Theory of Mind in Human-AI Interaction

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

Sarah E. Walsh
Georgia Institute of Technology
Atlanta, GA, USA
sewalsh@gatech.edu

Mei Si
Rensselaer Polytechnic Institute
Troy, NY, USA
SIM@rpi.edu

Jeffrey O. Kephart
IBM Research
Yorktown Heights, NY, USA
kephart@us.ibm.com

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

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

ABSTRACT

Theory of Mind (ToM), humans' capability of attributing mental states such as intentions, goals, emotions, and beliefs to ourselves and others, has become a concept of great interest in human-AI interaction research. Given the fundamental role of ToM in human social interactions, many researchers have been working on methods and techniques to equip AI with an equivalent of human ToM capability to build highly socially intelligent AI. Another line of research on ToM in human-AI interaction seeks to understand people's tendency to attribute mental states such as blame, emotions, and intentions to AI, along with the role that AI should play in the interaction (e.g. as a tool, partner, teacher, facilitator, and more) to align with peoples' expectations and mental models. The goal of this line of work is to distill human-centered design implications to support the development of increasingly advanced AI systems. Together, these two research perspectives on ToM form an emerging paradigm of "Mutual Theory of Mind (MToM)" in human-AI interaction, where both the human and the AI each possess the ToM capability. This workshop aims to bring together different research perspectives on ToM in human-AI interaction by engaging with researchers from various disciplines including AI, HCI, Cognitive Science, Psychology, Robotics, and more to synthesize existing research perspectives, techniques, and knowledge on ToM in human-AI interaction, as well as envisioning and setting a research agenda for MToM in human-AI interaction.

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

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3636308

ACM Reference Format:
Qiaosi Wang, Sarah E. Walsh, Mei Si, Jeffrey O. Kephart, Justin D. Weisz, and Ashok K. Goel. 2024. Theory of Mind in Human-AI Interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3613905.3636308

1 MOTIVATION

Theory of Mind (ToM) [2, 15, 27] 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 [e.g. 1, 6, 11, 37, 39]. In human-human interaction, 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 [27, 37] so that we could behave accordingly. Given the fundamental role of ToM in human social interactions, many AI researchers believe that *equipping AI with an equivalent of humans' ToM capability* is the key to building AI agents with heightened levels of social intelligence for them to work, play, and live with humans [5, 6, 37]. This vision has inspired a number of efforts to design and build a ToM-like capability for AI using different techniques, including recognizing and modeling people's non-verbal cues [21], emotional expressions [21], as well as people's beliefs, plans [32], and intents [16]. These studies typically leverage techniques such as machine learning (e.g., Bayesian network) [16, 21], computer vision [7], and cognitive modeling [16, 17, 25, 33] in contexts such as human-AI decision-making [16], human-AI collaborations [7, 11, 22], and multi-agent interactions [28, 33].

Whether AI can have a ToM capability, as well as how we should talk about ToM in AI, is a controversial topic in academic discourse. Some scholars argue that describing a machine's capability using the vocabulary of a uniquely human capability risks the danger of anthropomorphizing AI and misleading the public [31]. Some scholars argue that current AI systems may already possess some aspects of a ToM-like capability, given that certain advanced AI systems are already capable of making inferences about our beliefs, emotions, and intentions with relatively high accuracy. Other scholars have adopted a stronger stance by making controversial claims that ToM, a uniquely human capability, has spontaneously emerged in large language models (LLMs) [4, 20] [c.f. 29] as some of those models were able to pass different versions of the false-belief tasks used to assess ToM capability in children.

Alternatively, other researchers are examining ToM in human-AI interaction by focusing on *humans' ToM* when interacting with AI. There has been a lot of work done to understand humans' perceptions [37], mental models [3, 9, 13], and folk theories [8, 12] of AI. As AI sometimes gives the illusion of having an "(artificial) mind", researchers have also begun to examine people's reactions and engagements with their perceptions of such an artificial mind [e.g. 30]. Other work has explored people's tendencies to attribute human mental states such as blame [34], emotions [30], perspectives [40], intentions [26], and social motivations [28] to AI. Finally, some researchers are examining how different framings of an AI's role – as a collaborator or partner, a teacher, a facilitator, and more – impact peoples' perceptions and interactions with it and how to design effective human-AI teams [18, 19, 24]. As machines with a ToM-like capability are being developed, understanding people's ToM when interacting with AI systems that seemingly have a "(artificial) mind" offers critical insights into how such AI systems should be designed from a human-centered viewpoint.

Putting together these two perspectives of research on ToM in human-AI interaction, from the AI's side and from the human's side, there is an emerging paradigm that we call "Mutual Theory of Mind (MToM)" in human-AI interaction [36], where both the human and the AI possess the capability of ToM and continuously make inferences and attribute mental states to each other during an interaction. Although enabling MToM in human-AI interaction promises to make a great impact on achieving human-level 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 AI parties, have yet to be envisioned by the HCI and AI research communities.

2 WORKSHOP GOALS

The goal of this workshop is to bring together researchers from various disciplines studying ToM in human-AI interaction from the AI's ToM point of view and the human's ToM point of view to define a unifying research agenda on the human-centered design and development of MToM in human-AI interaction. This workshop will provide a platform for researchers to discuss techniques to build ToM-like capability in AI, as well as implications for designing AI based on human's ToM during human-AI interaction. Additionally, this workshop will also look at the phenomenon of MToM in human-AI interaction by envisioning the design and development of the interaction dynamic of MToM in human-AI interaction, as well as critically examining the consequences of having MToM in human-AI interaction. To support interdisciplinary discussions, we invite academic and industry researchers in disciplines including but not limited to cognitive science, AI, HCI, design, robotics, psychology, communication studies, and more to submit work that will inform our understanding of MToM in human-AI interaction.

For the purpose of this workshop, we define AI systems broadly to include any algorithmic-driven technical systems of varying complexity across different application contexts. Given the recent discourse on ToM in LLMs, we especially welcome submissions that discuss ToM and generative AI. 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-AI 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 interactions or other interaction contexts to help shape the discourse around the implications of MToM in human-AI interaction contexts.

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

(1) **Designing and building an AI's ToM-like capability**
   (a) What techniques, methods, models, and data can be used to build an AI's ToM-like capability? (e.g., machine learning techniques, cognitive models, prompt design for LLMs)
   (b) What information belongs in an AI's ToM model of a user, and how are such information updated? To what extent is an AI system able to adapt or personalize its responses to a user based on this information?
   (c) How can we measure, assess, and evaluate an AI's ToM-like capability?
   (d) What factors from an AI's design (e.g., physical appearance or voice) could influence people's perceptions of an AI's (artificial) mind?
   (e) What does it mean to design an AI's ToM-like capability in an ethical and human-centered manner?

(2) **Understanding and shaping humans' ToM in human-AI interaction**
   (a) What kind of mental states (e.g., beliefs, intentions, blame) do people attribute to AI? Why do people attribute mental states to AI?
   (b) How does people's attribution of mental states to AI relate to anthropomorphizing AI?
   (c) How does the role framing of an AI system (e.g., as a tool, partner, teacher, facilitator and more) impact people's expectations and perceptions for that system? What AI roles are appropriate for which use cases?
   (d) How do people perceive and react to AI systems that display ToM-like capabilities? What happens when the AI's ToM about the users is not accurate?

(3) **Envisioning MToM in human-AI interactions**
   (a) What will the interaction dynamic look like when having MToM in a human-AI interaction?
   (b) How does having MToM in human-AI interaction impact the quality of human-AI team outcomes?
   (c) What are the positive and negative consequences of having MToM in a human-AI interaction?

3 ORGANIZERS

In order 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 AI, HCI, Cognitive Science, and Robotics. We have collective experience in conducting online workshops (e.g. [10, 14, 23, 38]) and symposiums (e.g., [35]). In addition, many of us have experience participating and organizing hybrid conferences

and meetings at our respective institutions. We will use lessons learned from these experiences to conduct the workshop.

Qiaosi Wang (Chelsea) is a Ph.D. candidate in Human-Centered Computing at Georgia Institute of Technology. She conducts interdisciplinary research on human-AI interaction, cognitive science, and computer-supported cooperative work. Chelsea's Ph.D. dissertation work focuses on developing and empirically examining the theoretical framework of Mutual Theory of Mind [37] for human-AI communication, which explores how humans' and AI's perceptions of each other evolve through back-and-forth communications.

Sarah E. Walsh is a Robotics Ph.D. candidate at the Daniel Guggenheim School of Aerospace Engineering at Georgia Tech. She received her B.S. in Mathematics from Stockton University and her B.S. in Mechanical Engineering at Rutgers University. Her research focuses on the development of shared mental models at the intersection of AI interpretability and human behavior analysis to improve human-AI collaboration in team decision-making tasks.

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.

Jeffrey O. Kephart is a Distinguished Research Staff Member at the IBM Thomas J. Watson Research Center in New York. He received a B.S. in Electrical Engineering and Engineering Physics from Princeton University and a Ph.D. in Electrical Engineering (physics minor) from Stanford University. He leads an effort on multi-modal AI assistants that interact with humans via voice and gesture for data visualization, analytics, and decision making tasks, which have received Best Demo awards at AAAI and IJCAI. Kephart is an IEEE Fellow and a member of IBM's Academy of Technology.

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.

Ashok K. Goel is a Professor of Computer Science and Human-Centered Computing in the School of Interactive Computing at Georgia Institute of Technology, and the Chief Scientist with Georgia Tech's Center for 21st Century Universities. He is the Executive Director of the National AI Institute for Adult Learning and Online Education. He is a Fellow of AAAI and the Cognitive Science Society. Ashok's current research interests include AI agent's theory of mind of humans and itself, self-explanation, machine teaching, and mutual theory of mind between humans and AI agents.

4 WEBSITE

We will disseminate our workshop information and call for proposals through our website[1]. 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, pre-recorded paper talks, a finalized workshop schedule, speaker and talk descriptions, workshop agenda and other materials on our website. We will ask authors of accepted papers to record short paper talks and make them available prior to the workshop date to facilitate hybrid participation. We expect about 20 to 30 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 start a Slack or Discord channel to help participants promote their work and get to know each other.

We will post the call for participation on our website, social media, mailing lists in ACM, EUSSET, related professional societies, and organizers' respective institutions, as well as word-of-mouth. We are also assembling a program committee with researchers from both academia and industry to help us disseminate the call for participation message and submission review.

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 OC and PC will select submissions for inclusion in the workshop. 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 10-20 accepted submissions. All submissions will be subjected to single-blind peer-review by at least three experts from the organizing committee and the program committee.

6 DURING THE HYBRID WORKSHOP

We will hold a one-day hybrid workshop that will engage both in-person and remote attendees through a variety of activities. In accordance with our aim to promote interdisciplinary discussions and ideas, we plan to hold multiple sessions including a keynote, two paper sessions, two group activities, and an invited panel (with

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

breaks in between). We show a tentative 7.5-hour workshop schedule in Table 1, with a tentative time frame from 9 am to 4:30 pm local time where the conference will be held.

After brief welcome notes by the workshop organizers, the workshop will begin with a keynote to provide an initial exploration and overview of ToM in human-AI interaction. While we have not finalized the invited keynote speaker, the invited speaker will have extensive research experience and understanding of the research landscape around ToM in human-AI interaction, preferably with interdisciplinary research experience and background.

We will have two paper sessions (listed as "Paper Session I" and "Paper Session II" in Table 1) for authors to give short presentations of their accepted work and answer questions from the audience. To promote hybrid participation, we are considering playing prerecorded paper talks at the session and running a live moderated panel discussion with all the paper authors in each session, instead of individual Q&A, depending on how well the paper topics align with each other. The duration of each paper presentation will be determined by the total number of papers accepted.

Following each paper session will be group activities (listed as "Group Activity I" and "Group Activity II" in Table 1) aimed at facilitating discussions around MToM in human-AI interactions or other emerging topics from the submissions we received. The theme of the group activities will be to identify and address a grand challenge in MToM in human-AI interaction through small group discussions. These group discussions will be hybrid, participated by both in-person and remote attendees. Each group will be facilitated by at least one of the workshop organizers. At the beginning of each group activity, we will facilitate some short ice breaker activities (5 min) for the group to get to know each other and their work. Then the group will spend around 15 minutes to brainstorm and identify grand challenges in MToM in human-AI interaction, such as how to operationalize MToM or how to mitigate anthropomorphism enabled by MToM. Each group will then focus on one grand challenge and come up with solutions to address this challenge. These solutions could be proposed AI techniques, benchmarks, policies, research directions, etc. To structure the group discussions, we will encourage participants to use methods such as affinity diagrams and concept mapping through virtual collaboration tools (e.g., Miro and Mural) provided by us. Encouraging the use of these virtual collaboration tools will also help facilitate interactions and communications among the in-person and remote attendees within each small groups. Having two rounds of group activities will encourage participants to engage in discussions on the various topics and research opportunities and challenges regarding MToM in human-AI interaction with a broader group of interdisciplinary researchers.

We will host a panel discussion by inviting external panelists who are familiar with research in ToM and human-AI interaction. The panel discussion topics will draw on the workshop sub-topics that we outlined in the previous section, as well as other topics that emerged from the workshop submissions.

To facilitate engagements between the in-person and remote attendees, the keynote, the paper sessions, and the panel will all be live streamed to the remote attendees via online meeting software (e.g., Zoom) with live captioning enabled. If additional needs were brought up by the attendees (e.g., signing interpreters, in-room captioners), we will try to accommodate these needs by working with the workshop program chairs. Some members of our organizing committee have indicated that they would be attending remotely, and hence able to facilitate and monitor activities on the virtual meeting platform. The group activities will also be facilitated in a hybrid fashion by encouraging the use of virtual collaboration tools to facilitate interactions between remote and in-person attendees.

We will record (upon consent from workshop participants) the keynote, paper sessions, as well as the panel session and put the recordings up on our website to share with the broader research community and the public.

Table 1: Tentative schedule for the one-day hybrid 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 - 9:20 AM	20 min	Welcome
9:20 AM - 10:00 AM	40 min	Opening Keynote
10:00 AM - 11:00 AM	60 min	Paper Session I
11:00 AM - 11:15 AM	15 min	Coffee Break
11:15 AM - 12:00 PM	45 min	Group Activity I
12:00 PM - 1:00 PM	60 min	Lunch
1:00 PM - 2:00 PM	60 min	Paper Session II
2:00 PM - 2:15 PM	15 min	Coffee Break
2:15 PM - 3:00 PM	45 min	Group Activity II
3:00 PM - 4:00 PM	60 min	Invited Panel
4:00 PM - 4:30 PM	30 min	Closing Remarks

7 POST-WORKSHOP PLANS

First, we hope to organize a special issue on the topic of "Mutual Theory of Mind in Human-AI Interaction" in an HCI or HAI journal venue. We plan to invite strong workshop submissions to expand on their work to submit to this special issue. Second, we want to continue the discussion with our workshop attendees and build the community around this topic. To do this, we plan to start a mailing list/Slack group for our workshop attendees to post relevant updates, news, and encourage them to invite others who are also working on this topic to the group. Third, we also want to reach a broader audience to continue the discussion. Hence we hope to summarize the workshop discussions and outcomes in an online article that could be published in the Human-Centered AI publication on Medium (https://medium.com/human-centered-ai). We also want to share the recordings of some portions of this workshop, e.g., opening keynote and paper presentations, on YouTube and publicize the recordings on social media to reach a broader audience outside of the academic community.

8 CALL FOR PARTICIPATION

Theory of Mind (ToM) refers to humans' capability of attributing mental states such as goals, emotions, and beliefs to ourselves and others. This concept has become of great interest in human-AI interaction research. In this hybrid workshop (https://theoryofmindinhaichi2024.wordpress.com), we seek to bring together researchers working on different perspectives of ToM in human-AI interaction to define a unifying research agenda for Mutual Theory of Mind

(MToM) in human-AI interaction (i.e., where both humans and AI have ToM during interactions) through interdisciplinary discussions. We aim to explore three broad topics to inspire workshop discussions: (1) designing and building AI's ToM-like capability, (2) understanding and shaping human's ToM in human-AI interaction, (3) envisioning MToM in human-AI interaction. We encourage academic and industry researchers from various disciplines to contribute 2-6 pages ACM double-column format position papers, literature reviews, or in-progress empirical studies to shape the discourse around ToM in human-AI interaction. We welcome submissions that discuss ToM and advanced AI systems that give the illusion of "having a mind" such as large language models, as well as submissions that expand or propose new definitions of ToM in human-AI interaction. Papers should be submitted via EasyChair and will be evaluated based on quality and relevance to ToM in human-AI interaction. Upon acceptance, papers will be published on the workshop website. At least one author of each accepted submission must attend the workshop and all participants must register for both the workshop and for at least one day of the conference. For more information contact theoryofmindinhaichi24@easychair.org.

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] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[5] 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).

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Upol Ehsan, Philipp Wintersberger, Elizabeth A Watkins, Carina Manger, Gonzalo Ramos, Justin D Weisz, Hal Daumé Iii, Andreas Riener, and Mark O Riedl. 2023. Human-Centered Explainable AI (HCXAI): Coming of Age. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[11] 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*.

[12] 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.

[13] 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.

[14] Werner Geyer, Lydia B Chilton, Justin D Weisz, and Mary Lou Maher. 2021. Haigen 2021: 2nd workshop on human-ai co-creation with generative models. In *26th International Conference on Intelligent User Interfaces-Companion*. 15–17.

[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] Stephanie Houde, Siya Kunde, and Rachel Bellamy. 2023. Envisioning Generative AI Interaction for Collaborative Design. *Proceedings of DIS 2023 Workshops: Towards a Design (Research) Framework with Generative AI* (2023).

[19] Taenyun Kim, Maria D Molina, Minjin Rheu, Emily S Zhan, and Wei Peng. 2023. One AI Does Not Fit All: A Cluster Analysis of the Laypeople's Perception of AI Roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

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

[21] 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.

[22] Shih-Yun Lo, Elaine Schaertl Short, and Andrea L Thomaz. 2020. Planning with partner uncertainty modeling for efficient information revealing in teamwork. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 319–327.

[23] Mary Lou Maher, Justin D Weisz, Lydia B Chilton, Werner Geyer, and Hendrik Strobelt. 2023. HAI-GEN 2023: 4th Workshop on Human-AI Co-Creation with Generative Models. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 190–192.

[24] Christopher McComb, Peter Boatwright, and Jonathan Cagan. 2023. FOCUS AND MODALITY: DEFINING A ROADMAP TO FUTURE AI-HUMAN TEAMING IN DESIGN. *Proceedings of the Design Society* 3 (2023), 1905–1914.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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).

[30] 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.

[31] Ben Shneiderman and Michael Muller. 2023. On AI Anthropomorphism. *Human-Centered AI* (2023). https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae

[32] 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.

[33] 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.

[34] 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.

[35] Qiaosi Wang. 2023. *Mental Model in Human-AI Interaction.* Retrieved 2023-10-04 from http://qiaosiwang.me/mentalmodel-symposium-schedule.pdf

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

[37] 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.

[38] Justin D Weisz, Mary Lou Maher, Hendrik Strobelt, Lydia B Chilton, David Bau, and Werner Geyer. 2022. Hai-gen 2022: 3rd workshop on human-ai co-creation with generative models. In *27th International Conference on Intelligent User Interfaces*. 4–6.

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

[40] 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.