Explainable AI as a Crucial Factor for Improving Human-AI Decision-Making Processes

Regina de Brito Duarte

INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Abstract

A crucial aspect of AI-assisted decision making involves providing explanations for AI recommendations and predictions. Despite the optimism surrounding eXplainable AI (XAI) to improve transparency and trustworthiness, several studies have highlighted its shortcomings. My doctoral research aims to develop and validate a framework for human-AI decision making where explanations are central, serving as an enhancement factor for AI-assisted decision tasks. I hypothesize that a robust framework will elucidate underlying mechanisms and investigate the effects of AI explanations on decision outcomes. This research will advance our understanding of the combination of AI and human capabilities, informing the design of AI-assisted decision tasks for real-world scenarios.

Keywords

Human-AI Decision Making, eXplainable AI, Human-AI Interaction

1. Introduction

The development of Artificial Intelligence (AI) algorithms and their astonishing capabilities are the primary reasons for the rapid adoption of AI in our society. Specifically, in the realm of human decision-making, some decisions that were once solely the domain of humans are now being made with the assistance of AI decision support systems. Examples can be observed in courtrooms [1] or in healthcare, where regulators are concerned about the consequences of doctors using AI tools to support their clinical practice [2]. The full extent of the impacts of AI recommendations on human decision-making is not yet entirely understood, and there is currently a significant effort within the scientific community to comprehend its effects [3, 4, 5].

One prominent factor in the decision-making process assisted by AI algorithms is the provision of explanations for such AI recommendations and predictions. The field of eXplainable AI (XAI) has flourished due to the necessity of understanding why AI algorithms make specific predictions. Now, its utility is considered crucial to improve AI transparency and trustworthiness among human users [6, 7], especially in the context of human-AI decision making. Even with considerable optimism about XAI's potential to enhance transparency and human-AI interaction in decision-making, various studies underscore the limitations and risks tied to XAI explanations in AI-assisted decisions [8, 9, 10]. A notable issue is the overreliance paradigm, where humans depend heavily on AI algorithms, and the explanations do not improve trust

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

reginaduarte@tecnico.ulisboa.pt (R. d. B. Duarte)

0000-0003-0249-8319 (R. d. B. Duarte)

calibration, at times worsening overreliance patterns [11, 12, 13]. Some research indicates that explanations emphasizing feature importance might trigger over-reliance [3, 14], whereas certain explanations may not significantly impact decision performance and AI trust [15, 10]. Furthermore, explanations can have a placebo effect, where individuals trust the AI system more simply due to the presence of an explanation, regardless of its correctness or the trustworthiness of the AI system [8, 14].

There is an ongoing debate regarding the best types of explanations to present, considering multiple options and incorporating social theories of human explanation [16, 17, 18]. Two examples of new strategies to improve XAI in AI-assisted decision-making include providing a detailed narrative for full comprehension of the explanation and the context of the decision task [18] and offering a discussion of each option's pros and cons rather than explaining the reasoning behind an AI prediction [16]. Another possible enhancement involves cognitive forcing functions, which are design techniques meant to encourage active participation in the decision-making process [11, 19, 15]. This method forces decision-makers to engage with explanations, addressing the limitations of merely presenting them.

In parallel, several other factors may affect the AI-assisted decision-making process, not only influencing the decision itself but also impacting how explanations affect the final output. These factors include the difficulty of the task, the expertise of the human decision-maker, the stakes of the decision, and even the cognitive load required to understand the task [19, 13, 10]. Nevertheless, there is a clear lack of standardized metrics and structures needed to formulate a unified theory of human-AI decision making and compile all research findings [20]. The current research evidence in the literature is challenging to compare directly due to varying decision-making factors and contexts, making it difficult to discern patterns in AI-assisted decision making. Additionally, the methods and metrics applied to assess and comprehend these tasks vary widely and are difficult to align in the absence of standardization. For instance, it is not sufficiently clear when one should employ feature importance explanations, counterfactual explanations, or example-based explanations for a decision-making task with specific characteristics.

The purpose of my doctoral research is to add to the body of literature on AI-assisted decision-making by crafting a theory of Human-AI decision-making. This theory will lay the groundwork for understanding the decision-making process when assisted by an AI agent. My research has two primary goals. First, I intend to investigate the factors that could affect human-AI decision-making. While numerous studies have examined the impacts of various elements in AI-assisted decision-making, certain factors such as team composition and AI system embodiment have yet to be fully explored, especially regarding the influence of explanations in AI-aided decision-making. Second, I aim to develop and validate a framework to comprehend human-AI decision-making. For this objective, I will utilize both existing studies and my own research, which will introduce new insights and consider additional factors, as foundational inputs for the framework's development. In the subsequent section, I will elaborate on my research questions and targets.

2. Research Questions and Hypotheses

The research questions I aim to respond in my doctoral research are the following:

RQ.1 What constitutes the optimal framework for conceptualizing AI assisted decision-making processes?

RQ.2 How can efficiency and efficacy be promoted in AI-assisted decision-making processes?

RQ.3 How can explainable AI methods be enhanced to support AI assisted decision-making process?

The end goals of my research questions are twofold: Firstly, to investigate XAI explanations, both from technical and design perspectives, as a central element that can enhance the human-AI decision-making process. Secondly, to establish a validated framework for the analysis of AI-assisted decision-making tasks. For the first research question, I hypothesize the existence of a framework for studying AI-assisted decision-making processes that facilitates understanding of the underlying mechanisms involved. Illustrated in Figure 1 is a simplified framework for an AI-assisted decision process for a single decision task, comprising three main components: the human decider, responsible for making the decision and being accountable for its consequences; the decision task itself; and the AI recommendation provided to support the decision-making process.

Each component can exhibit various characteristics that influence the decision-making process. For instance, the decision task may vary in difficulty (task difficulty), operate within a high-stakes domain (risk), and require differing levels of cognitive effort from the user to comprehend and make the decision (cognitive load). Similarly, on the human decider side, factors such as expertise level and the presence of a group of decision-makers rather than an individual (number of deciders) can impact the process. This framework establishes the context in which the decision task is inserted and can inform the effects of the AI recomendations and explanations depending on the context. For instance, presenting counterfactual explanations to lay users might not be as effective as presenting counterfactual explanations to experts users in the task [14]. Hence, this framework is important to establish in what contexts the explanations are useful and what types of explanations are best, depending on the factors that characterize the decision-making process.

In AI-assisted decision-making processes, it is not enough to concentrate solely on task performance metrics, like efficiency, and fairness. It is crucial to also consider the human-AI relationship, including whether humans appropriately trust the AI and understand its recommendations, as these elements greatly affect task performance. By examining these three components and the factors influencing task performance, we can further develop the framework to comprehend the functioning of AI-assisted decision-making and the dynamics among the contributing factors.

This framework requires improvement in two ways: First, by incorporating a comprehensive list of factors that affect AI-assisted decision-making processes, and second, by understanding the effect of each factor on the final decision in terms of AI reliance, trust, efficiency, and efficacy. With a well-developed framework for AI-assisted decision-making addressed in RQ1, one can respond to our second RQ with a clearer mental model of the factors influencing the decision and how. Finally, concerning our third RQ, XAI explanations play a central role in

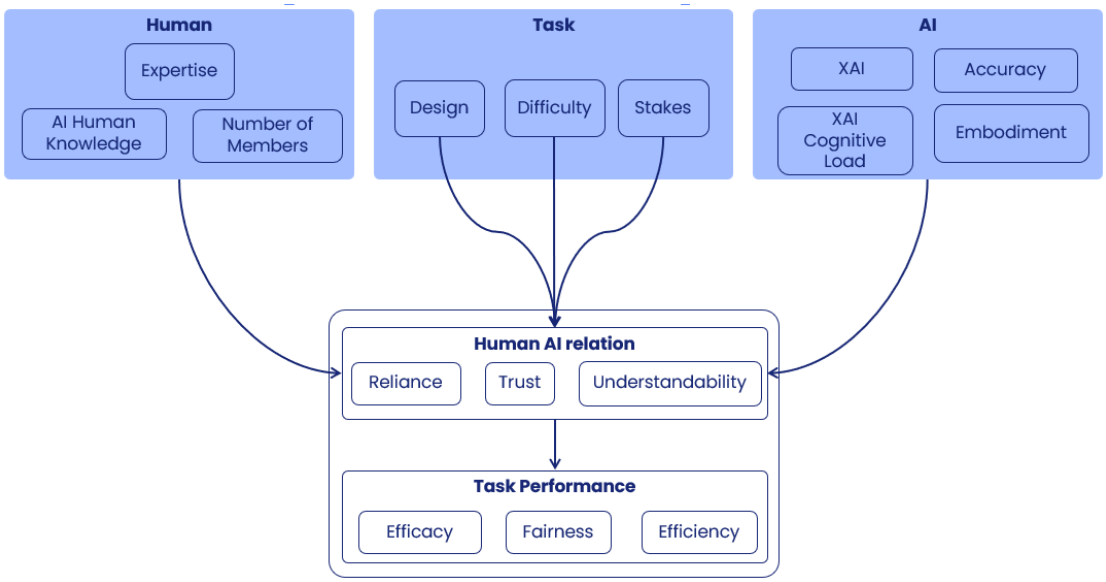

Figure 1: Preliminary framework of an AI-supported decision-making process consisting of three key elements: the human decision-maker, the decision task, and the AI recommendation/prediction. These three components collectively influence decision-making effectiveness and define the interaction between humans and AI.

the enhancement of AI-assisted decision-making process when all the factors are known. My hypothesis is that we can optimize XAI techniques and designs to enhance decision output and mitigate errors.

3. Methodology and Preliminary Results

To address my research questions, I will employ two primary methodologies. Initially, I will conduct multiple controlled user studies to explore various factors that may influence the AI-assisted decision-making process. These studies will serve two purposes. Firstly, they will aid in comprehending the underlying factors that impact AI-assisted decision-making, thus informing the development and posterior validation of a framework for AI-assisted decision-making. Second, these studies will consistently emphasize explanations as a predominant factor in understanding how they can mitigate errors and interact with other variables. These studies will encompass a wide range of decision-making contexts to thoroughly examine various factors of the AI-assisted decision-making framework. One set of studies will focus on decision tasks requiring low human expertise, while another will target tasks that demand high levels of expertise. Within each set, a baseline experiment will be conducted, followed by additional experiments that explore different factors, such as stress, AI embodiment, team composition, risk, and AI trustworthiness. Throughout all experiments, the type of explanations provided will be a factor to manipulate, but the presence of explanations will remain a constant factor. Finally, longitudinal studies will be conducted to understand the impact of time and repeated interaction on the human-AI decision-making process.

Finally, to construct a robust and validated framework for AI-assisted decision-making, I will conduct a comprehensive review of existing literature that explores this domain. By synthesizing insights and findings from these studies, I will develop an initial framework that integrates the key concepts and factors identified. This process will be iterative, involving continuous refinement of the framework through the dual approach of literature review and empirical validation via controlled studies. The goal is to ensure that the resulting framework is not only theoretically sound but also practically applicable across diverse decision-making contexts.

Up to this point, two controlled studies have been conducted. Both studies were designed as a mushrooms' edibility classification task with low expertise of the decision maker. The initial study explored the impacts of various types of explanation (counterfactual versus feature importance) on AI trust in varying risk settings and AI trustworthiness. The findings indicated that feature importance explanations enhance AI trust, while counterfactual explanations have no discernible effect on AI trust. This finding can be explained by the low expertise of the human deciders that did not have enough expert knowledge about mushrooms to understand properly counterfactual explanations. Moreover, in scenarios where AI trustworthiness is low, feature importance explanations can induce overtrust as the presence of the explanation by itself can have a priming effect on convincing the participant that the recommendation is trustworthy [14, 8]. The second study investigated the effects between the presence of explanations and the number of individuals making decisions. Preliminary results suggest that the impact of XAI is more pronounced in decision-making with groups of two users than in individual decision-making. Groups rely less on incorrect AI recommendations when explanations are available, but paradoxically, they rely more on incorrect AI recommendations when explanations are absent, compared to individual decision makers.

4. Future Work and Conclusion

The primary objective of my doctoral research is to establish a robust framework for AI-assisted decision-making tasks, which can significantly contribute to the field and serve as the foundation for a novel theory on human-AI decision-making processes where explanations are at the heart to enhance efficacy and AI trust. To achieve this goal, in future work, I plan to conduct additional controlled experiments, exploring novel types of explanations and investigating other factors in AI-assisted decision-making. These factors include variations in cognitive load associated with both the task and the explanations provided, diverse levels of expertise among users, and the integration of AI embodiment into decision-making scenarios. Additionally, refinement and validation of the AI-assisted decision-making framework will be pursued through several case studies drawn from existing literature. The final outcome of this research—a validated framework for AI-assisted decision-making—will enhance our understanding of combining AI and human capabilities, guiding the design of new human-AI tasks and AI explanations in real-world scenarios.

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] H. Margetts, C. Dorobantu, Rethink government with ai, Nature 568 (2019) 163–165.
[2] J. Christian, Regulators alarmed by doctors already using ai to diagnose patients, 2023. URL: https://futurism.com/neoscope/doctors-using-ai.
[3] 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.
[4] P. Hemmer, M. Schemmer, M. Vössing, N. Kühl, Human-ai complementarity in hybrid intelligence systems: A structured literature review., PACIS (2021) 78.
[5] 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).
[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. 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.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, R. Kr-

ishna, Explanations can reduce overreliance on ai systems during decision-making, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–38.
[14] 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).
[15] 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.
[16] 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.
[17] A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 624–635.
[18] 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.
[19] 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.
[20] 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.