Mental Models of AI: Bridging Elicitation Methods and Stochastic Systems

BHADA YUN, ETH Zürich, Switzerland
LAURA SCHÜTZ, Technical University of Munich, Germany
TÉO SANCHEZ, LMU Munich, Germany
PRERNA RAVI, Massachusetts Institute of Technology, USA
QIAOSI WANG, Carnegie Mellon University, USA
ANNA NEUMANN, University of Duisburg-Essen, Germany
OLIVER HUANG, ETH Zürich, Switzerland
GRACE WANG, Stanford University, USA
NADINE BIENEFELD, ETH Zurich, Switzerland
HARI SUBRAMONYAM, Stanford University, USA
LEV TANKELEVITCH, Microsoft Research, UK
APRIL YI WANG, ETH Zürich, Switzerland

Human-AI interaction researchers elicit people's mental models of how AI systems work, using methods such as think-alouds and free drawing. Each method captures different aspects of a mental model, so conclusions about the same user can differ across methods. As AI systems become more varied, multi-purpose, and embedded in people's lives, it is increasingly unclear how researchers can faithfully capture mental models of them. This workshop examines what elicitation methods make observable, discusses how AI systems complicate existing approaches, and identifies shared research priorities. In a world café, participants work with two organizer-prepared card decks, one of elicitation methods and one of AI systems. Groups draw one of each, design a study using the method to elicit mental models of the system, critique another group's study, and revise their own accordingly. Revised decks and a co-authored position paper will help researchers choose and compare elicitation methods across AI systems.

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

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

ACM Reference Format:
Bhada Yun, Laura Schütz, Téo Sanchez, Prerna Ravi, Qiaosi Wang, Anna Neumann, Oliver Huang, Grace Wang, Nadine Bienefeld, Hari Subramonyam, Lev Tankelevitch, and April Yi Wang. 2027. Mental Models of AI: Bridging Elicitation Methods and Stochastic Systems. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '27), May 10–14, 2027, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 8 pages.

Authors' Contact Information: Bhada Yun, bhayun@ethz.ch, ETH Zürich, Zürich, Switzerland; Laura Schütz, laura.schuetz@tum.de, Technical University of Munich, Munich, Germany; Téo Sanchez, teo.sanchez@lmu.de, LMU Munich, Munich, Germany; Prerna Ravi, prernar@mit.edu, Massachusetts Institute of Technology, Cambridge, MA, USA; Qiaosi Wang, qiaosiw@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Anna Neumann, anna.neumann1@uni-due.de, Research Center Trustworthy Data Science and Security (RC-Trust), University of Duisburg-Essen, Duisburg, Germany; Oliver Huang, haohuang@ethz.ch, ETH Zürich, Zürich, Switzerland; Grace Wang, gracewng@stanford.edu, Stanford University, Stanford, CA, USA; Nadine Bienefeld, nadine.bienefeld@unilu.ch, ETH Zurich, Zürich, Switzerland; Hari Subramonyam, harihars@stanford.edu, Stanford University, Stanford, CA, USA; Lev Tankelevitch, lev.tankelevitch@microsoft.com, Microsoft Research, Cambridge, UK; April Yi Wang, april.wang@inf.ethz.ch, ETH Zürich, Zürich, Switzerland.

2027. Manuscript submitted to ACM

1 Motivation

In cognitive science, mental models are internal representations of external reality that people use to reason, make decisions, and anticipate events [7, 9, 11]. HCI borrowed the term to describe people's internal representations of a system [4, 16]. These representations are built from direct experience, explanations, and analogy, and shaped by prior knowledge and expectations [8, 16]. They let people explain and predict a system's behavior and guide how they perceive and interact with it [16, 17, 22]. Well-formed mental models can also help people learn and retain a device's procedures, execute them faster, and infer procedures they were never taught [13], as well as get an intelligent agent to work to their satisfaction [14]. For HCI, understanding how people conceptualize computing systems is therefore central to designing systems they can learn to use effectively.

Because mental models are constructed and updated in people's minds, studying them requires a way to externalize them. Researchers therefore *elicit* mental models through methods such as think-alouds [15] and free drawing [10], which produce verbal or visual traces of these models that others can examine [2, 12]. Each of these methods brings particular aspects of a mental model into view and carries assumptions about what counts as evidence or an instantiation of one [20]. Depending on those assumptions, researchers may conclude that the same user's mental model is well-formed under one method but not under another [1, 6]. Comparing findings therefore requires knowing what each method can show, and whether differences between methods provide complementary evidence or lead to conflicting conclusions. This leads to our first question: *What does each elicitation method make observable, and how do its assumptions shape the claims a study can make?*

Importantly, generative and agentic AI systems produce variable outputs, adapt to context, and act with increasing agency, so what they can do extends well beyond what their labels or initial conceptualizations suggest [5, 23, 28, 29]. These systems also change quickly, as do the ways people use them and what they project onto them. This has direct effects on study design and interpretation, as users' mental models can shift as the novelty effect wears off or users get familiarized with the researcher's intended representation of a system over time [18]. This leads to our second question: *How do generative and agentic AI systems challenge the assumptions that mental model research inherited from studying more predictable systems?*

These two questions meet in study design. No study can be run without committing, even implicitly, to an elicitation method and a conceptualization of a specific AI system. The workshop therefore aims to make methods and conceptualizations of AI more explicit among researchers and pursues three goals:

G.1 **Examine What Elicitation Methods Make Observable.** Compare what methods capture and what they assume a mental model contains, and distinguish when differences across methods provide complementary evidence and when they lead to conflicting conclusions.

G.2 **Assess Study Designs Across AI Systems.** Use combinations of methods and AI systems to examine where existing approaches remain suitable for generative and agentic AI, where adaptation is needed, and how researchers' own conceptualizations of a system shape what a study of it can conclude.

G.3 **Identify Common Ground and Research Priorities.** Bring together perspectives across disciplines and between research and practice to identify which assumptions about mental models remain useful, where revision may be needed, and which conceptual and methodological questions warrant further research.

To pursue these goals, participants work in small groups, each designing a study sketch from two card decks: one of elicitation methods and one of AI systems. In a world café format, each group first specifies its study together with a *reference model* of what they believe constitutes a well-formed mental model of their chosen AI system, making

their own conceptualization of the system explicit (G.2). Other groups then critique the design, asking whether the method yields evidence specific to the chosen system (G.1), before the group returns to revise it based on the critiques. Comparing designs across tables surfaces which assumptions participants share and where they disagree, which we collect as research priorities (G.3).

This workshop continues a series that begins with our accepted half-day Mental Models of AI (MM/AI) workshop at ACM IUI '27 (Helsinki, February 2027), where participants will elicit one another's mental models of a coding agent and a text-to-image generator. Between IUI and CHI, organizers will build the two decks from a systematic review of 88 empirical studies [21] and from what the IUI exercises reveal about where methods and systems are hard to separate. At CHI, participants will use the card decks to brainstorm possible human-AI interaction (HAI) study designs, and critique and extend cards where definitions or method descriptions fall short. Afterward, we release the revised decks as an open resource for planning, reporting, and teaching mental model research, giving researchers from different traditions a shared reference point in the spirit of the CHI 2027 theme, *Bridges*.

2 Organizers

The team brings expertise in HAI, cognitive science, governance, and applied AI across institutions in Germany, Switzerland, the United Kingdom, and the United States, and together covers the workshop's elicitation methods and AI systems. Several have organized workshops at CHI, IUI, and CUI. Yun is the lead organizer and coordinates the position paper; Yun, Schütz, Sanchez, and Ravi design the card decks; organizers review submissions with the external program committee; and at least seven organizers will attend in person to host tables and facilitate discussion.

Bhada Yun is a PhD student at ETH Zürich. 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 well-being and agency. His research has been recognized with four CHI Honorable Mention Awards.

Laura Schütz recently completed her PhD in Computer Science at the Technical University of Munich. Her research focuses on modeling human perception and cognition for adaptive user interfaces. Laura is an incoming Postdoctoral 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.

Téo Sanchez is an MSCA Postdoctoral Fellow at the Munich Interactive Intelligence Initiative (MI3), LMU Munich. He recently conducted a systematic review of 88 empirical studies on mental models in human-AI interaction [21], which informed the framing of this workshop. His work has received awards at IUI 2022 (best paper), C&C 2023 (honorable mention), and the 2023 AFIHM Best PhD Dissertation Award.

Prerna Ravi is a PhD student at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL). 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 the Learning Sciences (ISLS) in 2023.

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. She is also the lead organizer of the Theory of Mind in Human-AI Interaction (ToMinHAI) workshop series at CHI and CUI [26, 27].

Anna Neumann is a PhD candidate in Computer Science at the Research Center Trustworthy Data Science and Security (RC-Trust), University of Duisburg-Essen. 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. Anna has organized workshops, retreats, and mini-conferences for the center's graduate school of around 35 PhD students.

Oliver Huang is a PhD student at ETH Zürich studying how novice users form and calibrate mental models of AI. He designs interactive techniques that engage users in externalizing and reflecting on their mental models, with work appearing at CHI, IUI, and VIS. His research spans the learning and visualization communities, where mental models are operationalized and evaluated in markedly different ways.

Grace Wang is a PhD student at Stanford University. She studies AI's long-term impacts on individual well-being and designs AI systems that foster well-being, growth, and authentic human experiences, publishing at venues like CHI, UIST, and TOCE. Her work often develops methods to elicit mental models, such as learners' mental models of self-regulation skills.

Nadine Bienefeld is a Senior Lecturer at ETH Zürich and a Senior Researcher at the University of Lucerne. She studies how AI transforms expert judgment, trust, and teamwork in high-risk settings, with a particular focus on shared mental models and transactive memory systems in human-AI teams. She brings a background in human factors of aviation safety and was a visiting fellow at NASA Ames.

Hari Subramonyam is an Assistant Professor (Research) at Stanford Graduate School of Education and Computer Science (by courtesy). He is also the Ram and Vijay Sriram Faculty Fellow at the Stanford Institute for Human-Centered AI. He studies ways to augment human learning using AI by engaging in cognitively informed design practices.

Lev Tankelevitch is a researcher at Microsoft Research within the *Tools for Thought* group [19]. His research uses metacognition as a lens to augment intentionality and human-AI interaction in collaborative knowledge work. He has a background in behavioural science and in cognitive neuroscience. He co-organises the *Tools for Thought* workshops at CHI [24, 25, 30].

April Yi Wang is an Assistant Professor of Computer Science at ETH Zürich, where she leads the Programming, Education, and Computer Human Interaction Lab (PEACH) and chairs the ACM SwissCHI local chapter. She examines how educational tools can scaffold accurate mental model formation and repair rather than obscure it. She has organized several workshops at CHI, IUI, and other venues.

3 Activities

3.1 Before the Workshop

We will launch the website by December 17, 2026, and distribute the call via SIGCHI and disciplinary mailing lists and organizer networks. We invite 2- to 8-page non-archival papers on mental models of AI or adjacent constructs, such as perceptions, sensemaking, and schemata, including reanalyses of existing data. Submissions are due February 11, 2027, and notifications will be sent by February 25, 2027, at least seven days before CHI's early-registration deadline. Each submission receives at least two double-blind reviews on OpenReview from organizers and external program committee members (provisionally Pat Pataranutaporn, MIT Media Lab; Magdalena Wischnewski, University of Duisburg-Essen; Mennatallah El-Assady, ETH Zürich), who judge relevance, methodological or theoretical insight, and potential to inform discussion of study design. Beyond selecting participants, submissions feed into the activities. Each names the AI systems it addresses and any elicitation methods it uses, and we add those missing from the decks. Four accepted

papers, chosen to span methods and systems, become five-minute lightning talks. Attendees without a submission can still take part fully, though authors of accepted papers receive priority.

The two printed, color-coded decks are the core material of the workshop. A *method card* names an elicitation method, the trace it produces (e.g., a sketch, a transcript, a set of predictions), what it assumes a mental model contains, and key references. An *AI-system card* describes a concrete usage scenario: the system, who uses it and for what, and one property that makes it hard to study (e.g., its output varies between runs, it remembers across sessions, or it acts with high proactivity). For example, the method card *free drawing* asks participants to sketch how a system works and assumes that its structure can be depicted spatially, while the AI-system card *AI companion used daily for emotional support* notes that the companion appears to remember past conversations, although what it actually retains is opaque to users. Accepted papers, accessible digital versions of the cards, and instructions will be shared with participants at least one week before the workshop.

3.2 Workshop Schedule

Session 1: Framing and research question formulation. On arrival, participants are assigned to five to six tables of about six people that mix disciplines, each hosted by an organizer. An introduction to mental models in HAI research is followed by two 10-minute invited talks by Dr. Zahra Ashktorab at Microsoft Research and a second invited speaker from industry (confirmation pending). Afterwards, we invite four 5-minute lightning talks from four selected papers with 2 minutes for questions. Organizers then walk each table through a worked example of the activity, described as follows: Each group draws one method card and one AI-system card. Groups can swap a card if they are unfamiliar with the method or system. The group then writes a single research question that the drawn method could plausibly answer about the drawn system. For instance, a group holding *free drawing* and *AI companion* might ask: *What do daily users of an AI companion believe it remembers about them, as expressed in free drawings?* The random card draw pushes groups toward pairings they would not choose themselves, while the swap keeps each design within reach of the group's expertise. This research question is the only output of Session 1 and is the starting point for the group's study sketch in Session 2.

Session 2: World café and study design. After the break, we begin the hands-on world café group activity. A world café structures discussion through three successive small-group rounds in which participants move between tables and build on earlier conversations [3]. Organizers stay at their table and groups move to different tables. Group members do not need to reach a consensus during any of the rounds, and ad-hoc revisions to the decks themselves, such as missing definitions or methods, are also encouraged as interesting findings and noted down by an organizer. In Round 1, each group designs a study around its research question, asking who the participants are, what they do under what method, what the method produces, what results they expect, and how they should interpret them. Toward the end of Round 1, each group also drafts a *reference model*: what a well-formed mental model of the system would contain. In Round 2, each group (excluding the organizers) switches to another table and critiques the study design. Groups are given prepared prompts by the organizer to start the discussions, e.g., *What can participants express through this method, and what not? Given the reference model, could the method tell a well-formed mental model from a poorly formed one? Would the findings hold if the system behaved differently next session?* Once the critique round is over, participants move on to Round 3, where they return to their home tables and revise the sketch in light of the critique. The discussion then turns to the larger group, where each group shares their initial research design, and how the critiques led to the revised study design. Finally, the organizer at each table facilitates a table reflection on how the other groups they observed

Table 1. Workshop schedule: two consecutive 90-minute sessions with a break. Times are elapsed within each session and include moves between tables.

Session 1: Framing and research question formulation		*Session 2: World café study design*	
0:00–0:05	Welcome and table assignment	0:00–0:20	Round 1: design a study and its reference model
0:05–0:15	Introduction to mental models in HAI	0:20–0:40	Round 2: critique another table's study design
0:15–0:35	Two invited talks	0:40–1:00	Round 3: revise a study, propose card changes
0:35–1:05	Four lightning talks	1:00–1:15	Table reports
1:05–1:15	Worked example with the decks	1:15–1:25	Group discussion of research priorities
1:15–1:30	Card draw and research question design	1:25–1:30	Closing and next steps

formulated, critiqued, and revised their study designs, and helps the group agree on its top two research priorities, which organizers collect for the position paper. The lead organizer closes the workshop activities by discussing next steps towards a co-authored position paper. For those interested, an interest form will be distributed through the mailing list and Slack.

3.3 After the Workshop

With contributors' permission, accepted papers and selected study sketches will be shared through the website and a public Zenodo community; participant papers will be non-archival. The revised decks, prompts, and a workshop summary will be freely downloadable from the workshop website. Participants will be invited to co-author a position paper synthesizing the study sketches and research priorities, preliminarily targeted at CHI '28 or ACM *Interactions*, with contributors credited as co-authors or, with permission, in the acknowledgments. The workshop does not involve research with human participants: the paper reports the group's collective output and does not analyze participants as research subjects.

4 Call for Participation

We invite researchers and practitioners from HCI, AI, psychology, cognitive science, design, and industry to *Mental Models of AI*. We ask: *How do elicitation methods and the AI systems they target shape what a study can conclude?* This in-person workshop combines invited talks, lightning talks, and a group activity in the form of three world café rounds across two 90-minute sessions. Using prepared elicitation method and AI-system cards, groups will design studies, imagine possible findings, and examine limitations and alternative interpretations. Discussions will address which claims study designs support, where interpretations diverge, and which conceptual and methodological questions should guide future mental model research. The cards provide referenced definitions and method descriptions; no prior experience with elicitation is required. We welcome empirical studies, reanalyses of existing data, methods, theory, reflexive accounts, and provocations on mental models of AI or adjacent constructs such as perceptions, sensemaking, and schemata. Papers should be 2 to 8 pages, excluding references, in ACM single-column format, and name the AI systems and any elicitation methods involved. Submit anonymized PDFs via OpenReview by February 11, 2027. Accepted papers will be shared with participants beforehand and, with authors' permission, posted non-archivally on Zenodo. Outputs include revised decks and a co-authored position paper. At least one author per accepted paper must register for CHI and attend in person, as will at least seven organizers. We expect 30–40 participants. Accepted-paper authors receive priority, though all attendees are welcome. Details and accessible materials: https://mentalmodelsofai.com/chi2027.

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] Juanita Brown, David Isaacs, and The World Café Community. 2005. *The World Café: Shaping Our Futures Through Conversations That Matter.* Berrett-Koehler Publishers.

[4] John M. Carroll and Judith Reitman Olson. 1988. Mental Models in Human-Computer Interaction. In *Handbook of Human-Computer Interaction*, Martin Helander (Ed.). North-Holland, Amsterdam, 45–65.

[5] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, New York, NY, USA. doi:10.1145/3593013.3594033

[6] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. 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, 307–317. doi:10.1145/3397481.3450644

[7] Kenneth J. W. Craik. 1943. *The Nature of Explanation.* Cambridge University Press, Cambridge, UK.

[8] Dedre Gentner and Donald R. Gentner. 1983. Flowing Waters or Teeming Crowds: Mental Models of Electricity. In *Mental Models*, Dedre Gentner and Albert L. Stevens (Eds.). Lawrence Erlbaum Associates, Hillsdale, NJ, 99–129. doi:10.4324/9781315802725-7

[9] Dedre Gentner and Albert L. Stevens (Eds.). 1983. *Mental Models.* Lawrence Erlbaum Associates, Hillsdale, NJ.

[10] 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

[11] Philip N. Johnson-Laird. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Harvard University Press, Cambridge, MA.

[12] 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), 46. doi:10.5751/ES-03802-160146

[13] 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:10.1207/s15516709cog0803_3

[14] 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* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/2207676.2207678

[15] Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. 2020. Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 183–191. doi:10.1145/3340631.3394841

[16] Donald A. Norman. 1983. Some Observations on Mental Models. In *Mental Models*, Dedre Gentner and Albert L. Stevens (Eds.). Lawrence Erlbaum Associates, Hillsdale, NJ, 7–14.

[17] Stephen J. Payne, Helen R. Squibb, and Andrew Howes. 1990. The Nature of Device Models: The Yoked State Space Hypothesis and Some Experiments with Text Editors. *Human-Computer Interaction* 5, 4 (1990), 415–444. doi:10.1207/s15327051hci0504_3

[18] Prerna Ravi, Carúmey Stevens, Beatriz Flamia Azevedo, Jasmine David, Brandon Hanks, Hal Abelson, Grace Lin, and Emma Anderson. 2027. Exploring Teachers' Perspectives on Using Conversational AI Agents for Group Collaboration. In *Artificial Intelligence in Education (AIED 2026) (Lecture Notes in Computer Science, Vol. 16584)*, Emmanuel G. Blanchard, Guanliang Chen, Min Chi, and Seiji Isotani (Eds.). Springer Nature Switzerland, Cham, 621–630. doi:10.1007/978-3-032-29763-1_43

[19] Microsoft Research. 2024. Tools for Thought. https://www.microsoft.com/en-us/research/project/t4t/. Accessed: 2024-09-26.

[20] William B. Rouse and Nancy M. Morris. 1986. On Looking into the Black Box: Prospects and Limits in the Search for Mental Models. *Psychological Bulletin* 100, 3 (1986), 349–363. doi:10.1037/0033-2909.100.3.349

[21] 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* (Paphos, Cyprus) *(IUI '26)*. Association for Computing Machinery, New York, NY, USA, 663–682. doi:10.1145/3742413.3789223

[22] Nancy Staggers and A.F. Norcio. 1993. Mental models: concepts for human-computer interaction research. *International Journal of Man-Machine Studies* 38, 4 (1993), 587–605. doi:10.1006/imms.1993.1028

[23] Hariharan Subramonyam, Roy Pea, Christopher Lawrence Pondoc, Maneesh Agrawala, and Colleen M. Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*

(CHI '24). ACM, New York, NY, USA, Article 1039, 19 pages. doi:10.1145/3613904.3642754

[24] Lev Tankelevitch, Elena L Glassman, Jessica He, Majeed Kazemitabaar, Aniket Kittur, Mina Lee, Srishti Palani, Advait Sarkar, Gonzalo Ramos, Yvonne Rogers, et al. 2025. Tools for Thought: Research and Design for Understanding, Protecting, and Augmenting Human Cognition with Generative AI. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[25] Lev Tankelevitch, Elena L Glassman, Jessica He, Aniket Kittur, Mina Lee, Srishti Palani, Advait Sarkar, Gonzalo Ramos, Yvonne Rogers, and Hari Subramonyam. 2025. Understanding, Protecting, and Augmenting Human Cognition with Generative AI: A Synthesis of the CHI 2025 Tools for Thought Workshop. *arXiv preprint arXiv:2508.21036* (2025).

[26] Qiaosi Wang, Sarah Walsh, Mei Si, Jeffrey Kephart, Justin D. Weisz, and Ashok K. Goel. 2024. Theory of Mind in Human-AI Interaction. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. ACM, New York, NY, USA, Article 493, 6 pages.

[27] 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)*. ACM, New York, NY, USA, Article 33, 6 pages.

[28] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, USA, Article 378, 22 pages. doi:10.1145/3613904.3642466

[29] Bhada Yun, Evgenia Taranova, and April Yi Wang. 2026. Does My Chatbot Have an Agenda? Understanding Human and AI Agency in Human-Human-like Chatbot Interaction. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 373, 32 pages. doi:10.1145/3772318.3791620

[30] Zelun Tony Zhang, Nick Von Felten, Leon Reicherts, Lev Tankelevitch, Zhitong Guan, Sean Rintel, Yue Fu, Jessica He, Kenneth Holstein, Advait Sarkar, et al. 2026. Tools for Thought: Understanding, Protecting, and Augmenting Human Cognition with Generative AI—From Vision to Implementation. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–7.