Design for Calibrated Trust in AI: Exploring Opportunities to Support An Appropriate Mental Models When Interacting With Conversational AI

Sarah Diefenbach*

LMU Munich, Department of Psychology, sarah.diefenbach@psy.lmu.de

Daniel Ullrich

LMU Munich, Department of Computer Science, daniel.ullrich@ifi.lmu.de

Paul Miles Preuschoff

RWTH Aachen University, Department of Computer Science, preuschoff@cs.rwth-aachen.de

Lara Christoforakos

LMU Munich, Department of Psychology, lara.christoforakos@psy.lmu.de

Artificial intelligence (AI) is increasingly embedded in daily life, offering convenient support in many tasks like inspiration for text production or answering everyday questions. However, its risks often remain less visible, ranging from data security concerns to more subtle effects like growing dependence, negative impacts on creativity and reduced critical thinking. While responsible AI use is widely advocated, we argue that this responsibility cannot be achieved through education alone but must be supported through AI design. Specifically, AI systems should foster appropriate mental models in users. This workshop explores how design choices, such as reducing human-like AI design or communicating (un)certainty of responses, can support more informed and balanced use. Inviting perspectives from disciplines including computer science, psychology, design, and ethics, the workshop aims to (1) deepen understanding of failures caused by misleading mental models, (2) generate initial design strategies to address them, and (3) build a research community for future collaboration in studies, funding, initiatives, and publications.

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

Additional Keywords and Phrases: AI, responsible use, trust, mental models, interaction design, psychological mechanisms.

ACM Reference Format:

First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 MOTIVATION AND OBJECTIVES

Artificial Intelligence (AI) has become part of our daily routines and conversation with chatbots based on Large Language Models (LLMs) such as ChatGPT has become one of the major forms of AI interaction in everyday life. Tasks like writing a formal letter, getting inspiration, or preparing a speech can be supported or delegated to AI. People address ChatGPT for various issues, such as health and medical diagnoses, advice on relationship issues and parenting, learning, news/politics, and many other questions they would not dare to ask anybody else [2]. Many people are delighted with how much easier AI has made their lives. But research also highlights the risks and side-effects of AI in many domains (e.g., education, journalism, governance) [6], [10], [12], [13], [14], [17], [18], [20], warning that people must use it "responsibly" and make "conscious choices" [1], [5], [8], [15], [18], [19].

1.1 Human-like AI design activates misleading mental models

However, we question that people know enough about AI to make such conscious choices and that current AI design supports the insights required for "responsible use". Responsible use requires an adequate mental model of the objective. To drive responsibly, one must have an idea of speed and braking distance and understand what kind of context factors (e.g., weather conditions, road conditions, sleepiness) affect the driver's control of the car. This can hardly be assumed for the average AI user. As argued by Ben Shneiderman [16], current AI design contradicts the traditional golden rules of good interface design, like making technology comprehensible, predictable and controllable. In fact, even the metaphor of a conversation applied in many AI contexts (e.g., ChatGPT) could be deceptive and trigger false assumptions (also see [3], [16], [9]). The reason this metaphor is quite established, is probably rooted in the fact that interacting with LLMs (on the surface) resembles what we know from human conversations. As known from research on social cognition in HCI (e.g., [11]), social patterns from human interaction can easily be activated in HCI, although individuals are aware of interacting with non-human entities. However, with LLMs that act very similarly to humans (e.g., through human language use), it can even be difficult to differentiate, whether one is chatting with a bot or a human. Still there are important differences ([3], also see Table 1).

Table 1: Differences between human interaction and interaction with conversational AI

When speaking to a human… 	**When interacting with AI…** 
Eloquent language hints at profound knowledge	Eloquent language results from clustering words occurring together with high probability, but not necessarily indicates meaningful content
One usually gets hints of how certain the person is about a statement	The answers include no reliable cues about certainty or suggests certainty which actually does not exist. One cannot infer whether statements are based on solid data, just single pieces of information on the web, or even pure hallucination
The human typically has a position itself	It tends to confirm you
You might not have the same opinion, get into heated debates or even stop talking to each other for some time	It is built to make you engage more, and prefer it over other services
A real friend would tell you the truth even if its uncomfortable	It always offers you a follow-up suggestion Reporting problematic behavior, no harsh judgment is imposed

1.2 Doublebind communication and responsibility paradox

Still, talking about responsible handling of AI, users are expected to constantly keep in mind that AI is only a machine, and that polished language is no indicator of quality. On the one side, the humanlike design of AI does a lot to build up

trust (and overtrust) [4]. On the other hand, disclaimers and regulations of AI use suggest full responsibility on the human side and double-checking AI results. This contradiction may also result from financial interests. Service providers create systems that are designed for financial gain, thus projecting the image of a competent, helpful system, while offloading responsibility for failures to users. Like with deceptive patterns [7], the current designs of AI do not align with the interests of users. In sum, the current framing and design of AI represents what psychotherapists would call a doublebind communication, including two contradictory messages: In the moment of interaction, the message is, *"just consider it as a human interaction partner. Speak to it like you would do to a human. See human in it."* In the moment of taking responsibility, the message is, *"remember it is only a machine that can make mistakes. You are responsible to doublecheck and reflect on everything*." From our perspective, we doubt whether a calibrated mental model of AI is something that can and should be addressed by simple calls for "AI literacy" in the sense of education or disclaimers. Instead **we suggest, that the design itself needs to speak to the user**. Just as you see in other areas (e.g., speed reduction in a street near a kindergarten), warning signs ("please remember, drive slowly") don't show the same effect as traffic bollards. While the concept of a (less) human-like metaphors in interaction design could be one starting point, there may be more. For example, less deceptive AI design could provide a chance to estimate the quality of statements and the data behind, e.g., ratings or visualizations that differentiate certain from not so certain statements. It also could encourage critical reflection, instead of confirming what feels good to hear.

1.3 Workshop objectives

Our workshop addresses the question of **how we can design AI in a way that it activates an appropriate mental model, namely, an idea of what AI is, its potentials, limits, motivations – so that people can use it responsibly, and not trap into biases, lacking trust or overtrust**. Given that AI interaction is introduced to various areas of daily life, we see this as a topic of high significance, with regards to individual and societal wellbeing and informed use, and thus, an important task for HCI research and design. We invite contributions from various disciplines and backgrounds such as computer science, psychology, design, ethics, and beyond. Main contributions of the workshop are (1) a broader understanding of AI failures due to misleading mental models, (2) first ideas how to address such difficulties in design, and (3) establishing a research community for next joint activities, e.g., joint studies, funding, initiatives, and publications.

2 WORKSHOP ACTIVITIES AND PROMOTION

The workshop is planned as half-day (3.5h incl. one 30-min coffee break) interactive venue with the focus on discussion and collaborative creation. A rough sketch of the workshop timeline is displayed below (see Table 2). On-site requirements are met by the standard equipment of workshop rooms (i.e., moveable tables, chairs, screen and projection). We will bring additional materials for the prototyping and interactive brainstorming sessions. Our promotional strategy to recruit participants is to actively invite researchers from our network (e.g., participants of related workshops at recent conferences such as CHI 2026) and to spread the call for participation via relevant groups in social media and mailings lists (for a draft see https://docs.google.com/document/d/1B6lMaJl7oI6G7sbu8ryXCGMUdFqNzm5BQxFKPg5snXg/edit?usp=sharing). Post-workshop plans include to publish workshop position papers and establishing a special issue and/or joint position paper with interested participants (e.g., special issue in the i-com – Journal of Interactive Media, where some of the workshop organizers sit on the editorial board). Moreover, we plan to use the NordiCHI 26 workshop as a kick-off for further workshops at other ACM conferences or seminars at other international venues (e.g. applying for a Dagstuhl seminar). In addition, we will discuss interests to apply for a larger funding to address the workshop issue more systematically over a larger time horizon.

Table 2: Rough sketch of the workshop schedule

Duration	Activity
15min	*Introduction*. Short presentation of the organizers, workshop outlook, intro of relevant basic theoretical/psychological concepts (i.e., mental models, trust, overtrust, calibrated trust, responsible use). Examples of failures/non responsible use due to inappropriate mental models in other domains
30min	*Common ground*, overview of perspectives: 90-sec pitches of participants
15min	*Activation:* Collecting examples of failures of AI use: individual work, post-its, then clustering joint discussion
30min	*Exploration:* Misleading mental models of AI: group work, 3 groups with 5-7 persons. Each group addresses specific clusters of failures, identify how these might be traced back to inaccurate mental models of AI
30min	Coffee break
30min	*Creativity*: Interactive part on design opportunities: group work, 3 groups with 5-7 persons. Each group addresses specific mental model inadequacies, identify how a more accurate mental model could be stimulated by design Rapid prototyping (creative process, from visual prototypes to role plays)
15min	*Presentation*: Presentation and discussion: 5 minute presentation from each group
15min	*Wrap-up*: most promising ideas, syntheses, take-aways from the workshop
15min	*Community building* and outlook: discussion next research steps, possible collaborations, joint publications, applying for joint funding, future workshops

3 ORGANIZERS

The organizing committee consists of researchers from various disciplines such as psychology, communication studies, computer science and interaction design, with previous experience in organizing workshops at national and international various conferences.

Sarah Diefenbach is professor for economic psychology and human-computer interaction at LMU Munich and addresses user experience and the design of interactive products from a psychological perspective. Her research has continuously stimulated discussions in the (Nordi)CHI community. Sarah will moderate and coordinate the workshop and will particularly contribute with her expertise in HCI design and evaluation from a psychological needs perspective.

Daniel Ullrich is postdoctoral researcher at the Chair of Human-Computer Interaction at LMU Munich. His research focuses on AI and Human-Robot-Interaction (HRI) and the critical analysis of societal impacts of technology. Daniel developed several methods for UX design and evaluation (e.g., INTUI questionnaire. robot impression inventory) the Cassandra method for responsible design based on dystopian visions which could also be applied in parts of the workshop.

Paul Preuschoff is a doctoral candidate in the field of Human-Computer Interaction at RWTH Aachen University. His research focuses on human centered design, social experiences, deceptive patterns and the impact of AI access on creativity, enjoyment and social interaction. He currently investigates different methods to incorporate AI indirectly into creative processes in order to expand them, while limiting AIs negative effects

Lara Christoforakos is a postdoctoral researcher at the chair of economic psychology and human-computer interaction at LMU Munich. Her research focuses on technologies for behavior change. She takes a critical perspective on emerging technologies, particularly highly agentic AI systems that may shift responsibility away from users and contributes to the workshop by raising critical questions about these developments and implications for user autonomy and design.

REFERENCES

[1] Glen Berman, Nitesh Goyal, and Michael Madaio. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, May 11, 2024. ACM, Honolulu HI USA, 1–24. https://doi.org/10.1145/3613904.3642398

[2] Sarah Diefenbach, Lara Christoforakos, Daniel Ullrich, and Marie Veihelmann. 2026. Sustainable AI: Exploring Gains and Losses of AI in Daily Routines. AHFE 2026 International Conference, July 20-24, Istanbul, Turkey.

[3] Sarah Diefenbach, and Daniel Ullrich (2026). Why We Need to Destroy the Illusion of Speaking to A Human: Critical Reflections On Ethics at the Front-End for LLMs. In Proceedings of the CHI 2026 Workshop: Ethics at the Front-End: Responsible User-Facing Design for AI Systems (CHI '26), April 13--17, 2026, Barcelona, Spain. http://arxiv.org/abs/2603.16633

[4] Sarah Diefenbach, Daniel Ullrich, and Marie Veihelmann. (2026). Biases Behind Overtrust In AI: Critical Reflections On An Appropriate Mental Model Of LLMs. In Proceedings of the CHI 2026 Workshop on Understanding, Mitigating, and Leveraging Cognitive Biases to Calibrate Trust in Evolving AI Systems (CHI '26), April 13--17, 2026, Barcelona, Spain. https://chi-bias-trust.github.io/participants.html

[5] Upol Ehsan, Elizabeth A Watkins, Philipp Wintersberger, Carina Manger, Nina Hubig, Saiph Savage, Justin D. Weisz, and Andreas Riener. 2025. New Frontiers of Human-centered Explainable AI (HCXAI): Participatory Civic AI, Benchmarking LLMs, XAI Hallucinations, and Responsible AI Audits. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, April 26, 2025. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3706599.3706713

[6] K. J. Kevin Feng, Rock Yuren Pang, Tzu-Sheng Kuo, Amy Winecoff, Emily Tseng, David Gray Widder, Harini Suresh, Katharina Reinecke, and Amy X. Zhang. 2025. Sociotechnical AI Governance: Challenges and Opportunities for HCI. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, April 26, 2025. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3706599.3706747

[7] Colin M. Gray, M., Cristiana Teixeira Santos, Nicole Tong, Thomas Mildner, Arianna Rossi, Johanna T. Gunawan, Caroline Sinders, 2023. Dark patterns and the emerging threats of deceptive design practices. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-4). https://dl.acm.org/doi/full/10.1145/3544549.3583173

[8] Ilka Hein and Sarah Diefenbach, 2025. Towards a comprehensive view on technology transparency: A cross-technology investigation of psychological and design factors around users' transparency need and perception. International Journal of Human–Computer Interaction, 1-18. https://doi.org/10.1080/10447318.2025.2502983

[9] Kashmir Hill, 2025. Why Do A.I. Chatbots Use 'I'? The New York Times, 19/12/2025. https://www.nytimes.com/2025/12/19/technology/why-do-ai-chatbots-use-i.html (accessed 25/04/2026)

[10] Joel Kiskola, Henrik Rydenfelt, Thomas Olsson, Lauri Haapanen, Noora Vänttinen, Matti Nelimarkka, Minna Vigren, Salla-Maaria Laaksonen, and Tuukka Lehtiniemi. 2025. Generative AI and News Consumption: Design Fictions and Critical Analysis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, April 26, 2025. ACM, Yokohama Japan, 1–18. https://doi.org/10.1145/3706598.3713804

[11] Youngme Moon. 2003. Don't Blame the Computer: When Self-Disclosure Moderates the Self-Serving Bias. *Journal of Consumer Psychology* 13, 1, 125–137. https://doi.org/10.1207/153276603768344843

[12] Sachita Nishal, Marianne Aubin Le Quéré, Brian James McInnis, Bronwyn Jones, Kristen Vaccaro, Tanja Aitamurto, Mor Naaman, and Nicholas Diakopoulos. 2025. News Futures: (Re-)Designing Socio-technical Systems for News Production and Consumption. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, April 26, 2025. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3706599.3706733

[13] Paul M. Preuschoff, Raghad Zaghal, Oliver Nowak, and Jan Borchers, 2025. Groups and AI: How Access to Generative AI Influences Creative Teamwork. In *Proceedings of the Mensch und Computer 2025* (pp. 407-421). https://dl.acm.org/doi/10.1145/3743049.3743073

[14] Gisela Reyes-Cruz, Velvet Spors, Michael Muller, Marianela Ciolfi Felice, Shaowen Bardzell, Rua Mae Williams, Karin Hansson, and Ivana Feldfeber. 2025. Resisting AI Solutionism: Where Do We Go From Here? In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, April 26, 2025. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3706599.3706732

[15] Malak Sadek, Marios Constantinides, Daniele Quercia, and Celine Mougenot. 2024. Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, May 11, 2024. ACM, Honolulu HI USA, 1–20. https://doi.org/10.1145/3613904.3642810

[16] Ben Shneiderman, 2022. Human-centered AI. Oxford University Press.

[17] Lev Tankelevitch, Elena L. Glassman, Jessica He, Majeed Kazemitabaar, Aniket Kittur, Mina Lee, Srishti Palani, Advait Sarkar, Gonzalo Ramos, Yvonne Rogers, and Hari Subramonyam. 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*, April 26, 2025. ACM, Yokohama Japan, 1–8. https://doi.org/10.1145/3706599.3706745

[18] Niels Van Quaquebeke, and Fabiola H. Gerpott, 2023. The Now, New, and Next of Digital Leadership: How Artificial Intelligence (AI) Will Take Over and Change Leadership as We Know It. Journal of Leadership & Organizational Studies, 30(3), 265-275. https://doi.org/10.1177/15480518231181731

[19] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 2023. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3581278

[20] Shaomei Wu, Kimi Wenzel, Jingjin Li, Qisheng Li, Alisha Pradhan, Raja Kushalnagar, Colin Lea, Allison Koenecke, Christian Vogler, Mark Hasegawa-Johnson, Norman Makoto Su, and Nan Bernstein Ratner. 2025. Speech AI for All: Promoting Accessibility, Fairness, Inclusivity, and Equity. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, April 26, 2025. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3706599.3706746