where both parties iteratively interpret each other's internal states. Building on the success of the 1st ToMinHAI workshop at CHI 2024, this installment invites researchers from AI, ML, HCI, and related fields to discuss ToM in human-CUI interactions to inform the future design of conversational AI.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Collaborative and social computing**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

theory of mind, mutual theory of mind, mental model, human-AI interaction, human-centered AI, social intelligence, conversational user interface, human-AI conversations

ACM Reference Format:
Qiaosi Wang, Joel Wester, Marvin Pafla, Minha Lee, Justin D. Weisz, and Mei Si. 2025. ToMinHAI at CUI'2025: Theory of Mind in Human-CUI Interaction. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 08–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3719160.3728627

1 Motivation

Recent advancements in AI are enabling CUIs to communicate with humans at various social capacities, transforming the way humans interact with AI agents through natural conversations. These developments are particularly evident in communication-based CUI applications such as customer service chatbots, semi-autonomous driving (e.g., Tesla), and AI companions. To enhance these conversational interactions, many researchers are turning to a key cognitive-social capability that enables human communications—Theory of Mind. Theory of Mind (ToM) [2, 15, 25] refers to humans' capability of attributing mental states such as intentions, goals, emotions, and beliefs to ourselves and others. This concept has become of great interest in human-AI interaction research to enhance human-AI communications [e.g. 1, 7, 12, 36, 38]. In human-human communication, a functioning ToM enables us to make conjectures about each others' minds through behavioral and verbal cues, which allows us to make predictions about each others' behaviors and perceptions of the world [25, 36] so that we could behave accordingly.

Given the fundamental role of ToM in human communications, many AI researchers believe that *equipping and assessing AI's ToM-like capability* is the key to building conversational AI agents with heightened levels of social intelligence for them to work, play, and live with humans [4, 7, 36]. Moving towards this vision, some scholars have built AI's ToM-like capability to recognize and model people's non-verbal cues [19], emotional expressions [19], as well as people's beliefs, plans [30], and intents [16] through machine learning (e.g., Bayesian network) [16, 19], computer vision [8], and cognitive modeling [16, 17, 23, 31] across varying contexts. Among these human-AI interaction contexts, human-CUI interaction remains one of the most prevalent and promising contexts to extract conversational cues from human utterances to construct AI's interpretations of the human's mental states—conversational cues such as human's pitch in voices [6] and use of words in varying complexities and lengths [36] have all been suggested to be indicative of people's perceptions of AI agents and others. Alternatively, other researchers are examining ToM in human-CUI interaction by focusing on *humans' ToM*, such as humans' perceptions [36], mental models [3, 10, 14], and folk theories [9, 13] of CUIs performing at varying social capacities. Scholars have explored people's

CUI '25, Waterloo, ON, Canada

ACM ISBN 979-8-4007-1527-3/25/07
https://doi.org/10.1145/3719160.3728627

tendencies to attribute human mental states such as blame [32], emotions [28], perspectives [39], intentions [24], and social motivations [26] to CUIs.

In conversational AI research, much prior work explored people's ToM-enabled behaviors towards conversational systems, such as voice agents, chatbots, or robots. For example, prior work have explored how people perceive and ascribe mind to virtual agents [20] or attribute artificial personality to CUIs [11]. Specifically, Li et al. showed that children who ascribed greater mind perception to voice AI agents were more proactive in repairing communication breakdowns compared to when communication breakdowns occurred in children's interaction with humans [22]. In human-chatbot interactions, Chaves et al. highlighted how language design plays a key role in people's perceptions of chatbots— chatbots' conversational capabilities coupled with their assumed social roles (e.g., a hotel concierge) could influence people's expectations towards these chatbots [5]. However, the effectiveness of CUI design on people's mind attribution behaviors could differ based on the nature of the interaction environment. Wallkötter et al. found that the impact of robot's framing and behavior on people's mind perception is present in virtual but not in real-world settings [33]. As AI-powered CUIs exhibit more human-like communication behaviors, researchers also raise concerns about people's mental state attribution behaviors towards CUIs. For example, Shiramizu et al. evidenced that voice pitch affect people's perception of synthetic voices, highlighting the possibility of controlling stereotypic perceptions via voice pitch [29]. Studies have also shown chatbots' human-like ability in eliciting deep self-disclosure from participants, revealing personal and sensitive information during conversations [21]. Recent advancements in Large Language Models (LLMs) have further enhanced chatbots' humanlikeness in generating human-like responses, making it possible for conversational AI agents to generate deceptive and misleading information during conversations [40, 41]

While much prior work covers people's ToM of CUIs and vice versa, we know little of how these can be intertwined to support the design of human-centered CUIs. Putting together these two perspectives (i.e., the technical and social) of research on ToM in human-CUI interaction, from the CUI's side and from the human's side, there is an emerging paradigm that we call "Mutual Theory of Mind (MToM)" interaction [34, 35], where both the human and the CUI possess the capability of ToM and continuously make inferences and attribute mental states to each other during an interaction. Although enabling MToM in human-CUI interaction promises to make a great impact on achieving human-level social interactions that are adaptive, continuous, constructive, and natural, the specific ways to operationalize MToM, as well as its consequences on the interaction between human and CUIs, have yet to be envisioned in the context of human-CUI interactions.

2 Workshop Goals

As part of the ToMinHAI workshop series, this ToMinHAI workshop at CUI aims to examine **the current practices, challenges, and opportunities in designing, building, and evaluating ToM in human-CUI interactions**. This workshop will provide a platform for researchers from various disciplines studying ToM in human-CUI interaction from the CUI's ToM point of view and the human's ToM point of view to discuss techniques, methods, theories, and knowledge to build and measure CUI's ToM-like capability during conversational interactions with humans, as well as implications for designing socially intelligent CUI based on human's mental state attribution to CUIs during human-AI interaction. Additionally, this workshop will continue to explore the evaluation, design, and ethical issues surrounding the phenomenon of MToM for human-CUI interactions across social contexts (e.g., human-AI teams, human-AI collaboration, AI assistants). To support interdisciplinary discussions, we invite academic and industry researchers and practitioners in disciplines including but not limited to cognitive science, AI, HCI, design, machine learning, robotics, psychology, communication studies, and more to submit work that will inform our understanding of ToM in human-CUI interaction.

For the purpose of this workshop, we are especially interested in conversational AI systems that perform text-based or voice-based conversations with human users across different application contexts. Given the recent discourse on ToM in LLMs, we especially welcome submissions that discuss ToM and CUIs powered by generative AI that can present human-level conversational capabilities during human-CUI interactions. Although the definition of ToM has been well-established in psychology and cognitive science, we encourage authors to submit work that can expand or propose new definitions of ToM in human-CUI interaction research and establish the role of such expanded or new definitions. Although we focus on human-AI interaction in this proposal, we invite researchers studying ToM in human-human conversations or other communication contexts to help shape the discourse around the implications of MToM in human-CUI interaction contexts.

We propose three broad topics that cover important perspectives on ToM in human-CUI interactions. Within each topic we outline a number of inspirational research questions for which we aim to solicit contributions to our workshop.

(1) **Building and measuring a CUI's ToM-like capability in human-AI conversations**
   (a) What are the AI application contexts that would benefit from CUIs having ToM-like capability? (e.g., autonomous driving, personal assistants)
   (b) What techniques, methods, models, and data can be used to build a CUI's ToM-like capability? (e.g., machine learning techniques, cognitive models, foundational models)
   (c) How can CUIs adapt or personalize its conversational responses to a user based on user's mental states?
   (d) What factors from a CUI's design (e.g., voice pitch and/or gender) could influence user's perceptions of a CUI's social roles?
   (e) How to evaluate CUI's ToM-like capabilities during human-CUI interactions?
   (f) What does it mean to design a CUI's ToM-like capability in an ethical and human-centered manner?

(2) **Understanding and shaping humans' ToM in human-CUI interactions**
   (a) What kind of mental states (e.g., beliefs, blame) do people attribute to CUI with varying social capacities?
   (b) How does people's mental states attribution to CUI relate to the CUI's conversational capabilities and features?

   (c) How does the design of the CUI (e.g., features, functionalities) influence people's mental state attribution behaviors to the CUI?
   (d) How do CUIs impact people's expectations and perceptions of CUI compared to non-conversational AI systems?
   (e) How do people perceive and react to CUIs that exhibit ToM-like capabilities through conversations?
   (f) What kind of users (e.g., personality traits) might benefit more from interacting with CUIs equipped with ToM-like capabilities?

(3) **Envisioning MToM in human-CUI interactions**
   (a) How can the vision of MToM inform the design of human-CUI interactions (e.g., turn-taking, mutual shaping of dialogue) in different contexts?
   (b) How does having MToM in human-CUI interactions impact the quality and outcome of human-CUI interactions?
   (c) What can we learn from human-human conversations to inform the design of MToM in human-CUI interactions?
   (d) What might MToM look like in conversations that involve multiple humans and multiple CUIs?
   (e) How can we measure, assess, and evaluate MToM in human-CUI interactions?
   (f) What are the positive and negative consequences of having MToM in human-CUI interactions?

3 Organizers

To encourage interdisciplinary discussions on ToM in human-AI interaction, our workshop organizers come from both academia and industry with research focuses on various relevant disciplines such as CUI, HCI, AI, Design and Cognitive Science. In addition to organizers from the first hybrid ToMinHAI workshop at CHI'2024 which successfully attracted about 40 attendees, we have welcomed new organizers especially from the CUI community to help situate the ToM research discourse within the CUI community discourse. In addition to our workshop organizing experience at the ToMinHAI hybrid workshop at CHI 2024, many of us also have experience participating and organizing workshops at international HCI conferences and internal symposiums at our respective institutions. We will use lessons learned from these experiences to conduct our in-person workshop.

Qiaosi Wang (Chelsea) is a Carnegie Bosch Postdoctoral Fellow at Carnegie Mellon University. She conducts interdisciplinary research on human-AI interaction, cognitive science, and responsible AI. Chelsea recently obtained her Ph.D. in Human-Centered Computing from Georgia Institute of Technology. Chelsea's Ph.D. dissertation proposed, developed, and empirically examined the theoretical framework of Mutual Theory of Mind [34] for human-AI communication, which explores how humans' and AI's interpretations of each other are shaped through continuous communication feedback.

Joel Wester is a PhD Fellow in the Human-Centred Computing group at Aalborg University. He has a background in cognitive science, psychology, and philosophy. Joel's work primarily focuses on AI-powered conversational user interfaces, such as chatbots or robots, and their impact on the wellbeing of everyday citizens. His research particularly centres around people's perceptions of AI systems' behavior and how these perceptions influence their user experiences.

Marvin Pafla is a PhD candidate in the David R. Cheriton School of Computer Science at the University of Waterloo. Drawing on his background in both Psychology and Computer Science, he investigates human–machine interaction, with a particular focus on neural network explainability and its implications for user trust and comprehension. His research explores the notion of "understanding understanding," examining how humans and AI can make sense of each other through a shared Theory of Mind. To advance this work, he studies the ontological foundations that support mutual understanding. Marvin is supervised by Professor Mark Hancock and Professor Kate Larson.

Minha Lee is an Assistant Professor at the Eindhoven University of Technology in the Department of Industrial Design, with a background in philosophy, digital arts, and HCI. Her research concerns morally relevant interactions with various agents like robots or chatbots. Her work explores how we can explore our moral self-identity through conversations with digital entities, e.g., via acting compassionately towards a chatbot. She co-leads the steering committee of the ACM CUI conference series after serving as one of the general chairs of the CUI 2023 conference.

Justin D. Weisz is a Senior Research Scientist, Manager, and Strategy Lead for Human-Centered AI at IBM Research in Yorktown Heights, NY. Dr. Weisz's research sits at the intersection of human-computer interaction (HCI) and artificial intelligence (AI), and he uses a mix of qualitative, quantitative, prototyping, crowdsourcing, and speculative methods to understand how to design AI systems that amplify and augment human capabilities. He was a co-organizer of the HAI-GEN workshops at IUI (2021-2023) and the HCXAI workshop at CHI (2023). Dr. Weisz is the PI of a project that explores how to help people work effectively with generative AI applications. He was appointed as an IBM Master Inventor in 2016, an ACM Senior Member in 2022, and he publishes in top-tier HCI and AI conferences including CHI, IUI, CSCW, AAAI, and NeurIPS. Dr. Weisz received his B.S., M.S., and Ph.D. in Computer Science from Carnegie Mellon University.

Mei Si is an associate professor in the Cognitive Science Department, Rensselaer Polytechnic Institute (RPI) and the graduate program director of the Critical Game Design program at RPI. Mei Si received a Ph.D. in Computer Science from the University of Southern California and an M.A. in Psychology from the University of Cincinnati. Her primary research interests are embodied conversational agents, interactive storytelling, cognitive robots, and AI in games.

4 Website

We will reuse and update our website from our first workshop[1]. The website address will be updated to https://tominhai-cui2025.gi

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

thub.io/ upon acceptance of the workshop. We will disseminate this workshop information and call for proposals through the updated website. We will put up the detailed workshop schedule and publish all the accepted workshop papers on our website upon authors' consent.

5 Pre-Workshop Plans

About two weeks prior to the workshop date, we will post accepted workshop papers, a finalized workshop schedule, speaker and talk descriptions, workshop agenda and other materials on our website. Based on the popular demand from our first ToMinHAI workshop which attracted 40 attendees, we expect about 10-15 participants. We will prioritize workshop registration for authors of accepted papers, then open up the remaining spots (if any) to the broader set of conference attendees on a first-come first-serve basis. To foster community-building prior to the workshop day, we will add workshop participants to our existing ToMinHAI Slack workspace (created for our first ToMinHAI workshop) to help them promote their work and get to know each other.

We will post the call for participation (see section 8) on our ToMinHAI Slack workspace, website, social media, mailing lists in ACM, related professional societies, and organizers' respective institutions, as well as word-of-mouth.

We will request that each submission be limited to 2-6 pages of content using the ACM double-column "sigconf" template; references will not be counted toward the page limit. Authors are welcome to submit in-progress or completed empirical research work as well as position papers or short literature reviews. The organizers will select submissions for inclusion in the workshop. If necessary, we will also assemble a program committee with researchers from both academia and industry to help select submissions. Selection will be based on uniqueness of content, engagement with the themes and topics in the workshop call, and potential for contribution to the research community. We anticipate about 5-10 accepted submissions. All submissions will be subjected to single-blind peer-review by at least two experts from the organizing committee and if necessary, the program committee.

6 During the Workshop

We will host a half-day, in-person workshop that will engage attendees through a variety of activities. In accordance with our aim to promote interdisciplinary discussions and ideas, the program will include interactive paper presentations and group activities. Table 1 provides an overview of the planned 4-hour schedule, with a tentative time frame from 9AM to 1PM local time (Waterloo, ON, Canada).

The opening remarks will introduce the motivation and background of this workshop through a brief overview of ToM research on both the CUI's side and the human's side in human-CUI interactions. We will clarify the workshop's objectives and underscore its interdisciplinary focus. Following the opening, we will facilitate a short ice-breaker activity for the workshop attendees to introduce themselves. While the specific ice-breaker activity is tentative at this point, we want this activity to be interactive and possibly taking the form of speed dating, where each workshop attendees will share their backgrounds, research goals, and their interests in ToM in human-CUI interactions with another workshop attendee, and then quickly switch to another attendee to talk to. By starting with this informal exchange, the workshop will foster a collaborative atmosphere, ensuring that participants feel comfortable contributing ideas and engaging with one another throughout the remainder of the day.

Table 1: Tentative schedule for our ToMinHAI at CUI workshop. The time shown in the table is based on local time of where the conference will be held.

Time	Duration	Session
9:00 AM - 9:10 AM	10 min	Opening Remarks
9:10 AM - 9:30 AM	20 min	Ice Breaker Activity
9:30 AM - 11:00 AM	90 min	Paper Session + QA
11:00 AM - 11:15 AM	15 min	Break
11:15 AM - 12:45 PM	90 min	Group Activity
12:45 PM - 1:00 PM	15 min	Closing Remarks

After the ice breaker, the workshop will transition into paper presentations. These will feature paper talks by authors with accepted workshop papers. Brief Q&A discussions after each presentation will enable attendees to share constructive feedback, position each paper in a broader context, and identify avenues for collaboration across multiple disciplines. The allotted length for each paper talk will be dependent on how many submissions were accepted, but we expect 5-10min for each talk, followed by a 5 min Q&A discussions. Depending on the number of accepted submissions, we will also explore the possibility of hosting a short panel discussion among the presenters to discuss common theme across the paper talks related to ToM and human-CUI interactions.

Once the paper session concludes, participants will take part in the group activity. Inspired by the recent research trend in designing ToM benchmarks for LLMs [18, 27], we are exploring the potential of facilitating the group activity to examine the opportunities and challenges of evaluating ToM in human-CUI interactions, especially in the age of LLM-powered CUIs. Building on the success of our first ToMinHAI workshop at CHI 2024, where participants primarily mapped out grand challenges and emerging directions, this activity will explore how to assess and measure ToM in practice. Groups will discuss potential benchmarks for ToM-like functions in CUIs, consider whether new dimensions of ToM are needed when designing CUIs, and investigate methods for gauging human perception and response to AI systems that exhibit varying degrees of social intelligence. A central question will be whether passing ToM benchmarks truly indicates that an AI possesses ToM, or if existing tools and metrics might oversimplify what counts as "mindreading." Short introductions within each group will help everyone remain aware of individual expertise and shared interests. By using collaborative tools and guided templates that we will provide, attendees will brainstorm concrete scenarios, propose evaluation frameworks, and consider domain-specific constraints, such as how ToM-based assessments might differ in educational versus health contexts.

In the closing portion of the workshop, we will gather everyone to synthesize the day's discussions. The workshop organizers will highlight recurring debates, unexpected points of consensus, and potential paths forward for research on MToM in CUIs. Participants will be invited to articulate final thoughts, share resources, and consider collaborative ventures that extend beyond the confines of this event. These discussions will be documented as part of the workshop summary which we will make available online through the human-centered AI Medium publication (https://medium.com/human-centered-ai).

7 Post-Workshop Plans

First, we are in the process of organizing a special issue for Frontier in AI journal on the topic of "Theory of Mind in Human-AI Interaction" and we plan to invite strong workshop submissions to submit an extended version of their submission to this special issue. Second, we want to continue the discussion with our workshop attendees and build the community around this topic. We already created a Slack group for our CHI 2024 workshop attendees which continues to be active today. We intend to invite attendees for this workshop into the same Slack workspace to expand our community, continue engaging discussions, and strengthen the connections between the CHI and CUI workshop communities. Third, we want our work to reach beyond the HCI and academic community by summarizing workshop discussions and outcomes in an online article. For our ToMinHAI workshop at CHI 2024, the organizers summarized all the paper sessions, panels, and activities in a summary article that was published in the Human-centered AI publication on Medium [37]. We plan to do the same thing for this CUI workshop.

8 Call for Participation

In this workshop (https://tominhai-cui2025.github.io/), we aim to bring together researchers investigating ToM from different research perspectives to define a unifying agenda for Mutual Theory of Mind (MToM) in human-CUI interactions, where both humans and CUIs exhibit ToM-like capabilities throughout their interactions.

We seek submissions that explore three broad topics: (1) building and measuring a CUI's ToM-like capabilities in human-AI conversations, (2) understanding and shaping human ToM in human-CUI interactions, and (3) envisioning MToM in various human-CUI interaction contexts. Beyond conceptual discussions, this year we particularly encourage work that examines how ToM-like abilities can be operationalized, designed, and evaluated, raising questions around which metrics and benchmarks truly capture whether a CUI 'has' ToM. We invite academic and industry researchers to contribute 2–6 page papers (in ACM double-column format) that include position statements, literature reviews, or in-progress empirical studies. Relevant work may involve CUIs taking on diverse social roles (e.g., teammates, tutors) or propose fresh directions for ToM research in human-CUI interaction.

All submissions will be assessed based on their quality and relevance to ToM in human-CUI interaction, and accepted papers will be made available on the workshop website. At least one author of each accepted submission must attend the workshop in person. For further information, contact Qiaosi Wang (Chelsea) at qiaosiw@andrew.cmu.edu.

References

[1] Jaan Aru, Aqeel Labash, Oriol Corcoll, and Raul Vicente. 2023. Mind the gap: challenges of deep learning approaches to Theory of Mind. *Artificial Intelligence Review* (2023), 1–16.

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

[3] Michelle Brachman, Qian Pan, Hyo Jin Do, Casey Dugan, Arunima Chaudhary, James M Johnson, Priyanshu Rai, Tathagata Chakraborti, Thomas Gschwind, Jim A Laredo, et al. 2023. Follow the Successful Herd: Towards Explanations for Improved Use and Mental Models of Natural Language Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 220–239.

[4] Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to tango: Towards theory of AI's mind. *arXiv preprint arXiv:1704.00717* (2017).

[5] Ana Paula Chaves, Jesse Egbert, Toby Hocking, Eck Doerry, and Marco Aurelio Gerosa. 2022. Chatbots Language Design: The Influence of Language Variation on User Experience with Tourist Assistant Chatbots. *ACM Trans. Comput.-Hum. Interact.* 29, 2, Article 13 (Jan. 2022), 38 pages. doi:10.1145/3487193

[6] Coralie Chevallier, Ira Noveck, Francesca Happé, and Deirdre Wilson. 2011. What's in a voice? Prosody as a test case for the Theory of Mind account of autism. *Neuropsychologia* 49, 3 (2011), 507–517.

[7] Fabio Cuzzolin, Alice Morelli, Bogdan Cirstea, and Barbara J Sahakian. 2020. Knowing me, knowing you: theory of mind in AI. *Psychological medicine* 50, 7 (2020), 1057–1061.

[8] Maryam Banitalebi Dehkordi, Reda Mansy, Abolfazl Zaraki, Arpit Singh, and Rossitza Setchi. 2021. Explainability in human-robot teaming. *Procedia Computer Science* 192 (2021), 3487–3496.

[9] Michael A DeVito, Jeremy Birnholtz, Jeffery T Hancock, Megan French, and Sunny Liu. 2018. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[10] Philip R Doyle, Leigh Clark, and Benjamin R Cowan. 2021. What do we see in them? identifying dimensions of partner models for speech interfaces using a psycholexical approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[11] Alexander Dregger, Maximilian Seifermann, and Andreas Oberweis. 2024. Language Cues for Expressing Artificial Personality: A Systematic Literature Review for Conversational Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 22, 17 pages. doi:10.1145/3640794.3665559

[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] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. " I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 153–162.

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

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

[16] O Can Görür, Benjamin Rosman, Fikret Sivrikaya, and Sahin Albayrak. 2018. Social cobots: Anticipatory decision-making for collaborative robots incorporating unexpected human behaviors. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 398–406.

[17] Nikolos Gurney, Stacy Marsella, Volkan Ustun, and David V Pynadath. 2021. Operationalizing theories of theory of mind: A survey. In *AAAI Fall Symposium*. Springer, 3–20.

[18] Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083* (2023).

[19] Jin Joo Lee, Fei Sha, and Cynthia Breazeal. 2019. A Bayesian theory of mind approach to nonverbal communication. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 487–496.

[20] Minha Lee, Gale Lucas, Johnathan Mell, Emmanuel Johnson, and Jonathan Gratch. 2019. What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 38–45. doi:10.1145/3308532.3329465

[21] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. " I hear you, I feel you": encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[22] Zhixin Li, Trisha Thomas, Chi-Lin Yu, and Ying Xu. 2024. "I Said Knight, Not Night!": Children's Communication Breakdowns and Repairs with AI Versus Human Partners. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (Delft, Netherlands) *(IDC '24)*. 781–788. doi:10.1145/3628516.3659394

[23] Dung Nguyen, Phuoc Nguyen, Hung Le, Kien Do, Svetha Venkatesh, and Truyen Tran. 2022. Learning Theory of Mind via Dynamic Traits Attribution. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 954–962.

[24] Giulia Peretti, Federico Manzi, Cinzia Di Dio, Angelo Cangelosi, Paul L Harris, Davide Massaro, and Antonella Marchetti. 2023. Can a robot lie? Young children's understanding of intentionality beneath false statements. *Infant and Child Development* 32, 2 (2023), e2398.

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

[26] Diogo Rato, Marta Couto, and Rui Prada. 2022. Attributing Social Motivations to Changes in Agents' Behavior and Appearance. In *Proceedings of the 10th International Conference on Human-Agent Interaction*. 219–226.

[27] Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding Language Models'(Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. *arXiv preprint arXiv:2306.00924* (2023).

[28] Daniel B Shank, Christopher Graves, Alexander Gott, Patrick Gamez, and Sophia Rodriguez. 2019. Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computers in Human Behavior* 98 (2019), 256–266.

[29] Victor Kenji M. Shiramizu, Anthony J. Lee, Daria Altenburg, David R. Feinberg, and Benedict C. Jones. 2022. The role of valence, dominance, and pitch in perceptions of artificial intelligence (AI) conversational agents' voices. *Scientific Reports* 12, 1 (28 Dec 2022), 22479. doi:10.1038/s41598-022-27124-8

[30] Maayan Shvo, Ruthrash Hari, Ziggy O'Reilly, Sophia Abolore, Sze-Yuh Nina Wang, and Sheila A McIlraith. 2022. Proactive Robotic Assistance via Theory of Mind. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9148–9155.

[31] 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.

[32] Michael T Stuart and Markus Kneer. 2021. Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.

[33] Sebastian Wallkötter, Rebecca Stower, Arvid Kappas, and Ginevra Castellano. 2020. A Robot by Any Other Frame: Framing and Behaviour Influence Mind Perception in Virtual but not Real-World Environments. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. 609–618. doi:10.1145/3319502.3374800

[34] Qiaosi Wang. 2024. *MUTUAL THEORY OF MIND FOR HUMAN-AI COMMUNICATION IN AI-MEDIATED SOCIAL INTERACTION.* Ph. D. Dissertation. Georgia Institute of Technology.

[35] Qiaosi Wang and Ashok K Goel. 2022. Mutual Theory of Mind for Human-AI Communication. *arXiv preprint arXiv:2210.03842* (2022).

[36] 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.

[37] Q. Wang, S. Walsh, M. Si, J. O. Kephart, J. Weisz, and A. Goel. 2024. ToMinHAI 2024: 1st Workshop on Theory of Mind in Human-AI Interaction. *Human-Centered AI on Medium* (19 August 2024). Retrieved 09-Oct-2024 from https://medium.com/human-centered-ai/tominhai-2024-1st-workshop-on-theory-of-mind-in-human-ai-interaction-dc1fd6331716

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

[39] Elmira Yadollahi, Marta Couto, Pierre Dillenbourg, and Ana Paiva. 2022. Do Children Adapt Their Perspective to a Robot When They Fail to Complete a Task?. In *Interaction Design and Children*. 341–351.

[40] Xiao Zhan, Yifan Xu, and Stefan Sarkadi. 2023. Deceptive AI Ecosystems: The Case of ChatGPT. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 3, 6 pages. doi:10.1145/3571884.3603754

[41] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.