Looking For Cognitive Bias In AI-Assisted Decision-Making

Regina de Brito Duarte[1,*], Joana Campos[1]

[1]INESC-ID, Instituto Superior Técnico, Lisbon, Portugal

Abstract

Artificial intelligence (AI) has been widely employed in decision-making contexts. However, AI-assisted decision-making continues to encounter several challenges, including prevalent patterns of over-reliance and under-reliance. This paper provides an analysis of the most common cognitive biases in AI-assisted decision-making, supported by multiple examples from the literature. Various solutions proposed in the literature to address the shortcomings of AI-assisted decision-making, such as Explainable AI techniques or cognitive forcing functions, may mitigate certain biases but potentially exacerbate others.

Keywords

AI-assisted decision-making, Cognitive bias, Human-AI interaction

1. Introduction

The rapid integration of Artificial Intelligence (AI) into society is driven by its remarkable capabilities, which enhance decision-making in fields such as law and healthcare [1]. However, the full impact of AI recommendations on human decisions remains an area of ongoing investigation [2, 3, 4, 5]. To address this, eXplainable AI (XAI) has emerged with the goal of making AI predictions more understandable [6, 7]. Despite this, the effectiveness of XAI faces challenges, such as overreliance, where users place excessive trust in AI [8, 9, 10].

To enhance AI-assisted decision-making, proposals include designing clearer explanations [11, 12, 13] and implementing cognitive forcing functions—techniques designed to increase user engagement in AI-assisted decision-making. These functions, such as decision checklists, delayed AI responses, or AI suggestions on demand, are intended to boost user attention [14]. While these approaches address cognitive biases inherent in AI-assisted decision-making [15], the same strategies that mitigate certain cognitive biases can unintentionally trigger others.

This extended abstract explores the identification and discussion of the most common cognitive biases in AI-assisted decision-making, along with their implications for the field. The aim is to highlight design considerations related to cognitive biases for XAI and human-AI interface designers and to provide a comprehensive perspective on how to approach cognitive biases in AI-assisted decision-making.

2. The AI-assisted decision-making process

Hoffman et al. propose a three-stage model to define the traditional decision-making process that includes situation assessment, interpretation, and selection [16]. The model begins with gathering and evaluating relevant information to define the problem and set goals, followed by analyzing this information to develop a plan of action, and concludes with selecting and committing to a specific course of action. These stages provide a structured approach that can vary depending on the decision task at hand.

HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 10—14, 2024, Malmö, Sweden

*Corresponding author.

reginaduarte@tecnico.ulisboa.pt (R. d. B. Duarte); joana.campos@tecnico.ulisboa.pt (J. Campos)

0000-0003-0249-8319 (R. d. B. Duarte); 0000-0002-0113-2211 (J. Campos)

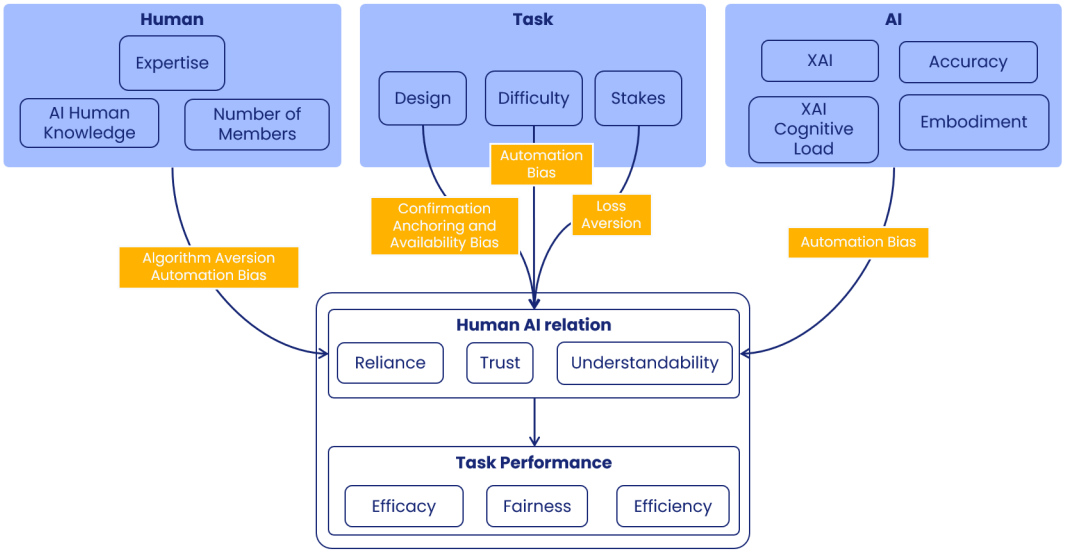

Figure 1: Simplified framework of the AI-assisted decision-making process. This framework includes the three main components that can affect the decision-making process: the Human Decider, Task Characteristics, and AI Agent. It also highlights the principal metrics used to evaluate the AI-assisted decision-making process, which can be influenced by all the components.

In AI-assisted decision-making, the traditional decision-making stages of situation assessment, interpretation, and selection/commitment are preserved. AI improves the interpretation stage by evaluating options, assessing their value, and considering potential outcomes [16, 12]. It typically offers high-accuracy recommendations that can be further enhanced with confidence intervals and explanations. Despite these advantages, the assumption that AI-assisted decision-making is always more efficient than human decision making is sometimes challenged [9, 14].

AI-assisted decision-making can be understood as a process involving three primary components: the human decision-maker, who is ultimately responsible for the final decision and its outcomes; the decision task with its specific characteristics; and the AI agent that provides recommendations to support the decision-making process. Each component can exhibit different characteristics that influence the decision-making process. For instance, a decision task may vary in complexity (task difficulty), be conducted in a high-stakes or low-stakes environment (risk), demand varying levels of cognitive effort from the user and be designed in different ways (design). Similarly, on the human side, factors like expertise level and whether the decision is made by a group or an individual (number of decision-makers) can impact the process. For the AI agent, aspects such as the accuracy of its recommendations and the types of explanations provided to the user can also influence the final outcome.

In AI-assisted decision-making processes, focusing solely on task performance — such as efficacy, efficiency, and fairness—is not sufficient. It is also essential to consider the human-AI relationship, including whether the human decision-maker relies on the AI appropriately and comprehends the AI's recommendations, as these factors significantly impact task performance. By considering these three components and the related factors that influence task performance, we can develop a framework to understand how AI-assisted decision-making processes function and the dynamics among the various contributing factors. Figure 1 illustrates the AI-assisted decision-making framework with these three components and the key decision metrics for evaluating the process. In the following sections, this framework will provide a clear mental model for understanding where cognitive biases might affect the AI-assisted decision-making process.

3. Cognitive Bias in AI-assisted Decision-Making

The scientific community recognizes that the human mind operates within a dual-process system, where certain cognitive processes are rapid, effortless, and intuitive — generated by System 1 — while others are slower and require greater mental effort — generated by System 2 [17]. This understanding is crucial for understanding human decision making, which often occurs under uncertainty with incomplete information. In such situations, decision-makers rely on heuristics—simple, quick judgments—as proxies for unknown answers. These heuristics are typically generated by System 1 and can lead to cognitive biases if not scrutinized by System 2.

While cognitive biases in classical decision-making have been extensively studied, those arising in AI-assisted decision-making are only now gaining attention [17]. This is due to the recent prominence of AI-assisted decision making and the previously unchallenged belief that AI tools inherently enhance decision efficiency [12]. However, recent studies suggest that AI can mitigate in one hand, and reinforce in another, cognitive biases in decision-making. This section provides an analysis of each cognitive bias in AI-assisted decision-making and how they can impact the decision making process, supported by various examples from the literature.

3.1. Confirmation Bias

Confirmation bias involves seeking information that confirms existing beliefs, disregarding contradictory data, and making decisions that reinforce initial beliefs [17]. In AI-assisted decision-making, it occurs when AI suggestions align with preexisting beliefs, reducing critical thinking [11]. Users may accept or reject recommendations solely on the basis of alignment, neglecting other factors. This bias is more common in lay users than in experts [18]. Additionally, when looking at explanations, users may selectively focus on parts confirming their beliefs [19].

3.2. Automation Bias

Automation bias is the tendency to favor decisions made by automated systems, even when they are prone to errors, leading to overreliance [20]. In AI-assisted decision-making, this cognitive effect occurs, especially when the cognitive load of the decision is high [21, 22] or when the expertise of the human decider is low. Explainable AI [11] and cognitive forcing functions [23, 14] are viewed as solutions that can mitigate this bias.

3.3. Algorithm Aversion Bias

In contrast to automation bias, algorithm aversion bias leads humans to dismiss algorithmic decisions just because it is a machine [24]. In AI-assisted decision-making, users may prefer human recommendations as they perceive them as easier to understand [25]. In critical tasks, individuals may favor human discretion over algorithmic application of fairness principles, as humans can transcend these principles if necessary [26]. This bias can lead to under-reliance and disuse of AI systems.

3.4. Anchoring Bias

The anchor effect, occurs when individuals estimate uncertain quantities, influenced by initial reference points called anchors [27]. These anchors, whether informative or randomly assigned, bias final estimates. This effect is prominent in various contexts, particularly in quantitative estimations like real estate pricing, where initial listing prices affect subsequent estimates [28]. It also affects qualitative judgments, such as sentencing decisions in judicial settings, evidenced by studies that show significant variations based on initial sentencing demands [29].

Cognitive biases that arise from the anchor effect in AI-assisted decision-making stem from direct and indirect anchoring processes. An immediate anchor is the AI's suggestion, influencing decisions by guiding towards similar options and potentially neglecting other factors. This can yield varied

outcomes. If the AI system surpasses human capabilities, it improves decision accuracy [30]. In contrast, reliance on less accurate AI recommendations can lead to overreliance [31]. Additionally, when AI recommendations come after humans initiate decision-making, the original human estimate can act as an anchor. Two possibilities emerge: If the AI suggestion aligns with the initial estimate, confirmation bias may prompt immediate adoption, as previously discussed. On the contrary, if the AI suggestion differs, individuals tend to stick to their original estimate and may not rely on the system.

Anchoring bias may manifest itself indirectly in situations involving ordering and framing effects. For example, when individuals receive accuracy information about an AI assistant, it can act as an anchor, reducing trust compared to scenarios without disclosure [32]. Additionally, in repeated use of AI assistants, users may initially perceive high accuracy, leading to inflated trust and anchoring future assessments to this impression, increasing reliance on the system [33]. The opposite scenario may also occur.

3.5. Loss Aversion

One notable human behavioral trait is loss aversion, where losses hold more weight than equivalent gains [17]. This bias can extend to AI-assisted decision-making. Humans may focus more on false positives than false negatives in AI errors [15], leading to algorithm aversion bias and under-reliance. Additionally, in risky decision tasks, individuals tend to trust their beliefs over AI, contributing to a lack of trust, which is challenging to mitigate [31].

3.6. Availability Bias

Availability bias leads to an overestimation of event frequencies based on easily recalled instances [17]. In human decision-making with AI recommendations, users can incorrectly estimate the frequency of AI suggestions due to memory recall [11], affecting AI reliance. Wang et al. propose presenting base frequencies to mitigate this bias [11]. Furthermore, explanations can also induce availability bias if users recall relevant knowledge. Users may perceive explanations as more or less plausible based on recalled knowledge, potentially reinforcing biased perceptions [19].

3.7. The Effects of Cognitive Bias

The cognitive biases that arise and their effects can vary depending on the characteristics of the decision-making task. Within the AI-assisted decision-making framework described in section 2, several biases may occur in relation to the three components. For example, an individual with limited knowledge of AI but high expertise in the relevant field may exhibit algorithm aversion, leading to a lack of trust in AI recommendations [24]. Conversely, a lack of experience on the part of the human decision-maker can result in automation bias. In group decision-making scenarios, there is a tendency toward groupthink—a bias where individuals conform to the majority opinion, potentially increasing overreliance on the AI system [34, 35].

The task's characteristics also play a significant role. High-stakes decisions are more prone to loss aversion bias, which may lead to under-reliance on the AI system [15, 31]. Conversely, highly complex tasks may result in automation bias, as the human decision-maker might rely more on the AI system due to the task's difficulty [22]. Even task design can influence the decision-making process and the emergence of specific biases. For instance, if the AI recommendation is presented at the outset, alongside the collection of all relevant information, it could trigger anchoring bias, where the AI recommendation serves as an anchor [30]. However, a cognitive forcing function that delays showing the recommendation until after a certain period could lead to confirmation bias, where the human decision-maker has already formed an opinion, and the AI recommendation merely reinforces this decision, reducing critical thinking [18].

Finally, regarding the AI component, a high cognitive load required to interpret the explanations—or even just the presence of explanations—might induce automation bias in the human decision-maker [22].

Each decision-making task is defined by the unique characteristics of its components—human, task, and AI—and the interplay of these factors results in different cognitive biases for each task. Therefore, it is crucial to analyze various cognitive forcing function designs and explanations in each scenario. This analysis should identify not only the biases that need to be mitigated but also those that could potentially be introduced by the explanations or the new design.

4. Conclusion

The paper focuses on common cognitive biases in AI-assisted decision-making rather than covering all possible biases. Techniques such as XAI and cognitive forcing functions can help address some of these biases, but they can also unintentionally introduce new ones. For instance, explanations can trigger the mere exposure effect, leading to overreliance [15, 14]. Additionally, complex explanations that aim for completeness [13] or present arguments for and against each option [12] can induce automation bias due to their high cognitive demands [22].

Cognitive forcing functions are designed to enhance user engagement in AI-assisted decision-making. These kinds of techniques can also trigger cognitive biases similar to those caused by explanations. For example, introducing AI suggestions after the user's initial decision can lead to anchoring effects or algorithm aversion [30]. Moreover, if not carefully designed, these functions can inadvertently reduce engagement by making the decision process overly complex. In conclusion, while these techniques offer valuable solutions, they also present challenges, requiring a nuanced approach to effectively manage potential cognitive biases for each case of AI-assisted decision-making task.

Acknowledgments This research was funded by INESC-ID (UIDB/50021/2020), as well as the projects CRAI C628696807-00454142 (IAPMEI/PRR) and TAILOR H2020-ICT-48-2020/952215 and HumanE AI Network H2020-ICT-48-2020/952026.

References

[1] J. Christian, Regulators alarmed by doctors already using ai to diagnose patients, 2023. URL: https://futurism.com/neoscope/doctors-using-ai.

[2] V. Chen, Q. V. Liao, J. Wortman Vaughan, G. Bansal, Understanding the role of human intuition on reliance in human-ai decision-making with explanations, Proceedings of the ACM on Human-computer Interaction 7 (2023) 1–32.

[3] P. Hemmer, M. Schemmer, M. Vössing, N. Kühl, Human-ai complementarity in hybrid intelligence systems: A structured literature review., PACIS (2021) 78.

[4] Z. Li, Z. Lu, M. Yin, Decoding ai's nudge: A unified framework to predict human behavior in ai-assisted decision making, arXiv preprint arXiv:2401.05840 (2024).

[5] V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, C. Tan, Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1369–1385.

[6] S. S. Kim, E. A. Watkins, O. Russakovsky, R. Fong, A. Monroy-Hernández, " help me help the ai": Understanding how explainability can support human-ai interaction, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–17.

[7] M. Schemmer, N. Kuehl, C. Benz, A. Bartos, G. Satzger, Appropriate reliance on ai advice: Conceptualization and the effect of explanations, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 410–422.

[8] M. Schemmer, P. Hemmer, M. Nitsche, N. Kühl, M. Vössing, A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 617–626.

[9] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, Does the whole exceed

its parts? the effect of ai explanations on complementary team performance, in: Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–16.
[10] M. Eiband, D. Buschek, A. Kremer, H. Hussmann, The impact of placebic explanations on trust in intelligent systems, in: Extended abstracts of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–6.
[11] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing theory-driven user-centric explainable ai, in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–15.
[12] T. Miller, Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 333–342.
[13] A. Jacovi, J. Bastings, S. Gehrmann, Y. Goldberg, K. Filippova, Diagnosing ai explanation methods with folk concepts of behavior, Journal of Artificial Intelligence Research 78 (2023) 459–489.
[14] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–21.
[15] A. Bertrand, R. Belloum, J. R. Eagan, W. Maxwell, How cognitive biases affect xai-assisted decision-making: A systematic review, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 78–91.
[16] R. R. Hoffman, J. F. Yates, Decision making [human-centered computing], IEEE Intelligent Systems 20 (2005) 76–83.
[17] D. Kahneman, Thinking, Fast and Slow, Farrar, Straus Giroux, NY, 2011.
[18] M. Szymanski, M. Millecamp, K. Verbert, Visual, textual or hybrid: the effect of user expertise on different explanations, in: 26th international conference on intelligent user interfaces, 2021, pp. 109–119.
[19] T. Kliegr, Š. Bahník, J. Fürnkranz, A review of possible effects of cognitive biases on interpretation of rule-based machine learning models, Artificial Intelligence 295 (2021) 103458.
[20] K. Goddard, A. Roudsari, J. C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, Journal of the American Medical Informatics Association 19 (2012) 121–127.
[21] D. Lyell, E. Coiera, Automation bias and verification complexity: a systematic review, Journal of the American Medical Informatics Association 24 (2017) 423–431.
[22] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, R. Krishna, Explanations can reduce overreliance on ai systems during decision-making, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–38.
[23] K. Z. Gajos, L. Mamykina, Do people engage cognitively with ai? impact of ai assistance on incidental learning, in: 27th international conference on intelligent user interfaces, 2022, pp. 794–806.
[24] E. Jussupow, I. Benbasat, A. Heinzl, Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion (2020).
[25] M. Yeomans, A. Shah, S. Mullainathan, J. Kleinberg, Making sense of recommendations, Journal of Behavioral Decision Making 32 (2019) 403–414.
[26] J. Jauernig, M. Uhl, G. Walkowitz, People prefer moral discretion to algorithms: Algorithm aversion beyond intransparency, Philosophy & Technology 35 (2022) 2.
[27] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty., science 185 (1974) 1124–1131.
[28] G. B. Northcraft, M. A. Neale, Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions, Organizational behavior and human decision processes 39 (1987) 84–97.
[29] B. Englich, T. Mussweiler, F. Strack, Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making, Personality and Social Psychology Bulletin 32 (2006) 188–200.
[30] F. Cabitza, A. Campagner, L. Ronzio, M. Cameli, G. E. Mandoli, M. C. Pastore, L. M. Sconfienza,

D. Folgado, M. Barandas, H. Gamboa, Rams, hounds and white boxes: Investigating human–ai collaboration protocols in medical diagnosis, Artificial Intelligence in Medicine 138 (2023) 102506.
[31] R. de Brito Duarte, F. Correia, P. Arriaga, A. Paiva, et al., Ai trust: Can explainable ai enhance warranted trust?, Human Behavior and Emerging Technologies 2023 (2023).
[32] T. Kim, H. Song, The effect of message framing and timing on the acceptance of artificial intelligence's suggestion, in: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–8.
[33] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, V. Gogate, Anchoring bias affects mental model formation and user reliance in explainable ai systems, in: 26th International Conference on Intelligent User Interfaces, 2021, pp. 340–350.
[34] C.-W. Chiang, Z. Lu, Z. Li, M. Yin, Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–18.
[35] I. L. Janis, Groupthink, IEEE Engineering Management Review 36 (2008) 36.