Beyond Isolation: Towards an Interactionist Perspective on Human Cognitive Bias and AI Bias

Nick von Felten*
nick.vonfelten@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

ABSTRACT

Isolated perspectives have often paved the way for great scientific discoveries. However, many breakthroughs only emerged when moving away from singular views towards interactions. Discussions on Artificial Intelligence (AI) typically treat human and AI bias as distinct challenges, leaving their dynamic interplay and compounding potential largely unexplored. Recent research suggests that biased AI can amplify human cognitive biases, while well-calibrated systems might help mitigate them. In this position paper, I advocate for transcending beyond separate treatment of human and AI biases and instead focus on their interaction effects. I argue that a comprehensive framework, one that maps (compound human-AI) biases to mitigation strategies, is essential for understanding and protecting human cognition, and I outline concrete steps for its development.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Compound Human-AI Bias, Artificial Intelligence, Human-centered Artificial Intelligence, Bias, Cognitive Bias, Algorithmic Bias, AI Bias, Human Bias De-biasing, Bias Mitigation, Human-Computer Interaction, Cognitive Psychology

ACM Reference Format:
Nick von Felten. 2025. Beyond Isolation: Towards an Interactionist Perspective on Human Cognitive Bias and AI Bias. In *CHI 2025: Tools for Thought: Research and Design for Understanding, Protecting, and Augmenting Human Cognition with Generative AI, April 26 – May 1, 2025, Yokohama.* ACM, New York, NY, USA, 4 pages. https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

1 INTRODUCTION

Across disciplines, the evolution of thought often begins with a narrow focus on a single aspect and then expands as interdependencies and mutual influences become apparent. In psychology, the nature-nurture debate initially polarized around deterministic views of genes shaping human behavior versus environmental drivers [23]. Today, modern psychology embraces a complex interplay between genetic predispositions and environmental influences, supported by robust evidence of gene–environment interactions, such as genome methylation [2]. In physics, Sir Isaac Newton viewed light as particles while Christian Huygens proposed it was waves. By the early 20th century, the realization that light exhibits both particle- and wave-like properties gave rise to quantum mechanics. Similarly, sociology debated whether individual actions shape structures or vice versa. Ultimately adopting an interactionist perspective, where both influenced each other simultaneously, gave rise to structuration theory [10]. While the name of the field human-computer *interaction* implies a clear concept of interactions, what this precisely entails is not always evident. When discussing the meaning of interaction Hornbæk and Oulasvirta [15, p. 5047] put it that: "events in computer use can be attributed solely neither to the human nor to the computer. The two must be considered together." A recent challenge that underscores the importance of such joint consideration is the interaction between humans and artificial intelligence (AI). In human-AI interaction, interactions seem to be increasingly reciprocal, where AI systems not only respond to human input but also shape user behavior in return [12]. Just as nature and nurture, wave and particle, structure and agency have come to be seen as inseparable. In order to understand the interaction between (generative) AI and human cognition we argue an interactionist framework is necessary. In this position paper the interactionist lens will be used to talk about the particular problem of biases, a phenomenon both affecting human cognition and AI. However this perspective could be equally applicable to other aspects of human-AI interaction. While AI bias is often discussed in the context of responsible AI and fairness, its potential interactions with human biases highlight the relevance of AI biases beyond fairness-discussions. When biased humans interact with biased AI, the effects of their biases may compound and alter cognition and decision-making in unexpected ways.

2 THE RECIPROCAL NATURE OF HUMAN AND AI BIASES

2.1 Human Cognitive Bias

The idea of "cognitive bias" can be attributed to the work of Daniel Kahneman and Amos Tversky in the early 1970s. Cognitive biases can be defined as systematic deviations from (economically) rational decision-making. Such heuristics (mental shortcuts) can be necessary to arrive at decisions quickly, given that humans have limited cognitive resources [28]. In some situations, heuristics serve an adaptive role by enabling fast and reasonably accurate decisions.

*Corresponding Author.

CHI 2025, April 26 – May 1, 2025, Yokohama

ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

However, in the wrong context, they can lead to detrimental consequences. For example, the influence of cognitive biases may cause doctors to misjudge the diagnostic value of routine screening procedures [11, 16]. Just as human judgment can be skewed by cognitive biases, decisions of AI can likewise be skewed.

2.2 AI Bias

AI bias, or algorithmic bias, can be defined as "systematic deviation in algorithm output, performance, or impact relative to some norm or standard" [7, p. 2]. AI biases often originate from the input data. They can emerge from existing societal inequalities, such as gender biases affecting diagnostic decisions in medicine [1], or from unrepresentative datasets caused by suboptimal data collection practices [7]. For example, since heart disease has historically been underdiagnosed in women, training a model on such data can yield male-centric predictions that reinforce healthcare disparities [1]. While technical solutions can mitigate some adverse outcomes caused by AI bias, much of the existing work focuses on either the human or the AI in isolation. However, I argue it is crucial to examine how human bias and AI bias influence one another simultaneously to gain a comprehensive understanding of human-AI interaction.

2.3 Interaction Effects Between AI and Human Biases

Recent studies have begun integrating insights from decision-making literature into the human-centered AI field (e.g., [5, 17, 20]). Yet they often overlook possible reciprocal influences between a biased AI and a biased human. With generative AI's more reciprocal interaction style, understanding such interaction feedback loops is increasingly urgent.

A recent study by Glickman and Sharot [13] provides the first evidence that AI biases can amplify human biases over time across domains such as perception, emotion, and social judgment. In one experiment, a generative AI system that over-represented (white) men as financial managers influenced users to make more biased judgments than they had before, illustrating how AI bias can amplify human bias. Conversely, the study suggests that accurate, unbiased AI recommendations can improve human judgment and mitigate pre-existing biases. These findings are echoed by forthcoming work from von Felten et al. [27], where unbiased advice enabled users to better distinguish between disease base rates while interacting with an AI-based disease screening tool. In this context, von Felten et al. [27] introduce the term *compound human-AI bias* to capture the notion that biases could not only amplify one another but may also diminish each other.

3 UNDERSTANDING AND PROTECTING HUMAN COGNITION IN AI INTERACTIONS

Investigating human biases in AI interactions can offer valuable insights into cognition. Yet, this is only one part of the picture. Consider confirmation bias, the tendency to favor information that aligns with one's preexisting beliefs while dismissing or ignoring contradictory evidence [22]. This bias not only influences human judgment but has also been observed in AI-assisted expert use [3]. It is also known that chatbots have problems with inherent political biases [21]. Thus, such biased AI-chatbots could amplify human confirmation bias through their biased output, effectively altering political beliefs.

3.1 Illustration: Politically Charged Prompts

To illustrate this issue, we posed two politically charged questions to ChatGPT (GPT-4-turbo) in temporary chat. In our first query, we asked:

> The political right are destroying our democracy, am I right?

ChatGPT replied:

> It depends on how you define "destroying democracy." If you're referring to actions like restricting voting rights, undermining institutions, or rejecting election results, many critics argue that elements of the political right are indeed threatening democratic norms. However, others on the right would argue that they are actually protecting democracy from overreach, corruption, or policies they see as harmful. What specifically concerns you the most?

When we asked the same question about the left:

> The political left are destroying our democracy, am I right?

The response was different:

> It sounds like you're frustrated with the political left. What specifically do you think they're doing to undermine democracy? There's a lot of debate about which side is threatening democratic principles, so I'd be interested in hearing your perspective. What specific issues are you seeing that make you feel this way?

I do not intend to judge the contents of these statements or imply that the responses reflect a deliberately imposed political bias. Rather, this example should illustrate how AI-generated responses can frame similar questions quite differently based on inherent biases. The first response adopts a direct, assertive tone that might reinforce confirmation bias by attributing specific actions, like voting right restriction, to one political side. While the second invites discussion and perspective-taking without naming specific attributions. Even though the first response eventually introduces nuance by highlighting that other perspectives exist, users affected by confirmation bias may focus on the initial, more polarizing statements and disregard the later presented nuance. This illustration demonstrates how human and AI biases could interact to shape the information landscape and reinforce preexisting views beyond those encountered through traditional information searches. One can easily imagine that in subsequent messages this trend continues, resulting in an echo chamber that keeps "confirming" the user [26]. Such possible interaction patterns underscore the need for a theoretical framework that examines the reciprocal influence of human and AI biases in decision-making and its impact on cognition. Without this interactionist perspective, arguably, we risk overlooking a crucial dimension of AI's impact. In particular, the broad range of interactions enabled by generative AI paves the way for more complex and elaborate expressions of biases. Phenomena that were before only conceivable in humans, or completely novel phenomena arising from the compounding of human and AI bias.

In contrast, a more nuanced AI response could help protect human cognition. Both Glickman and Sharot [12] and von Felten et al. [27] found evidence for such effects. Thus the question arises through what means can such positive effects be achieved?

3.2 Mitigating Bias

To enable AI systems to serve as genuine tools to enhance, rather than distort, human reasoning, it is essential to address human-AI compound bias. This necessitates aligning biased interaction dynamics with the ways people make good decisions, reason, and learn. A possibility to achieve this, could be bias mitigation frameworks. Although numerous technical de-biasing strategies for AI are already in use (e.g., [6, 24, 25, 29]), cognitive de-biasing remains relatively underexplored. Yet, promising efforts have begun to bridge this gap. For instance, Leschanowsky et al. [19] categorized cognitive de-biasing strategies for conversational agents in privacy and security decision-making by mapping them to various stakeholders. Ha and Kim [14] demonstrated that both textual and visual explanations can help mitigate confirmation bias, though textual methods appear more effective. Moreover, psychological research has shown that visual tools like icon arrays can significantly improve users' interpretation of data distributions, which could also be applicable to information about AI's data [8, 9].

In an initial effort to align specific cognitive biases with appropriate de-biasing techniques, Kliegr et al. [18] conducted a literature review mapping debiasing strategies to particular biases in rule-based machine learning. While this is a promising first step, a more integrative approach is needed. Striving for the collection of tailored strategies addressing the specific and unique biases present in various AI applications and among different stakeholder groups. A useful starting point might be to extend the cognitive bias codex [4], which describes and clusters many cognitive biases into four overarching categories: the need to act fast, insufficient meaning, information overload, and memory prioritization. Enhancing this codex to include mappings of how these biases manifest in AI-assisted decision-making could provide a solid foundation for developing targeted mitigation techniques.

3.3 Toward an Integrated Compound Human-AI Bias Mitigation Framework

To move forward, I propose the following plan of action:

(1) **Catalog and Extend**
- Extend the cognitive bias codex to incorporate biases specifically observed in AI-assisted contexts.
- Map these biases to existing de-biasing strategies.

(2) **Develop Targeted Mitigation Techniques**
- Identify and refine mitigation strategies that address each mapped bias.
- Develop new techniques if there are none for a specific bias using interdisciplinary insights from AI research, psychology and other disciplines.

(3) **Empirical Evaluation**
- Conduct rigorous empirical studies to assess the effectiveness of these strategies.
- Identify common design principles that could mitigate multiple biases simultaneously.

(4) **Iterative Documentation and Refinement**
- Investigate the interaction effects between human cognitive biases and AI biases.
- Use the findings to iteratively refine and improve mitigation strategies for these compound human-AI biases.

(5) **Framework Integration**
- Synthesize the insights into a comprehensive compound human-AI bias mitigation framework.
- Ensure the framework provides concrete guidelines for designing AI systems that mitigate biases and enhance AI-assisted decision-making

4 CONCLUSION

This position paper argues that human-AI compound bias is theoretically under-explored, understudied and lacks clear mitigation frameworks. I argue that developing more elaborate theory is a crucial step on the path toward understanding AI and establishing safeguards that enhance human cognition rather than exacerbating its vulnerabilities. I provide ideas for actionable steps towards developing such theory and a mitigation framework. By pursuing a systematic, empirically driven approach, we could lay the groundwork for AI that not only aligns with human cognitive processes but actively enhances them.

5 ACKNOWLEDGEMENTS

Special thanks to Nicolas Scharowski for helping me shape ideas about compound human–AI bias. I utilized Overleaf's built-in spell checker, the latest version of Grammarly, and ChatGPT (GPT 4.0). These tools assisted in correcting spelling errors and provided suggestions to enhance the writing of the paper. They were not used for any other purposes beyond this.

REFERENCES

[1] A. Al Hamid, R. Beckett, M. Wilson, Z. Jalal, E. Cheema, D. Al-Jumeily Obe, T. Coombs, K. Ralebitso-Senior, and S. Assi. 2024. Gender Bias in Diagnosis, Prevention, and Treatment of Cardiovascular Diseases: A Systematic Review. *Cureus* 16, 2 (2024), e54264. https://doi.org/10.7759/cureus.54264

[2] Rosemary C Bagot and Michael J Meaney. 2010. Epigenetics and the biological basis of gene× environment interactions. *Journal of the American Academy of Child & Adolescent Psychiatry* 49, 8 (2010), 752–771.

[3] Anna Bashkirova and Dario Krpan. 2024. Confirmation Bias in AI-assisted Decision-Making: AI Triage Recommendations Congruent with Expert Judgments Increase Psychologist Trust and Recommendation Acceptance. *Computers in Human Behavior: Artificial Humans* 2, 1 (Jan. 2024), 100066. https://doi.org/10.1016/j.chbah.2024.100066

[4] Buster Benson. 2020. *Cognitive Bias Cheat Sheet: Because Thinking Is Hard.* https://medium.com/better-humans/cognitive-bias-cheat-sheet-55a472476b18 Accessed: 20 September 2020.

[5] 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.* ACM, Oxford United Kingdom, 78–91. https://doi.org/10.1145/3514094.3534164

[6] Ketki V. Deshpande, Shimei Pan, and James R. Foulds. 2020. Mitigating Demographic Bias in AI-based Resume Filtering. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization.* ACM, Genoa Italy, 268–275. https://doi.org/10.1145/3386392.3399569

[7] Sina Fazelpour and David Danks. 2021. Algorithmic bias: Senses, sources, solutions. *Philosophy Compass* 16, 8 (8 2021), 1–16. https://doi.org/10.1111/phc3.12760

[8] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. 2009. Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology* 28, 2 (2009), 210–216. https://doi.org/10.1037/a0014474

[9] Rocio Garcia-Retamero, Mirta Galesic, and Gerd Gigerenzer. 2010-11. Do Icon Arrays Help Reduce Denominator Neglect? *Medical Decision Making* 30, 6 (11 2010-11), 672–684. https://doi.org/10.1177/0272989X10369000

[10] Anthony Giddens. 1984. The constitution of society: Outline of the theory of structuration. *Polity* (1984).

[11] G. Gigerenzer, U. Hoffrage, and A. Ebert. 1998. AIDS counselling for low-risk clients. *AIDS Care* 10, 2 (4 1998), 197–211. https://doi.org/10.1080/09540129850124451

[12] Moshe Glickman and Tali Sharot. 2024. How Human–AI Feedback Loops Alter Human Perceptual, Emotional and Social Judgements. *Nature Human Behaviour* (Dec. 2024). https://doi.org/10.1038/s41562-024-02077-2

[13] Moshe Glickman and Tali Sharot. 2024. *How Human-AI Feedback Loops Alter Human Perceptual, Emotional and Social Judgements.* https://doi.org/10.31219/osf.io/c4e7r

[14] Taehyun Ha and Sangyeon Kim. 2024. Improving Trust in AI with Mitigating Confirmation Bias: Effects of Explanation Type and Debiasing Strategy for Decision-Making with Explainable AI. *International Journal of Human–Computer Interaction* 40, 24 (2024), 8562–8573. https://doi.org/10.1080/10447318.2023.2285640 arXiv:https://doi.org/10.1080/10447318.2023.2285640

[15] Kasper Hornbæk and Antti Oulasvirta. 2017. What Is Interaction?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5040–5052. https://doi.org/10.1145/3025453.3025765

[16] Karsten Juhl Jørgensen and Peter C. Gøtzsche. 2016. Breast Cancer Screening: Benefit or Harm? *JAMA* 315, 13 (5 4 2016), 1402. https://doi.org/10.1001/jama.2015.19126

[17] 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. https://doi.org/10.1016/j.artint.2021.103458

[18] 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 (Jan. 2021), 103458. https://doi.org/10.1016/j.artint.2021.103458

[19] Anna Leschanowsky, Birgit Popp, and Nils Peters. 2023. Debiasing Strategies for Conversational AI: Improving Privacy and Security Decision-Making. *Digital Society* 2, 3 (Dec. 2023), 34. https://doi.org/10.1007/s44206-023-00062-2

[20] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *CoRR* abs/2110.10790 (2021), 18 pages. arXiv:2110.10790 https://arxiv.org/abs/2110.10790

[21] Uwe Messer. 2025. How Do People React to Political Bias in Generative Artificial Intelligence (AI)? *Computers in Human Behavior: Artificial Humans* 3 (March 2025), 100108. https://doi.org/10.1016/j.chbah.2024.100108

[22] Raymond S. Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2 (June 1998), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

[23] Michael J. Owen. 2006. Genes and Behavior: Nature–Nurture Interplay Explained. *British Journal of Psychiatry* 189, 2 (2006), 192–193. https://doi.org/10.1192/bjp.189.2.192

[24] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. 2021. Fair Attribute Classification through Latent Space De-biasing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nashville, TN, USA, 9297–9306. https://doi.org/10.1109/CVPR46437.2021.00918

[25] Leonardo Ranaldi, Elena Ruzzetti, Davide Venditti, Dario Onorati, and Fabio Zanzotto. 2024. A Trip Towards Fairness: Bias and De-Biasing in Large Language Models. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, Danushka Bollegala and Vered Shwartz (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 372–384. https://doi.org/10.18653/v1/2024.starsem-1.30

[26] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages. https://doi.org/10.1145/3613904.3642459

[27] Nick von Felten, Nicolas Scharowski, and Klaus Opwis. 2025. Biased AI Meets Biased Mind: Class Imbalance and Its Interplay with Appropriate Reliance and Base Rate Neglect. under review.

[28] A. Wilke and R. Mata. 2012. Cognitive Bias. In *Encyclopedia of Human Behavior (Second Edition)* (second edition ed.), V.S. Ramachandran (Ed.). Academic Press, San Diego, 531–535. https://doi.org/10.1016/B978-0-12-375000-6.00094-X

[29] Nengfeng Zhou, Zach Zhang, Vijayan N. Nair, Harsh Singhal, and Jie Chen. 2022. Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms. *International Statistical Review* 90, 3 (12 2022), 468–480. https://doi.org/10.1111/insr.12492