[MM/AI] Mental Models in Human-AI Interaction: Methods and Challenges in the Generative and Agentic AI Era (Workshop)

TÉO SANCHEZ, Ludwig Maximilian University, Germany
BHADA YUN, ETH Zürich, Switzerland
PRERNA RAVI, CSAIL, Massachusetts Institute of Technology, USA
LAURA SCHÜTZ, Technical University of Munich, Germany
ANNA NEUMANN, Research Center Trust, University Duisburg-Essen, Germany
ROBIN SHING MOON CHAN, ETH Zürich, Switzerland
APRIL YI WANG, ETH Zurich, Switzerland
QIAOSI (CHELSEA) WANG, HCII, Carnegie Mellon University, USA
SUMIT ASTHANA, Microsoft, USA

The mental model construct is widely used in HCI to refer to the knowledge structure people hold in order to reason about and interact with computing systems. Yet it is often operationalized intuitively: the construct is often used interchangeably with related concepts (e.g. folk theories, sensemaking) and methods of studying it (e.g. through elicitation) are many and diverse, with each method resting on distinct assumptions about what counts as a mental model. Generative and agentic AI systems may further complicate mental model formation and elicitation as such systems are opaque by design and increasingly act on users' behalf across files, applications, and on the web. Together, these challenges may hinder the commensurability of research on people's mental models of AI systems. The MM/AI workshop calls for a critical reassessment of how we understand and study mental models in human-AI interaction research. It aims to foster theoretical and methodological exchange on mental models in human–AI interaction, identify open challenges, and develop directions for future research. We invite short papers on users' or stakeholders' mental models of AI systems, particularly contributions that reflect on the conceptual and methodological foundations of the construct. The half-day workshop combines lightning talks, hands-on elicitation exercises, and structured discussions on key questions concerning the future of the mental model for human-AI interaction research.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; **HCI theory, concepts and models**.

Additional Key Words and Phrases: mental models, human-AI interaction, elicitation methods, generative AI, agentic AI

ACM Reference Format:
Téo Sanchez, Bhada Yun, Prerna Ravi, Laura Schütz, Anna Neumann, Robin Shing Moon Chan, April Yi Wang, Qiaosi (Chelsea) Wang, and Sumit Asthana. 2027. [MM/AI] Mental Models in Human-AI Interaction: Methods and Challenges in the Generative and Agentic

Authors' Contact Information: Téo Sanchez, Ludwig Maximilian University, Munich, Germany, teo.sanchez@lmu.de; Bhada Yun, ETH Zürich, Zürich, Switzerland, bhayun@ethz.ch; Prerna Ravi, CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA, prernar@mit.edu; Laura Schütz, Technical University of Munich, Munich, Germany, laura.schuetz@tum.de; Anna Neumann, Research Center Trust, University Duisburg-Essen, Duisburg, Germany, ; Robin Shing Moon Chan, ETH Zürich, Zürich, Switzerland, chanr@ethz.ch; April Yi Wang, april.wang@inf.ethz.ch, ETH Zurich, Zurich, Switzerland; Qiaosi (Chelsea) Wang, qiaosiwatandrew.cmu.edu, HCII, Carnegie Mellon University, Pittsburgh, USA; Sumit Asthana, sumitasthana@microsoft.com, Microsoft, Redmond, USA.

Manuscript submitted to ACM

AI Era (Workshop). In *32nd International Conference on Intelligent User Interfaces (IUI '27), February 8–11, 2027, Helsinki, Finland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Description

Background and motivation The use of the mental model construct in human-AI interaction research, including at IUI, has grown steadily since 2018 [17]. Borrowed from cognitive psychology [7], it refers to people's internal representation of a system, constructed through interaction and shaped by prior knowledge and expectations [4]. Mental models support people's ability to reason about and predict system behavior, and ultimately guide their actions. HCI and human factors research has shown that understanding a system's inner mechanisms can improve performance, learning, retention, skill transfer, and satisfaction [9, 10, 15, 20]. Conversely, knowing what mental models users actually hold is a means to improve the system's design. This has motivated work on mental model elicitation, i.e., capturing people's reasoning about a system in a form that can be examined by others [2]. However, the construct's relevance for our research community faces three important challenges. First, studies elicit mental models in heterogeneous ways: *researcher-led verbalizations* such as interviews and open-text questions [3, 16, 19], *participant-led verbalizations* such as think-aloud and teach-back [5, 6], *structured judgments* such as comprehension scales and prediction tests [1, 14, 18], and *artifact-based methods* such as free drawing and card sorting [11–13]. Each method rests on different assumptions about what constitutes a mental model, making it difficult to compare the results. Second, the construct is often used interchangeably with adjacent concepts such as folk theories and sensemaking [8], blurring what exactly is being studied. Third, and most importantly, generative and agentic AI systems may challenge mental model formation and elicitation: these systems are opaque by design, adaptive, and increasingly act on users' behalf across files, applications, and the web. Misconceptions about their behavior can have concrete consequences. For instance, a user believing a coding agent cannot read files listed in a .GITIGNORE may inadvertently send sensitive data to a third-party provider. **These three challenges may ultimately hinder the legibility and commensurability of mental model research at IUI and beyond. We argue it is therefore timely to dedicate a space for theoretical and methodological exchange on the construct at IUI '27.**

Workshop goals This workshop aims to foster dialogue among researchers and practitioners about mental models in light of contemporary generative and agentic AI. It offers a venue for exchanging first-hand methodological experiences and perspectives on how people's mental models of these systems can or should be studied. The workshop's objectives are threefold:

Obj 1. Map elicitation practices. Document and compare how participants conceptualize and study mental models of AI systems across different research contexts, through their paper submissions.

Obj 2. Operationalize mental models for AI. Surface methodological challenges and opportunities associated with studying mental models of generative and agentic AI through participants' first-hand research experiences, as reflected in their submissions, and through hands-on elicitation exercises.

Obj 3. Reflect on the future of the construct. Advance discussion on how the mental model construct should be dynamically conceptualized and applied to remain relevant in the context of contemporary AI systems through structured group discussions.

Target audience The workshop targets researchers and practitioners, from academia or industry, ideally with first-hand experience with people's mental model of AI in the context of HCI.

2 Previous history

This is the inaugural edition of the workshop under the proposed organizing team.

3 Organizers

The organizing team combines first-hand research experience about mental models in human-AI interaction, experience in community-building activities, and a strong motivation to reopen and shape future research on these topics within the IUI community.

Name	Affiliation	Email	Web Page	Role	Attend.
Téo Sanchez	LMU Munich	teo.sanchez@lmu.de	https://teo-sanchez.github.io	Main contact	Yes
Bhada Yun	ETH Zürich	bhayun@ethz.ch	https://bhadayun.com	Co-organizer	Yes
Prerna Ravi	MIT	prernar@mit.edu	https://prernaravi.com	Co-organizer	Yes
Laura Schütz	TUM	laura.schuetz@tum.de	https://lauraschuetz.github.io/	Co-organizer	Yes
Anna Neumann	University Duisburg-Essen	anna.neumann1@uni-due.de	https://annaneumann.carrd.co/	Co-organizer	TBD
Robin Chan	ETH Zürich	chanr@ethz.ch	https://chanr0.github.io/	Co-organizer	Yes
April Yi Wang	ETH Zürich	april.wang@inf.ethz.ch	https://peachlab.inf.ethz.ch/	Co-organizer	Yes
Qiaosi Wang	CMU HCII	qiaosiw@andrew.cmu.edu	https://www.qiaosiwang.me/	Co-organizer	Yes
Sumit Asthana	Microsoft	sumitasthana@microsoft.com	https://sumitasthana.xyz/	Co-organizer	Yes

Table 1. Workshop organizers. Short biographies are provided in appendix A.

4 Workshop Program Committee.

Submissions will undergo double-blind peer review on OpenReview, supervised by a program committee. Each submission will receive at least two reviews from a pool of reviewers comprising co-organizers and externally recruited researchers, with conflicts of interest managed through OpenReview. Program committee members will oversee review quality and making final recommendations. Besides quality, reviews will prioritize relevance to the workshop themes, methodological and theoretical insight, and potential to stimulate discussion. Provisional members of the program committee are: **Prof. Dr. Simone Stumpf** (University of Glasgow), **Magdalena Wischnewski** (University of Duisburg-Essen), **Prof. Dr. April Yi Wang** (ETH Zürich), **Dr. Qiaosi (Chelsea) Wang** (Carnegie Mellon University), and **Dr. Sumit Asthana** (Microsoft research).

5 Participants

We expect 25–30 participants, including organizers, authors of accepted submissions, and attendees admitted through open registration. We welcome researchers and practitioners from HCI, AI, psychology, cognitive science, design, recruited through call for paper posted on a dedicated workshop website[1], and disseminated through ACM SIGCHI communication channels, the organizers' professional networks, and social media. We invite short papers operationalizing mental models in human-AI interaction, particularly contributions discussing its conceptual and methodological foundations.

[1]https://mentalmodelsofai.com/

6 Workshop activities

The program is designed to progressively move from sharing participants' insight and identifying methodological challenges through paper presentations and hands-on exercises, toward reflecting on the future of the mental model construct in human-AI interaction through structured group discussion.

0:00–0:15 — Introduction (Dr. Téo Sanchez): The workshop will begin with a brief introduction to its goals and themes, followed by a short genealogy of the mental model construct, tracing its origins in cognitive psychology and its adoption within HCI and human-AI interaction research.

0:15–0:50 — Selected lightning talks: Five submissions selected for their quality and diversity will be invited to present their work in a 5-minute lightning format, followed by 2 minutes for questions. The objective is to highlight the variety of ways mental models are conceptualized and operationalized across current research, while providing a common set of examples and points of reference for the subsequent activities.

0:50–1:30 — Peer elicitation exercise: Participants will work in pairs where one will elicit the other's mental model of an AI system, focusing on one of two systems to facilitate comparison across groups: (1) a coding agent (e.g., Claude Code) or (2) a consumer-oriented text-to-image generator (e.g., DALL-E), two widely adopted and but contrasting forms of contemporary AI systems. Participants acting as elicitors will be free to choose among several elicitation approaches presented by the organizers. Method descriptions and supporting materials will be provided.

1:30–2:00 — Coffee break. If participants wish, the previous activity may continue during the coffee break.

2:00–3:00 — Breakout discussions: Participants will be assigned to breakout groups, each focusing on a different open question concerning the future of mental model research in the context of contemporary AI systems. Groups will draw on insights from the presentation, the elicitation exercise, and their own research experience. The goal is not necessarily to reach consensus, but to articulate competing perspectives and promising research directions. The questions under consideration are, and are not limited to: **Q1. Operationalization methods.** Can mental model elicitations be meaningfully compared within and across studies? What assumptions underlie different elicitation methods? **Q2. Contemporary AI systems.** How do contemporary AI systems challenge existing assumptions about mental models and their elicitation? What are (un)suitable class of methods for these systems? **Q3. Conceptual boundaries.** What makes the mental model construct distinctive from adjacent constructs such as folk theories, sensemaking, schemata, and anthropomorphism? **Q4. Dymanics.** How can researchers capture the dynamic nature of mental models given mental models continue to evolve through interaction with a system? **Q5. Research and design value.** What insights does mental model elicitation provide for the design of intelligent user interfaces? Could they be obtained through other approaches?

3:00–3:30 — Report-back and closing discussion: Each breakout group will briefly report their insights, points of disagreement, and the priorities they foresee for future research.

7 Planned outcomes

A public Zenodo archive will gather accepted papers, workshop materials, discussion summaries, and consented elicitation artifacts. The workshop website will stay online after the event and serve as a hub for accessing these materials. We also plan to produce a short workshop report synthesizing key insights and future research directions identified during the discussions. The report will be deposited on Zenodo and HAL.

8 Length

Half-day

References

[1] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 2 (2020), 1–37. doi:10.1145/3366485

[2] Robert W. Andrews, J. Mason Lilly, Divya Srivastava, and Karen M. Feigh. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science* 24, 2 (2023), 129–175. doi:10.1080/1463922X.2022.2061080

[3] Michelle Brachman, Siya Kunde, Sarah Miller, Ana Fucs, Samantha Dempsey, Jamie Jabbour, and Werner Geyer. 2025. Building Appropriate Mental Models: What Users Know and Want to Know about an Agentic AI Chatbot. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 247–264. doi:10.1145/3708359.3712071

[4] John M. Carroll and Judith Reitman Olson. 1988. Chapter 2 - Mental Models in Human-Computer Interaction. In *Handbook of Human-Computer Interaction*, MARTIN HELANDER (Ed.). North-Holland, Amsterdam, 45–65. doi:10.1016/B978-0-444-70536-5.50007-5

[5] Giuseppe Desolda, Andrea Esposito, Florian Müller, and Sebastian Feger. 2023. Digital Modeling for Everyone: Exploring How Novices Approach Voice-Based 3D Modeling. In *Human-Computer Interaction – INTERACT 2023*, José Abdelnour Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno, and Marco Winckler (Eds.). Springer Nature Switzerland, Cham, 133–155.

[6] Aike C. Horstmann, Clara Strathmann, Lea Lambrich, and Nicole C. Krämer. 2023. Alexa, What's Inside of You: A Qualitative Study to Explore Users' Mental Models of Intelligent Voice Assistants. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (Würzburg, Germany) *(IVA '23)*. Association for Computing Machinery, New York, NY, USA, Article 5, 10 pages. doi:10.1145/3570945.3607335

[7] Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness.* Number 6. Harvard University Press.

[8] Natalie A. Jones, Helen Ross, Timothy Lynam, Pascal Perez, and Anne Leitch. 2011. Mental Models: An Interdisciplinary Synthesis of Theory and Methods. *Ecology and Society* 16, 1 (2011). doi:10.5751/ES-03802-160146

[9] David E. Kieras and Susan Bovair. 1984. The role of a mental model in learning to operate a device. *Cognitive Science* 8, 3 (1984), 255–273. doi:10.1016/S0364-0213(84)80003-8

[10] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/2207676.2207678 event-place: Austin, Texas, USA.

[11] Johannes Kunkel, Thao Ngo, Jürgen Ziegler, and Nicole Krämer. 2021. Identifying Group-Specific Mental Models of Recommender Systems: A Novel Quantitative Approach. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer International Publishing, Cham, 383–404.

[12] Sunok Lee and Sangsu Lee. 2022. "Hey Alexa, Where Are You?" A Drawing Study to Understand Users' Mental Models of the Environments Surrounding Conversational Agents. In *[ ] With Design: Reinventing Design Modes*, Gerhard Bruyns and Huaxin Wei (Eds.). Springer Nature Singapore, Singapore, 2001–2016.

[13] Liping Li, Hsinwen Chang, Weihan Sun, Jin Guo, and Jianchao Gao. 2020. How Drivers Categorize ADAS Functions. In *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments*, Pei-Luen Patrick Rau (Ed.). Springer International Publishing, Cham, 606–615. doi:10.1007/978-3-030-49788-0_46

[14] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. doi:10.1145/3397481.3450639

[15] Stephen Payne, Helen Squibb, and Andrew Howes. 1990. The Nature of Device Models: The Yoked State Space Hypothesis and Some Experiments With Text Editors. 5, 4 (1990), 415–444. doi:10.1207/s15327051hci0504_3

[16] Savvas Petridis, Michael Xieyang Liu, Alexander J. Fiannaca, Carrie J. Cai, and Michael Terry. 2026. Compass vs Railway Tracks: Unpacking User Mental Models for Communicating Long-Horizon Work to Humans vs. AI. In *Proceedings of the 2026 Designing Interactive Systems Conference (DIS '26)*. Association for Computing Machinery, New York, NY, USA, 1188–1204. doi:10.1145/3800645.3812956

[17] Téo Sanchez, Oleksandra Vereschak, and Ophelia Deroy. 2026. Mental Models in Human-AI Interaction: Systematic Review of Empirical Methodologies and Guidelines. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*. 663–682.

[18] Vaynee Sungeelee, Nathanaël Jarrassé, Téo Sanchez, and Baptiste Caramiaux. 2024. Comparing Teaching Strategies of a Machine Learning-based Prosthetic Arm. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 715–730. doi:10.1145/3640543.3645170

[19] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 31–40. doi:10.1145/1240624.1240630

[20] Richard M. Young. 2014. Surrogates and mappings: Two kinds of conceptual models for interactive devices. In *Mental models*. Psychology Press, 35–52. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315802725-4/surrogates-mappings-two-kinds-conceptual-models-interactive-devices-richard-young

A Biographical and research backgrounds

Téo Sanchez is a MSCA Postdoctoral Fellow at the Munich Interactive Intelligence Initiative (MI3), LMU Munich. His HCI research rethinks users as proactive machine *teachers* and investigates the psychological consequences of framing human-AI interaction as a teacher–student relationship. He recently conducted a systematic review of 88 empirical studies on mental models in human-AI interaction [17], which contributed to ground the framing of this workshop. His work has received awards at IUI 2022 (best paper), C&C 2023 (honorable mention), as well as the AFIHM Best Ph.D. Dissertation Award in 2023.

Bhada Yun is a PhD student at ETH Zürich studying Machine Intelligence and Visual & Interactive Computing. His work develops phenomenological methods for studying human-AI interaction, eliciting how people subjectively perceive, make sense of, and relate to AI systems, with the aim of aligning AI to human wellbeing and agency. His research has been recognized with four CHI Honorable Mention Awards.

Prerna Ravi is a PhD student at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL). Her research focuses on designing generative AI agents that augment team collaboration in education, creative practice, and collective decision-making. She has studied how users' mental models of conversational AI evolve over time, and how they both shape and are shaped by perceptions of the agent's trustworthiness, agency, and anthropomorphism. Her research has been recognized with the Best Paper Honorable Mention and Nomination awards at ACM CHI and Learning at Scale. Prerna was the main organizer for a tutorial workshop at the International Society of Learning Sciences (ISLS) in 2023.

Laura Schütz is a PhD candidate in Computer Science at Technical University of Munich. She also holds an MS in Design from Stanford. Her research focuses on modeling human perception and cognition for adaptive user interfaces. Laura is an incoming Postdoc Fellow at the ETH AI Center, where she will focus on learning user mental models from multimodal interaction traces during human-AI interaction. She has previously organized multiple interdisciplinary workshops, e.g., at ISMAR, Harvard Kennedy School, and Stanford Graduate School of Business.

Anna Neumann is a PhD candidate in Computer Science at the Research Center for Trustworthy Data Science and Security. Her research focuses on transparency and governance requirements of technical systems to give stakeholders meaningful agency and recourse over systems, which has been recognized with a CHI Best Paper Award. She is part of the Compliant and Accountable Systems Group, which also focuses on perceptions of technological systems towards better governance of them. Anna has organized workshops, retreats, and mini-conferences for the RC Trust Graduate School which encompasses around 35 PhD students.

Robin Shing Moon Chan is a PhD candidate in Computer Science at ETH Zürich. His research focuses on understanding linguistic user behavior during LLM-assisted task solving and designing interactive algorithms and systems that effectively support such human–LM collaboration. His work has been recently recognized with a CHI Best Paper Award. He has previously co-organized a tutorial workshop at ACL attended by 100+ NLP researchers.

April Yi Wang is an Assistant Professor of Computer Science at ETH Zurich, where she leads the Programming, Education, and Computer Human Interaction Lab (PEACH) and is the chair of ACM SwissCHI local chapter. Through projects on AI tutors, classroom chatbots, learner-LLM problem decomposition, and agency in AI-mediated software engineering, she examines how educational tools can scaffold accurate mental model formation and repair rather than obscure it. She has served on the organizing committee of IEEE VL/HCC and regularly on the program committees of CHI, UIST, and Learning@Scale, as well as organized several workshops at CHI, IUI, and other venues.

Qiaosi (Chelsea) Wang is a Carnegie Bosch Postdoctoral Fellow at Carnegie Mellon University, and an incoming assistant professor at UC Berkeley's School of Education. Her work focuses on theorizing, designing, and examining

people's perceptions and mental models of conversational AI agents through the socio-cognitive lens of Mutual Theory of Mind across online learning, everyday life, and recently mental health contexts. Chelsea has served on the program committees of premier HCI conferences such as CHI, DIS, and CSCW. She's also the lead organizer of the Theory of Mind in Human-AI Interaction (ToMinHAI) workshop series at CHI and CUI.

Sumit Asthana is a Senior Applied Scientist at Microsoft. His work focuses on advancing AI assistants and agentic systems, with a focus on human-AI interaction, decision support, evaluation frameworks, and synthetic data generation for AI-assisted experiences. His research spans human-centered AI, Bayesian modeling, AI-assisted decision-making, and the alignment of AI systems with human mental models. He has served on committees related to Human-centered AI in HCI and ML conferences such as CHI, CSCW, ICLR, NeurIPS.