COURSE

REFLECT: Tutorial on Reflecting on Bias in LLMs through Human-Centered Perspectives

ANTONELA TOMMASEL, Johannes Kepler University Linz, Linz, Upper Austria, Austria

MARKUS SCHEDL, Johannes Kepler University Linz, Linz, Upper Austria, Austria

RALPH HERTWIG, Max Planck Institute for Human Development, Berlin, Germany

Open Access Support provided by:

Max Planck Institute for Human Development

Johannes Kepler University Linz

Published: 22 March 2026

Citation in BibTeX format

IUI '26: 31st International Conference on Intelligent User Interfaces
March 23 - 26, 2026
Paphos, Cyprus

Conference Sponsors:
SIGCHI
SIGAI

https://doi.org/10.1145/3742414.3794952
ISBN: 9798400719851

REFLECT: Tutorial on Reflecting on Bias in LLMs through Human-Centered Perspectives

Antonela Tommasel
Institute of Computational Perception
Johannes Kepler University
Linz, Austria
ISISTAN
CONICET-UNICEN
Tandil, Argentina
antonela.tommasel@jku.at

Markus Schedl
Institute of Computational Perception
/ Multimedia Mining and Search Group
Johannes Kepler University
Linz, Austria
markus.schedl@jku.at

Ralph Hertwig
Adaptive Rationality
Max Planck Institute for Human Development
Berlin, Germany
hertwig@mpib-berlin.mpg.de

Abstract

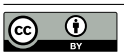
Large Language Models (LLMs) increasingly shape how people access, produce, and reason with information. These models do not merely generate language, they mirror the data, discourse, and cognitive patterns on which they are trained. As a result, they often reflect and amplify existing social and cognitive biases, and even stereotypical representations, influencing what information is surfaced, how perspectives are represented, and which voices are privileged or silenced. Understanding these reflections requires going beyond technical bias detection toward examining how bias emerges in LLM outputs, how users perceive and react to it, and how design choices can reinforce or mitigate biased interactions. REFLECT offers a human-centered exploration of bias in LLMs, connecting perspectives from computer science, human–computer interaction, and cognitive psychology. REFLECT examines multiple ways in which bias manifests in LLMs (from selection effects to acquiescence and stereotypical associations) and discusses what these manifestations reveal about the interaction between human data, model training, and generative processes, also exploring interaction and design strategies that can help make such reflections visible and open to critical interpretation. Designed as a half-day session, REFLECT provides a concise yet reflective introduction to understanding bias as an emergent property of LLMs. By the end, participants will be equipped with conceptual and practical tools to identify and interpret how LLMs reflect biases, fostering more transparent, accountable, and trustworthy human–AI interactions.

CCS Concepts

• **Human-centered computing** → **Natural language interfaces**;
• **Computing methodologies** → **Natural language generation**.

Keywords

Large Language Models, Cognitive effects, Human–AI Interaction, Bias in AI



IUI Companion '26, Paphos, Cyprus

ACM ISBN 979-8-4007-1985-1/26/03
https://doi.org/10.1145/3742414.3794952

ACM Reference Format:
Antonela Tommasel, Markus Schedl, and Ralph Hertwig. 2026. REFLECT: Tutorial on Reflecting on Bias in LLMs through Human-Centered Perspectives. In *Companion Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI Companion '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3742414.3794952

1 Description of the tutorial topic

Large Language Models (LLMs) increasingly mediate how people access, generate, and reason with information. Beyond their technical capabilities, LLMs act as mirrors of data, reproducing linguistic patterns, social representations, cultural framings, and stereotypical associations [6, 7]. As LLMs become central to applications in decision-making, education, and healthcare, it becomes crucial to understand not only how biases emerge in model behavior, but also how users perceive, interpret, and react to them, and how interaction design can reinforce or mitigate their effects [2, 8].

Biases in LLMs can be understood as systematic deviations (cognitive, linguistic, or social) that influence how information is generated, interpreted, and prioritized [3]. These biases do not originate from a single source, they can stem from human cognition, model behavior shaped by data and training, and the dynamics of interaction itself. Cognitive tendencies such as recency effects (overweighting recent information) shape how users interpret outputs, while selection and acquiescence effects reflect artifacts of model training and alignment. At the interaction level, stereotypical associations can emerge when human expectations and model patterns reinforce one another. Nonetheless, not all such tendencies are inherently undesirable, some, like acquiescence effects, may be intentionally introduced to support alignment and user satisfaction. In this sense, bias should not be seen only as a flaw, but as a systematic tendency whose implications depend on context, purpose and use. Viewed together, these biases reveal that what appears as a single "biased output" often results from intertwined cognitive and computational processes that jointly shape how users perceive objectivity and trust in generative systems.

While much prior work has focused on measuring or mitigating certain biases at the model level, user-centered studies reveal a persistent gap between algorithmic definitions of bias and how individuals experience and interpret it in practice [6, 9]. Current approaches to evaluating bias in LLMs tend to focus on model behavior over user experience, paying less attention to how users make sense of biased outputs, how interface framing shapes their

judgments, and how trust and reliance evolve through interaction [6, 9]. Addressing this gap requires reframing bias as an emergent property of human–AI interaction, shaped by user perceptions, interaction design, and sociocultural context [1, 5].

REFLECT introduces participants to current approaches for examining how LLMs reproduce and amplify biases, how users identify and perceive such biases, and how design interventions (e.g., reflective prompts, contrastive examples or contextual guardrails) can support more transparent, accountable, and trustworthy human–AI interactions. The topic aligns with IUI 2026 themes on human-centered AI, trust and reliance in intelligent systems and end-user interaction with LLMs. By bridging empirical bias research, interaction design, and cognitive insights, REFLECT offers the IUI audience a timely and interdisciplinary perspective on making generative AI systems more transparent, accountable, and socially aware. It complements ongoing discussions in IUI on explainable and interactive AI by highlighting bias reflection and user interpretation as critical dimensions of intelligent user interface design.

2 Previous history

To our knowledge, REFLECT will be the first tutorial at the ACM IUI conference centered on examining how LLMs reflect biases through their outputs and interactions. While recent tutorials and workshops at venues such as FAccT and CHI have explored issues of fairness, explainability, and responsible AI, they have primarily emphasized technical mitigation and model auditing rather than analyzing bias as an inherent characteristic of LLM reasoning and generation. At FAccT[1][2] and CHI[3][4], for example, tutorials on responsible prompting and fairness in generative AI focused on deployment practices but did not analyze how bias manifests or is perceived during interaction. Similarly, a KDD's 2025 tutorial[5] expanded the discussion toward multi-agent ecosystems and system-level trustworthiness. While this tutorial provided an important technical taxonomy for understanding LLM agents, it did not address how biases are experienced, perceived, or made interpretable in human interaction with generative models.

REFLECT builds on this foundation by positioning bias not merely as a statistical artifact, but as a semantic and interpretive reflection of human data, discourse, and cognition. Through this focus, REFLECT contributes a missing perspective to the IUI community, connecting the technical behavior of LLMs with human-centered analyses, aligning directly with IUI's themes on bias in LLMs and trust, user control, and intelligent interfaces for generative AI.

3 Organizers

Antonela Tommasel is an Assistant Professor at Johannes Kepler University Linz (JKU, Austria) and at the National University of the Center of the Province of Buenos Aires (UNCPBA, Argentina), as well as a researcher at the National Scientific and Technical Research Council (CONICET, Argentina). Her research lies at the intersection of artificial intelligence and social computing, with a focus on recommender systems. Her recent work examines fairness and biases in LLM–based recommendations, and the use of agentic LLM-based systems to support decision-making processes in software engineering. Beyond algorithmic design, she studies the role of user engagement and misinformation dynamics in online environments, aiming to build systems that are not only accurate but also responsible. She has published in leading venues in recommender systems and user modeling. She has 15 years of teaching and mentoring experience, including work with Latin American and international students through workshops and summer schools.

Markus Schedl is a Full Professor at the Johannes Kepler University Linz (JKU), affiliated with the Institute of Computational Perception, leading the Multimedia Mining and Search group and the Human-centered AI group at the Linz Institute of Technology (LIT) AI Lab. His research interests include information retrieval, recommender systems, natural language processing, and trustworthy AI, with a particular focus on detecting and mitigating bias in retrieval and recommendation algorithms and on psychological aspects of retrieval and recommendation processes. He (co-)authored more than 300 refereed conference papers, journal articles, books, and book chapters, including publications at ACL, EACL, EMNLP, SIGIR, RecSys, Nature Scientific Reports, and RSOS. Markus has 20 years of experience as a lecturer and has given numerous tutorials, for example, on "Psychology-informed Recommender Systems" (*RecSys 2022*) and "Psychological Aspects in Retrieval and Recommendation" (*SIGIR 2025*).

Ralph Hertwig is the Director of the Center of Adaptive Rationality (ARC) at the Max Planck Institute for Human Development in Berlin. His research focuses on models of bounded rationality such as simple heuristics, the importance of learning and decisions from experience, the measurement of risk preferences, and on systematic and scalable ways to change behavior and decisions by boosting people's competences. He has also investigated systematic differences in experimental practices in fields such as economics and psychology, and how different practices shape experimental results including those on cognitive biases. He has published in the leading journals in psychology and beyond including Science, Nature Human behavior, PNAS, Nature Communications, behavioral and Brain Sciences, Psychological Review, Psychological Bulletin, Psychological Science, Perspectives on Psychological Science, Trends in Cognitive Sciences.

4 Participants

We anticipate a moderately sized group of participants (20 to 35). The tutorial is designed for researchers and practitioners in HCI, AI, and cognitive sciences. It will also be relevant to designers and engineers developing LLM-based interfaces who wish to explore strategies to surface, explain, or mitigate bias, as well as to scholars working on fairness, accountability, and transparency who seek to connect technical auditing with human-centered evaluation. Students and early-career researchers entering the field of human–AI interaction and generative AI will also find the session accessible and engaging. The topic's strong relevance on ongoing discussions on bias in LLMs and human-AI interaction suggests that REFLECT will attract substantial interest from the IUI community and beyond.

[1] https://sites.google.com/view/responsible-gen-ai-tutorial/
[2] www.bertforhumanists.org
[3] https://ibm.github.io/responsible-prompting-course/
[4] https://www.hcitogenai.com
[5] https://ymm-cll.github.io/TrustAgent.github.io/

5 Tutorial activities

REFLECT is structured as a *half-day interactive session* combining lectures, demonstrations, and discussions. The agenda balances conceptual grounding, empirical evidence, and reflective engagement.

Opening and conceptual framing (40 minutes). The tutorial begins with a concise introduction to bias in LLMs, presenting it as both a computational and cognitive phenomenon. Through examples, participants examine how LLMs reflect the patterns of their training data, from subtle imbalances in representation to the reinforcement of prevailing narratives, illustrating how these models act as mirrors of collective discourse and worldview. Participants will also discuss how certain systematic tendencies may be intentionally embedded in model design, raising the question of when a bias functions as a feature rather than a flaw. This reflection helps situate bias not as inherently negative, but as a context-dependent phenomenon shaped by both technical goals and human interpretation. This framing establishes the foundation for understanding bias as an emergent property rather than a purely technical defect.

Empirical insights into bias perceptions (30 minutes). This part synthesizes findings from recent user studies on bias perception, trust, and interaction with LLMs. Drawing on empirical research in human–AI interaction, communication studies, and cognitive psychology, the discussion highlights how users interpret biased outputs, how interface framing affects perceived fairness and credibility, and how trust calibration evolves through interaction. Recent studies on bias perception [6, 9], transparency-by-design [5, 7], and user trust in generative AI [2, 8] serve as illustrative cases. Participants will reflect on what these studies reveal about the limits of awareness and the conditions under which users overlook, question, or rationalize bias in AI-generated content.

Exploring bias manifestations (60 minutes). Participants will be guided through demonstrations of how biases emerge in outputs under different prompts and contexts. Participants will be encouraged to compare generated examples, observe the effects of small input variations, and discuss how linguistic framing or topical context influences model behavior. This exploration connects theoretical notions of bias with their concrete realizations in text generation.

Designing for reflection and transparency (40 minutes). This part focuses on how interaction and design strategies can make model biases visible and open to critical interpretation. Through examples from recent studies (such as [1, 4, 10]), participants will discuss how elements such as contrastive outputs, reflective prompts, contextual guardrails or bias disclosures can help users recognize and reason about bias. This discussion connects design practices to broader questions of trust, interpretability, and accountability in intelligent interfaces.

Synthesis and next steps (10 minutes). REFLECT concludes with a short collective reflection summarizing key insights and identifying open research challenges. Participants will be encouraged to consider how the tutorial's ideas could inform their own research on evaluation, visualization, and human-centered design for LLMs.

6 Planned outcomes of the tutorial

The main goal of REFLECT is to foster a shared understanding of how biases emerge, are reflected, and can be critically examined in LLMs through human-centered perspectives. By the end of the session, participants will have gained conceptual and practical tools to analyze bias both as a computational phenomenon and as an interactional experience. They will leave with an informed perspective on how design and evaluation choices can shape users' awareness, trust, and interpretation of LLM-generated content. The tutorial also aims to build a small interdisciplinary network of researchers and practitioners interested in human-centered approaches to bias in generative AI. All tutorial materials will be made publicly available[6] to support reuse in teaching, research, and evaluation studies.

7 Length

This tutorial is proposed as a half-day session (3 hours).

Acknowledgments

This tutorial is funded in whole or in part by the Austrian Science Fund (FWF): https://doi.org/10.55776/COE12.

References

[1] Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. 2024. A taxonomy for human-llm interaction modes: An initial exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.

[2] Minh-Tay Huynh. 2024. Using generative AI as decision-support tools: unraveling users' trust and AI appreciation. *Journal of Decision Systems* (2024), 1–32.

[3] Daniel Kahneman. 2017. *Thinking, Fast and Slow.* Penguin Books Limited.

[4] Jiayang Li, Jiale Li, and Yunsheng Su. 2024. A map of exploring human interaction patterns with LLM: Insights into collaboration and creativity. In *International conference on human-computer interaction*. Springer, 60–85.

[5] Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* (2023).

[6] Hyunseung Lim, Dasom Choi, and Hwajung Hong. 2025. How Do Users Identify and Perceive Stereotypes? Understanding User Perspectives on Stereotypical Biases in Large Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 3241–3253.

[7] Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong C Park. 2024. Ask llms directly," what shapes your bias?": Measuring social bias in large language models. *arXiv preprint arXiv:2406.04064* (2024).

[8] Nargess Tahmasbi, Elham Rastegari, and Minh Truong. 2025. A Conceptual Model of Trust in Generative AI Systems. (2025).

[9] Lu Wang, Max Song, Rezvaneh Rezapour, Bum Chul Kwon, and Jina Huh-Yoo. 2023. People's perceptions toward bias and related concepts in large language models: a systematic review. *arXiv preprint arXiv:2309.14504* (2023).

[10] Mingqian Zheng, Wenjia Hu, Patrick Zhao, Motahhare Eslami, Jena D Hwang, Faeze Brahman, Carolyn Rose, and Maarten Sap. 2025. Let Them Down Easy! Contextual Effects of LLM Guardrails on User Perceptions and Preferences. *arXiv preprint arXiv:2506.00195* (2025).

[6] https://github.com/hcai-mms/uiu2026-tutorial-llm-biases