Understanding, Mitigating, and Leveraging Cognitive Biases to Calibrate Trust in Evolving AI Systems

Saumya Pareek
The University of Melbourne
Parkville, VIC, Australia

Nattapat Boonprakong
National University of Singapore
Singapore

Naja Kathrine Kollerup
Aalborg University
Aalborg, Denmark

Si Chen
University of Notre Dame
Notre Dame, IN, USA

Simo Hosio
University of Oulu
Oulu, Finland

Koji Yatani
University of Tokyo
Tokyo, Japan

Yi-Chieh Lee
National University of Singapore
Singapore

Ujwal Gadiraju
Delft University of Technology
Delft, Netherlands

Niels van Berkel
Aalborg University
Aalborg, Denmark

Jorge Goncalves
The University of Melbourne
Parkville, VIC, Australia

Abstract

Despite decades of advancements in Artificial Intelligence (AI), fostering appropriate trust in AI systems remains a challenge. Cognitive biases — systematic deviations from rational judgement — profoundly influence human decision-making, and reliance on such "mental shortcuts" can make AI systems appear more or less trustworthy than they really are, often undermining collaboration outcomes. As AI evolves with more sophisticated and persuasive natural language outputs, particularly through Generative AI (GenAI) and Large Language Models (LLMs), these biases may manifest in new and unpredictable ways, calling for their comprehensive examination. This workshop brings together diverse researchers from HCI, human-centred AI, cognitive psychology, interaction design, and related fields to collaboratively explore how cognitive biases influence trust calibration in human-AI interaction and establish a research agenda. We will explore how biases emerge across the human-AI interaction pipeline, what design strategies can mitigate or even harness these heuristics, and what methods are needed to study these dynamics effectively. Through a highly interactive 90-minute session, participants will map out open challenges, brainstorm tensions and solutions, chart future research directions, and share perspectives from their own diverse disciplinary lenses. Through this workshop, we aim to build a shared understanding of how cognitive biases influence trust in evolving AI systems, and derive a forward-looking, bias-aware research agenda that promotes appropriate trust in human–AI interaction.

CCS Concepts

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

Keywords

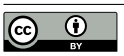
cognitive biases, heuristics, trust, trust calibration, human-AI interaction, cognitive limitations

ACM Reference Format:
Saumya Pareek, Nattapat Boonprakong, Naja Kathrine Kollerup, Si Chen, Simo Hosio, Koji Yatani, Yi-Chieh Lee, Ujwal Gadiraju, Niels van Berkel, and Jorge Goncalves. 2026. Understanding, Mitigating, and Leveraging Cognitive Biases to Calibrate Trust in Evolving AI Systems. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3772363.3778722

1 Motivation

AI systems have evolved far beyond mere tools; they now act as collaborators, advisors, and decision-makers in critical domains such as healthcare, hiring, and transportation [9, 43], as well as social companions in everyday life [34, 45, 48]. The past few years have seen an unprecedented leap in AI capabilities: where traditional models once produced terse predictions, conversational and voice agents such as GPT-5 [38] now generate fluent, persuasive, and seemingly confident natural language responses, fundamentally transforming human-AI interaction. This paradigm shift raises a critical question: *How can users learn to trust AI when appropriate, and equally, to distrust it when necessary?*

Appropriate trust is pivotal to effective human-AI interaction. Achieving this balance relies on how well end-users can *calibrate their trust*, i.e., align their trust in AI with the AI's actual capabilities and limitations [30, 46]. Miscalibration can lead to *over-reliance*, where users forgo their decision agency and accept AI decisions uncritically [41, 54], or *under-reliance*, where users remain sceptical of AI advice, dismissing valuable assistance [24, 44].



CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/2026/04
https://doi.org/10.1145/3772363.3778722

It is noteworthy that **AI systems have been around for decades, yet the challenge of fostering appropriate trust in these systems persists.** *Why have we struggled to design approaches that help users trust AI only when warranted, and what aspects of human cognition contribute to this miscalibration?* Human cognition is shaped by bounded rationality [47]: our limited capacity for attention, memory, knowledge, effort, time, and reasoning leads us to depend on heuristics, or mental shortcuts, to navigate decisions and make judgements [18, 31]. These heuristics often serve us well, allowing us to manage complexity and act effectively in everyday contexts [19]. However, heuristics can also distort perceptions and behaviours, giving rise to **cognitive biases** — defined by Tversky and Kahneman [25, 52] as systematic, unconscious tendencies to deviate from rational decision-making. For example, *anchoring bias* leads individuals to overly rely on the first piece of information they encounter [52], while *confirmation bias* drives people to seek and interpret information that supports their existing beliefs [37]. To date, over 180 cognitive biases have been documented [2, 17], with many more *"systematic"* behavioural effects yet to be classified as cognitive biases [26]. Recent discourses in psychology suggest these biases are not flaws but **intrinsic features of human cognition** [21, 39]. Yet, AI systems are rarely designed with them in mind [3, 36], often resulting in persistent challenges of miscalibrated trust.

Recent discussions in the HCI community suggest that computing systems—including AI—tend to **trigger and amplify existing cognitive biases** in users [3, 7, 8]. Subsequently, these biases systematically shape, or skew, the way individuals interact with systems, nudging them towards certain behaviours or decisions [10]. If not properly designed, AI systems can trigger undesired biases, and thus, make users over- or under-rely on AI [22, 40, 41]. For instance, *automation bias* can amplify over-trust in (incorrect) AI recommendations [20, 41], while *anchoring bias* can cause initial impressions of AI performance to disproportionately shape subsequent judgements, regardless of actual AI accuracy [1, 42, 51].

We posit that AI systems can and should be designed to **alleviate the harmful effects of biases**. One promising direction is to equip AI systems with cognitive assistance tools [13, 32, 53, 55] that help users reflect and make informed decisions. More recently, researchers [8, 33] have argued that AI systems can **benefit from existing cognitive biases** in users, harnessing them to foster appropriate reliance on AI. For example, Ma et al. [33] utilised *anchoring bias* to promote people's trust in AI. Danry et al. [12] leveraged the *framing effect* to improve human discernment of logically flawed statements. By explicitly accounting for these biases in AI design, and perceiving them as *features* rather than limitations of the human mind, AI systems can avoid triggering their undesired effects, and harness useful heuristics. Such approaches may not only help individuals manage decision-making but also achieve more optimal human-AI interactions. However, this is to be noted with the caveat that mitigating or leveraging the effects of cognitive biases may sacrifice user agency or experience [6], as these biases are generally automatic and happen without our awareness.

The advent of Large Language Models (LLMs) has further intensified the challenge of trust calibration in AI. Unlike traditional, task-specific AI systems, LLMs engage users in dynamic, back-and-forth conversational interactions. This paradigm shift raises pressing questions about how cognitive biases manifest in such fluid, personalised, and persuasive settings. For example, the *authority bias*—where people defer to confident or authoritative-seeming sources [35]—may be exacerbated as LLMs deliver responses with fluent and persuasive language. Similarly, biases like *anchoring* and *confirmation bias* may surface in new ways, as users' first impressions of an LLM's output or the model mirroring users' own beliefs can disproportionately shape subsequent trust dynamics. Recent research [11, 29] further argues that LLMs not only mimic human-like behaviour but also express cognitive biases themselves. Such interactions could reinforce existing biases, exacerbate over-reliance, or shape users' trust in unpredictable ways.

These challenges make clear that understanding and addressing cognitive biases in human–AI trust calibration is both urgent and unfinished work. It is also a pressing issue in HCI to design technologies that are calibrated to aspects of human cognition. This workshop will create a dedicated space for researchers from HCI, cognitive science, psychology, AI, and interaction design to come together to examine how cognitive biases shape trust calibration in emerging AI systems. Participants will share discipline-specific perspectives, map open challenges, and identify concrete research directions. The intended outcome is a shared foundation for ongoing collaboration and a clearer research agenda for re-framing AI system design — treating cognitive biases not as shortcomings or limitations of end-users, but rather as features of human cognition that must be acknowledged and meaningfully integrated into design.

1.1 Workshop Topics

We structure the workshop around **three** focused topics sitting at the intersection of cognitive biases and trust calibration in AI systems. Each topic addresses a critical challenge in aligning human trust with AI capabilities, while accounting for the realities of human cognition. We designed these topics to spark interdisciplinary dialogue and encourage participants to contribute insights and questions from their respective lenses. Together, these discussions will lay the groundwork for a shared research agenda for designing human–AI trust calibration approaches that account for cognitive biases as features of human cognition.

(1) **Understanding and Mapping Biases in Human-AI Interaction:** Despite decades of research, the ways cognitive biases shape user trust and reliance in AI systems remain under-characterised, and are also consistently evolving as AI evolves. Scholars [8, 26] also argue that there is a plethora of systematic effects not accounted for as cognitive biases. We ask: In what ways can cognitive biases manifest in human–AI interaction across domains and tasks (e.g., structured decision support vs. conversational LLM settings)? Where do biases typically arise in the human–AI interaction pipeline? How are these biases currently operationalised in research, or where might their role be overlooked? What patterns of *over-* and *under-reliance* emerge as a result of these biases? Can behaviours observed in existing empirical human-AI research be explained, at least in part, through these heuristics and biases? To support this mapping, we urge participants to draw on established cognitive bias clusters such as *Need*

to Act Fast, *Too Much Information*, *Too Little Information*, and *Memory Limitations* [2, 17].

(2) **Design Strategies — Mitigating Biases *and* Leveraging Heuristics:** This topic considers how to effectively respond to biases through meaningful design, without falling into overly corrective approaches or those that sacrifice user agency or experience. What interventions (e.g., transparency scaffolds, system framing, delays before seeing AI response, just-in-time cues, or behavioural "nudges") could mitigate the harmful effects of biases *without* undermining user autonomy, understanding, or accuracy? What are some design tensions at play here, and what ethical trade-offs emerge? When and (should) cognitive biases and heuristics be harnessed to *improve* trust calibration? How can we design AI systems with cognitive biases and heuristics in mind? What novel design solutions could LLMs offer to mitigate/leverage cognitive biases?

(3) **Methods and Case Studies:** Studying how biases and heuristics shape trust is methodologically complex: these phenomena are often tacit and context-sensitive. The HCI community lacks a common standard to study cognitive biases [8]. This topic invites reflection on how best to empirically capture, quantify, or observe trust miscalibration driven by biases—whether in the lab or even in the wild. What tools, study designs, measures, and signals may capture, quantify, or observe bias-driven trust dynamics? What blind spots exist in our empirical approaches today, and what would more robust, ecologically valid, and interdisciplinary approaches entail?

2 Organisers

Our team comprises a mix of junior and senior scholars working in different areas of expertise – HCI, AI, Social Computing, and Cognitive Psychology – with a shared agenda of designing AI to better support human cognition and foster appropriate trust. The organisers are located across Europe, America, Asia, and Oceania.

• **Saumya Pareek** is a PhD candidate at the University of Melbourne, Australia. Her research focuses on identifying and understanding the factors that shape users' trust in AI systems, and leveraging these insights to design strategies that help users calibrate their trust in AI systems effectively. She has presented her work at leading HCI conferences and is involved in interdisciplinary collaborations in the field of HAI [40–42].

• **Nattapat Boonprakong** is a Postdoctoral Research Fellow at National University of Singapore. He investigates how cognitive biases manifest in the interaction between humans and computing systems and how they influence the designs of systems and user interfaces. He has published a series of works on the issue of cognitive biases in HCI [4, 6–8] and has led a successful workshop at CSCW'23 pertaining to cognitive biases in human-AI collaboration [5].

• **Naja Kathrine Kollerup** is a PhD candidate at Aalborg University, Denmark. She investigates the dynamics of trust between users and AI systems, particularly in how these interactions influence the understanding and design of AI within healthcare. She has published and presented her work at HCI conferences [27, 28].

• **Si Chen** is a Postdoctoral Research Fellow at the University of Notre Dame, USA. Her research focuses on the intersection of AI in education, human-computer interaction and AI ethics, particularly emphasising how emerging technologies impact/support educational equity and accessibility through human-centred design. She has co-organised workshops on cognitive biases at CSCW.

• **Simo Hosio** is an Associate Professor at the University of Oulu, Finland, where he leads the Crowd Computing Research Group. His research interests are in digital health, crowd computing, and human-AI interaction. Hosio has experience in successfully organising several past workshops at the CHI conference, as well as at UbiComp and CSCW.

• **Koji Yatani** is an Associate Professor at The University of Tokyo, where he leads the Interactive Intelligent Systems Laboratory. His research lies in Human-Computer Interaction and ubiquitous computing, with a focus on Human-AI Interaction and human well-being support. He also leads an international research initiative, called Mental Well-being Intelligence, supported by JST ASPIRE for Top Scientists.

• **Yi-Chieh Lee** is an Assistant Professor in the Department of Computer Science at NUS. His research lies at the intersection of HCI, social computing, and human-centered AI, with a focus on designing and deploying Conversational AI (CAI) for social good. His work has been published at top-tier venues and recognised with multiple prestigious awards, including Best Paper and Honorable Mention awards and a Google Research Scholar Award.

• **Ujwal Gadiraju** is an Associate Professor at Delft University of Technology, Netherlands, and a Director of the Delft "Design@Scale" AI Lab. His work lies at the intersection of HCI, AI, and information retrieval. His goal is to help people far and wide by fostering meaningful reliance on AI.

• **Niels van Berkel** is a Professor at Aalborg University, Denmark. His work seeks to support and enhance human cognition through digital technology, typically in collaborative and real-world settings. He has (co-)organised various workshops at CHI, UbiComp, and CSCW.

• **Jorge Goncalves** is an Associate Professor at the University of Melbourne, Australia. His research interests include crowdsourcing, social computing, affective computing and human-centred AI. He has also served as Workshops Co-Chair for CHI'19 and CHI'20, and co-organised many successful workshops at leading HCI venues, such as CHI, CSCW, and UbiComp [23, 49, 50].

3 Plan to Publish Proceedings

Accepted submissions are non-archival. The workshop website[1] will host all accepted papers prior to the conference, fostering early community engagement. With the authors' consent, these papers will remain available on the website as a resource for the broader research community. Additionally, we will encourage participants to upload their position papers to arXiv. Authors of the position papers can reuse their work for future peer-reviewed venues.

4 Workshop Mode, Materials, and Accessibility

The workshop will be conducted entirely in person, structured as a **single 90-minute session** prioritising hands-on activities

[1] https://chi-bias-trust.github.io/

and group reflection around the core workshop themes. During the workshop, we will introduce activities and topics through a projected presentation and prompt cards, and use a Miro board as the central tool to visualise and collate insights from all participant activities. This board will serve both as a collaborative workspace during the session and as a shared artefact afterwards.

To allow asynchronous engagement, the Miro board will remain accessible after the workshop, so participants can revisit and review the insights should they wish to carry these ideas forward beyond the in-person session. We will create a Slack channel for this workshop to facilitate communication before, during, and after the workshop. All workshop materials, including activity prompts, summaries, and outputs from the activities, will be shared on the workshop website and through our Slack channel, allowing those unable to attend the full workshop or those who need additional time to reflect and share their input.

To ensure accessibility and inclusivity, we will provide automatic live transcription of presentations and discussions within the workshop using Zoom. All workshop materials and prompt cards will be made available in large-print and audio-friendly formats. The Miro boards, workshop materials, and our website will comply with accessibility standards/screen-reader requirements.

5 Workshop Activities

5.1 Pre-Workshop Plans

We will promote the workshop across multiple channels, including social media, existing mailing lists and Slack servers of HCI conferences/workshops, university networks, relevant research groups, and upcoming HCI conferences. We will share the Call for Participation (Section 6) through these channels, aiming to invite a diverse range of participants.

We will invite participants to submit a position paper in the form of: (1) an essay (1–2 pages) stating their research background and motivation for attending this workshop, or (2) a short / abridged paper (2–4 pages excluding references) presenting research contributions that align with one or more of the workshop topics. The workshop organisers will review the submissions, selecting those that can spark meaningful discussions, provide diverse perspectives, and contribute to the workshop's goals. We will focus on inviting participants with relevant expertise in human-AI interaction, cognitive biases, and trust calibration, and/or those from related fields such as psychology, behavioural science, and AI system design. In addition, we will qualitatively analyse submissions to identify themes and participant expertise, so we can form complementary participant groups that cover a range of disciplines, research methods, and career stages for richer perspectives during activities.

Two weeks prior to the workshop, we will invite all prospective participants to join our Slack channel, where they can interact with one another and the organisers. Participants will also receive a short prompt pack with key bias clusters [2, 17], provocative scenarios, and optional readings to familiarise themselves ahead of the session. We will qualitatively analyse the position papers and sort the participants into small working groups that feature a mix of perspectives and disciplines.. These base groups will be used for Activity-1 before we rotate participants into new mixed groups for subsequent activities to encourage cross-group exchange and interaction. We will share, in advance via our Slack channel, how groups will be formed and rotated (e.g., base groups for Activity-1 and remixed groups for subsequent activities), so that attendees know what to expect and how they will interact with others during the session.

5.2 Workshop Structure

The workshop will span **90 minutes** and be structured around key interactive sessions (see Table 1). Workshop organisers will facilitate each activity to ensure engagement and a smooth flow. Below is an outline of the workshop's structure:

- **Welcome and "Bias Bingo" (minutes 0-10):** We will briefly introduce the motivation and structure of the workshop, framing the challenge of trust calibration and cognitive biases in human-AI interaction. Participants will then take part in a quick interactive icebreaker: "Bias Bingo," a prompt-card-based activity where participants match AI interaction scenarios with potential cognitive biases (e.g., confirmation, availability, or automation bias), to activate prior knowledge and introduce key concepts in a playful way.
- **Activity 1 — Mapping Biases in Human-AI Interaction (minutes 10-30):** In small groups (formed beforehand based on participants' submitted position papers to ensure a mix of perspectives and disciplines), participants will be given fictional human-AI interaction scenarios (e.g., healthcare decision support, LLM chatbots, traditional AI systems). Using colour-coded Miro templates and artefacts, groups will map: which bias clusters are relevant to this interaction context (e.g., *"Need to Act Fast"*, *"Too Much Information"*), where in the interaction pipeline biases are most likely to emerge, and how these biases could be envisioned to cause over- or under-trust. Groups will share 1–2 key insights back with the whole room.
- **Activity 2 — Design Interventions: Mitigating Biases and Leveraging Heuristics (minutes 30–60):** Participants will then move into newly mixed groups (rotated from Activity-1 so each new table combines different backgrounds and prior discussions) and tackle new scenario prompts with a new lens: how can AI systems and/or interfaces be designed to better support trust calibration without undermining user autonomy? They will draw on insights from their own position papers, and/or choose from a menu of design elements (e.g., nudges, delays, confidence cues, transparency scaffolds, etc.), and/or brainstorm new interventions. Groups will discuss how these interventions could mitigate adverse effects or even leverage useful aspects of the identified biases. Groups will also annotate their board with the trade-offs or ethical tensions that become salient as they think deeply about their interventions. They will add their outputs to our Miro board to be synthesised later.
- **Activity 3 — Methods and Future Research Directions (minutes 60–80):** Groups will be reshuffled again, and will reflect on the methods they have used in their own submissions (or envision using) to study how cognitive biases impact trust. Reflective prompts include: *"What is hard to observe or measure about cognitive biases?"*, *"What have existing methods missed?"*,

Table 1: Proposed Workshop Schedule (90 minutes)

Duration	Session
Pre-Workshop	
-	Participants introduce themselves in the workshop's Slack channel and access pre-reading materials and a provocative prompt pack.
Workshop Session (90 min)	
10 mins	Welcome and "Bias Bingo" Icebreaker
20 mins	Activity 1 — Mapping Biases in Human-AI Interaction
30 mins	Activity 2 — Design Interventions: Mitigating or Leveraging Heuristics
20 mins	Activity 3 — Methods and Future Research Directions
10 mins	Collective Synthesis and Next Steps
Post-Workshop	
-	Post-Workshop: Gathered insights and results posted on the workshop website, initiating follow-ups/reflections.

"If resources were unlimited, what study would you design to understand this phenomenon?" Groups will post their *critiques and wish-list* methods to the Miro board, which we will cluster in real time to identify broader themes.

- **Collective Synthesis and Next Steps (minutes 80–90):** We will close the workshop with a final plenary discussion where we summarise key patterns, ideas, and tensions that emerged from the group activities. This will explicitly surface connections across groups so that insights can be shared beyond individual tables. Participants, having spent the session immersed in mapping biases, debating interventions, and thinking about methods to observe biases and their impact on trust calibration, will be asked to post the most urgent open questions or research directions on a live board (e.g., PollEv) projected in the room. We will conclude by encouraging participants to stay connected through our Slack channel to promote ongoing engagement and collaboration. Finally, an optional social event will follow, where we can continue networking and extending our conversations beyond the workshop.

5.3 Post-Workshop Plans

We have a three-part plan to ensure the continuation and amplification of our workshop's impact. First, we will maintain community engagement through our dedicated Slack channel, enabling participants to continue conversations, ask questions, share follow-up work, foster collaborations around cognitive biases and AI trust calibration, and seek mentorship from the senior members of the organising committee. Second, our workshop website will serve as an archival repository, hosting all materials such as papers and essays, workshop probes and prompt cards, Miro boards, and summaries of discussions, providing a lasting resource for participants and the broader community to return to. We will also share updates on any follow-on activities (e.g., collaborative projects) initiated by the community. Third, the organising team will collect and synthesise the insights gathered across activities (bias mapping, generating design interventions, methodological reflections, and the identification of future research directions and important questions for the field) into a collaboratively-written publication. Participants will be asked, before the workshop commences, to indicate whether they consent to the use of their workshop inputs (Miro boards, PollEv responses) as research data for this synthesis. All consenting participants will be acknowledged in the resulting publication. In addition, participants will be given the option to join a post-workshop writing group: those who opt in and make substantial contributions to framing, analysis, and/or drafting or revising the manuscript will be invited as co-authors, in line with common authorship guidelines. Importantly, these options around data use, acknowledgement, and potential co-authorship will be clearly communicated in pre-workshop materials and on our workshop website so participants are informed and their contributions treated in a transparent way. In the resulting publication, we will identify key patterns, tensions, and open questions, offering the community a shared foundation and agenda for advancing research on cognitive biases and trust calibration in emerging AI systems.

6 Call For Participation

Calibrating trust in AI systems remains a persistent challenge: users often struggle to decide when to accept or question AI recommendations. Cognitive biases—systematic deviations from rational judgement—play a significant role in these difficulties. As AI grows more sophisticated with generative models and LLMs, these biases may manifest in novel ways, influencing trust in ways we do not yet fully understand. This workshop explores how cognitive biases shape trust calibration in AI, and how they might be mitigated, or even leveraged, to design more trustworthy and human-aligned systems. We invite researchers and practitioners from HCI, human-centred AI, cognitive psychology, and related fields to join our 90-minute, in-person workshop at CHI2026. The session will feature interactive activities where we collaboratively identify and map biases in human–AI interaction, brainstorm design strategies and solutions, and reflect on challenges and tensions. Together, we will identify pressing research directions and create a shared agenda for advancing this space. To participate, please submit either a motivation essay (1–2 pages) or an abridged short paper (2–4 pages) surrounding one or more workshop themes: understanding and mapping biases in human-AI interaction, designing strategies to mitigate adverse effects of biases or leverage useful

heuristics, and methodologies or case studies for capturing bias-driven trust dynamics. Page limits exclude references. Submissions should follow the CHI Extended Abstract template, be submitted via Google Forms https://forms.gle/8gCa1TAffyNEyA2T9 and will be reviewed based on quality and relevance. Accepted papers will be published on our website with participants' consent. We expect 25–35 participants and welcome diverse perspectives, including those underrepresented in the HCI community. Members of the organising team will attend the workshop. For more information, please visit: https://chi-bias-trust.github.io/.

7 Expected Size of Attendance

We aim to invite around 25–35 participants (excluding organisers) from co-authors of the accepted workshop submissions. We expect participants to be a good mix of senior and junior academics, spanning diverse expertise (i.e., HCI, AI, cognitive science, and psychology) and research methodologies (i.e., both quantitative and qualitative). In addition, we will extend invitations to leading researchers in human-AI interaction and related fields to participate in our workshop and enrich the discussions by sharing their perspectives during group activities.

8 Note About Past Workshops

This workshop builds on a series of prior workshops at CHI [14, 15], CSCW [5], and Dagstuhl [16], which examined the design of technologies that support human cognition and decision-making through the lens of cognitive biases. While these earlier workshops established the importance of recognising biases in decision-making, this proposed workshop extends the conversation to the specific and timely context of *trust calibration in AI systems*, particularly in light of emerging paradigms such as Generative AI and LLMs. As a specialised extension, it represents the continuation of an established HCI research agenda, while also advancing it to proactively shape how cognitive biases are understood, mitigated, and leveraged in the design of trustworthy AI systems.

References

[1] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels van Berkel. 2023. "If I Had All the Time in the World": Ophthalmologists' Perceptions of Anchoring Bias Mitigation in Clinical AI Support. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–14. doi:10.1145/3544548.3581513

[2] Buster Benson. 2016. Cognitive Bias Cheat Sheet. https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18. Accessed: 2024-07-15.

[3] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-Assisted Decision-Making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 78–91. doi:10.1145/3514094.3534164

[4] Nattapat Boonprakong, Xiuge Chen, Catherine Davey, Benjamin Tag, and Tilman Dingler. 2023. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases When Facing Different Opinions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). doi:10.1145/3544548.3580917 Publisher: Association for Computing Machinery.

[5] Nattapat Boonprakong, Gaole He, Ujwal Gadiraju, Niels Van Berkel, Danding Wang, Si Chen, Jiqun Liu, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2023. Workshop on Understanding and Mitigating Cognitive Biases in Human-AI Collaboration. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 512–517. doi:10.1145/3584931.3611284

[6] Nattapat Boonprakong, Saumya Pareek, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. Assessing Susceptibility Factors of Confirmation Bias in News Feed Reading. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 418, 19 pages. doi:10.1145/3706598.3713873

[7] Nattapat Boonprakong, Benjamin Tag, and Tilman Dingler. 2023. Designing Technologies to Support Critical Thinking in an Age of Misinformation. *IEEE Pervasive Computing* (2023), 1–10. doi:10.1109/MPRV.2023.3275514

[8] Nattapat Boonprakong, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. How Do HCI Researchers Study Cognitive Biases? A Scoping Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 473, 20 pages. doi:10.1145/3706598.3713450

[9] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 104:1–104:24. doi:10.1145/3359206

[10] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–15. doi:10.1145/3290605.3300733 Publisher: Association for Computing Machinery.

[11] Vanessa Cheung, Maximilian Maier, and Falk Lieder. 2025. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences* 122, 25 (2025), e2412015122. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2412015122 doi:10.1073/pnas.2412015122

[12] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages. doi:10.1145/3544548.3580672

[13] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW048 (May 2025), 30 pages. doi:10.1145/3710946

[14] Tilman Dingler, Benjamin Tag, Evangelos Karapanos, Koichi Kise, and Andreas Dengel. 2020. Workshop on Detection and Design for Cognitive Biases in People and Computing Systems. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3334480.3375159

[15] Tilman Dingler, Benjamin Tag, Philipp Lorenz-Spreen, Andrew W. Vargo, Simon Knight, and Stephan Lewandowsky. 2021. Workshop on Technologies to Support Critical Thinking in an Age of Misinformation. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–5. doi:10.1145/3411763.3441350

[16] Tilman Dingler, Benjamin Tag, and Andrew Vargo. 2022. Technologies to Support Critical Thinking in an Age of Misinformation (Dagstuhl Seminar 22172). *Dagstuhl Reports* 12, 4 (2022), 72–95. doi:10.4230/DagRep.12.4.72

[17] Cognitive Bias Foundation. 2024. Bias Cheat Sheet. http://bias.transhumanity.net/bias-cheat-sheet/. Accessed: 2024-09-13.

[18] Gerd Gigerenzer. 2004. *Fast and Frugal Heuristics: The Tools of Bounded Rationality.* John Wiley & Sons, Ltd, Chapter 4, 62–88. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470752937.ch4 doi:10.1002/9780470752937.ch4

[19] Gerd Gigerenzer, Peter M. Todd, and A. B. C. Research Group. 1999. *Simple Heuristics That Make Us Smart.* Oxford University Press USA, New York, NY, USA.

[20] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2014. Automation bias: empirical results assessing influencing factors. *International Journal of Medical Informatics* 83, 5 (May 2014), 368–375. doi:10.1016/j.ijmedinf.2014.01.001

[21] Martie G. Haselton, Daniel Nettle, and Paul W. Andrews. 2015. *The Evolution of Cognitive Bias.* John Wiley & Sons, Ltd, Chapter 25, 724–746. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470939376.ch25 doi:10.1002/9780470939376.ch25

[22] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. doi:10.1145/3544548.3581025

[23] Danula Hettiachchi, Mark Sanderson, Jorge Goncalves, Simo Hosio, Gabriella Kazai, Matthew Lease, Mike Schaekermann, and Emine Yilmaz. 2021. Investigating and Mitigating Biases in Crowdsourced Data. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21)*. Association for Computing Machinery, New York, NY, USA, 331–334. doi:10.1145/3462204.3481729

[24] S Mo Jones-Jang and Yong Jin Park. 2023. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication* 28, 1 (Jan. 2023), zmac029. doi:10.1093/jcmc/

zmac029
[25] Daniel Kahneman. 2011. *Thinking, Fast and Slow.* Macmillan.
[26] Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz. 2021. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence* 295 (2021), 103458. doi:10.1016/j.artint.2021.103458
[27] Naja Kathrine Kollerup, Stine S. Johansen, Martin Grønnebæk Tolsgaard, Mikkel Lønborg Friis, Mikael B. Skov, and Niels van Berkel. 2024. Clinical needs and preferences for AI-based explanations in clinical simulation training. *Behaviour & Information Technology* 0, 0 (2024), 1–21. arXiv:https://doi.org/10.1080/0144929X.2024.2334852 doi:10.1080/0144929X.2024.2334852
[28] Naja Kathrine Kollerup, Joel Wester, Mikael B. Skov, and Niels van Berkel. 2024. How Can I Signal You To Trust Me: Investigating AI Trust Signalling in Clinical Self-Assessments. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 525–540. doi:10.1145/3643834.3661612
[29] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking Cognitive Biases in Large Language Models as Evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 517–545. doi:10.18653/v1/2024.findings-acl.29
[30] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. doi:10.1518/hfes.46.1.50_30392
[31] Falk Lieder and Thomas L. Griffiths. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* 43 (2020), e1. doi:10.1017/S0140525X1900061X
[32] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 261, 23 pages. doi:10.1145/3706598.3713423
[33] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). doi:10.1145/3544548.3581058 Publisher: Association for Computing Machinery.
[34] Takuya Maeda and Anabel Quan-Haase. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1068–1077. doi:10.1145/3630106.3658956
[35] Stanley Milgram. 1963. Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology* 67, 4 (1963), 371–378. doi:10.1037/h0040525
[36] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. doi:10.1016/j.artint.2018.07.007
[37] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.
[38] OpenAI. 2025. *Introducing GPT-5.* https://openai.com/index/introducing-gpt-5/
[39] Lionel Page. 2022. *Optimally Irrational: The Good Reasons We Behave the Way We Do.* Cambridge University Press.
[40] Saumya Pareek, Sarah Schömbs, Eduardo Velloso, and Jorge Goncalves. 2025. "It's Not the AI's Fault Because It Relies Purely on Data": How Causal Attributions of AI Decisions Shape Trust in AI Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–18. doi:10.1145/3706598.3713468
[41] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction* CSCW2 (2024), 383:1–383:31. doi:10.1145/3686922
[42] Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2024. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM, 546–561. doi:10.1145/3630106.3658924
[43] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 125–134. doi:10.1609/hcomp.v7i1.5281
[44] Andrew Prahl and Lyn Van Swol. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36, 6 (2017), 691–702. doi:10.1002/for.2464
[45] Replika. 2024. *Replika: The AI companion who cares.* https://replika.com/
[46] Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422. doi:10.1145/3581641.3584066 arXiv:2302.02187 [cs].
[47] Herbert A Simon. 1957. A behavioral model of rational choice. *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting* (1957), 241–260.
[48] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies* 149 (2021), 102601. doi:10.1016/j.ijhcs.2021.102601
[49] Benjamin Tag, Sarah Webber, Greg Wadley, Vanessa Bartlett, Jorge Goncalves, Peter Koval, Petr Slovak, Wally Smith, Tom Hollenstein, Anna L Cox, and Vassilis Kostakos. 2021. Making Sense of Emotion-Sensing: Workshop on Quantifying Human Emotions. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Virtual, USA) *(UbiComp '21)*. Association for Computing Machinery, New York, NY, USA, 226–229. doi:10.1145/3460418.3479272
[50] Garreth W. Tigwell, Zhanna Sarsenbayeva, Benjamin M. Gorman, David R. Flatla, Jorge Goncalves, Yeliz Yesilada, and Jacob O. Wobbrock. 2019. Addressing the Challenges of Situationally-Induced Impairments and Disabilities in Mobile Interaction. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3290607.3299029
[51] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)*. ACM. doi:10.1145/3450613.3456817
[52] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.
[53] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300831
[54] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3544548.3581197
[55] Junti Zhang, Zicheng Zhu, Jingshu Li, and Yi-Chieh Lee. 2025. Mining Evidence about Your Symptoms: Mitigating Availability Bias in Online Self-Diagnosis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 93, 23 pages. doi:10.1145/3706598.3713805