Comment les biais cognitifs affectent la prise de décision assistée par l'IA explicable

Rafik Belloum, Astrid Bertrand, James R. Eagan, Winston Maxwell

Comment les biais cognitifs affectent la prise de décision assistée par l'IA explicable

Rafik Belloum*, Astrid Bertrand**,
James R. Eagan**, Winston Maxwell***

*Univ. Polytechnique Hauts-de-France LAMIH, CNRS, UMR 8201 F-59313 Valenciennes, France
rafik.belloum@uphf.fr
**LTCI, Institut Polytechnique de Paris France
astrid.bertrand@telecom-paris.fr
james.eagan@telecom-paris.fr
***i3, CNRS, Institut Polytechnique de Paris France
winston.maxwell@telecom-paris.fr

Résumé. Ce papier résume une revue de la littérature sur les biais cognitifs influençant la prise de décision assistée par l'IA explicable (XAI). Il va au-delà de la simple identification des biais cognitifs en XAI, offrant une vision stratégique, illustrée par une carte heuristique qui guide le futur développement de systèmes XAI plus en phase avec les processus cognitifs humains. Il convient de noter que ce résumé synthétise un article déjà publié par Bertrand et al. (2022).

1 Introduction

Le domaine de l'Intelligence Artificielle Explicable vise à apporter de la transparence aux systèmes d'IA complexes. Bien qu'il soit généralement considéré comme un domaine essentiellement technique, des efforts ont récemment été déployés pour mieux comprendre les méthodes d'explication humaine des utilisateurs et les contraintes cognitives. Malgré ces avancées, la communauté manque d'une vision générale de la manière dont les biais cognitifs affectent les systèmes d'explicabilité. Cet article, déjà paru et que nous souhaitons résumer ici, comble cette lacune en présentant une cartographie heuristique novatrice, alignant les biais cognitifs humains avec les techniques d'explicabilité issues de la littérature XAI, et structurée autour de la prise de décision assistée par la XAI.

2 Cartographie des Biais Cognitifs en XAI

L'article propose une cartographie de biais cognitifs identifiés dans la littérature XAI, offrant une vue détaillée de leur présence et de leur impact. L'utilisation du guide PRISMA (Moher et al., 2009) pour la revue de la litterature garantit la rigueur méthodologique. Ces biais sont catégorisés en fonction de différents contextes, tels que le type d'explicabilité utilisé, le domaine d'application, la tâche assistée par l'IA et le type d'utilisateur (expert du domaine, expert en IA ou utilisateur lambda).

Biais cognitifs affectant la conception des méthodes XAI. Les résultats du papier soulignent comment certains biais cognitifs influent sur la conception des méthodes XAI. Ces biais (Miller, 2019), présentés dans des boîtes jaunes sur la carte heuristique de la figure 1, sont liés aux heuristiques explicatives que les individus utilisent lors de l'explication ou de la réception d'une explication. Contrairement à d'autres biais, ces heuristiques ne sont pas considérées comme des "erreurs", mais plutôt comme des contraintes à prendre en compte lors de la conception des techniques d'explicabilité.

Biais cognitifs impactant l'évaluation des techniques XAI dans les études utilisateur. Une autre catégorie de biais cognitifs, présentée dans une boîte marron sur la carte heuristique de la figure 1, concerne la distorsion potentielle dans l'évaluation des techniques XAI lors d'études utilisateur. Cela découle des préoccupations croissantes quant à la nécessité de tester les explications avec les utilisateurs. Certains chercheurs insistent sur cette approche, tandis que d'autres la déconseillent, craignant que les biais cognitifs ne faussent les évaluations et trompent le domaine XAI (Herman, 2017).

Atténuation et Exacerbation par les Techniques XAI. L'article identifie également des biais cognitifs qui peuvent être atténués avec succès par les techniques XAI (couleur orange sur la figure 1) (Wang et al., 2019; Bertrand et al., 2023), tout en soulignant que certaines méthodes peuvent également exacerber certains biais (couleur rouge). Cette distinction est cruciale pour guider le développement de futures techniques XAI qui minimisent les effets négatifs sur la prise de décision humaine.

Références

Bertrand, A., R. Belloum, J. R. Eagan, et W. 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*, pp. 78–91.

Bertrand, A., T. Viard, R. Belloum, J. R. Eagan, et W. Maxwell (2023). On selective, mutable and dialogic xai : a review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21.

Herman, B. (2017). The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv :1711.07414*.

Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial intelligence 267*, 1–38.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman, et P. Group* (2009). Preferred reporting items for systematic reviews and meta-analyses : the prisma statement. *Annals of internal medicine 151*(4), 264–269.

Wang, D., Q. Yang, A. Abdul, et B. Y. Lim (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15.

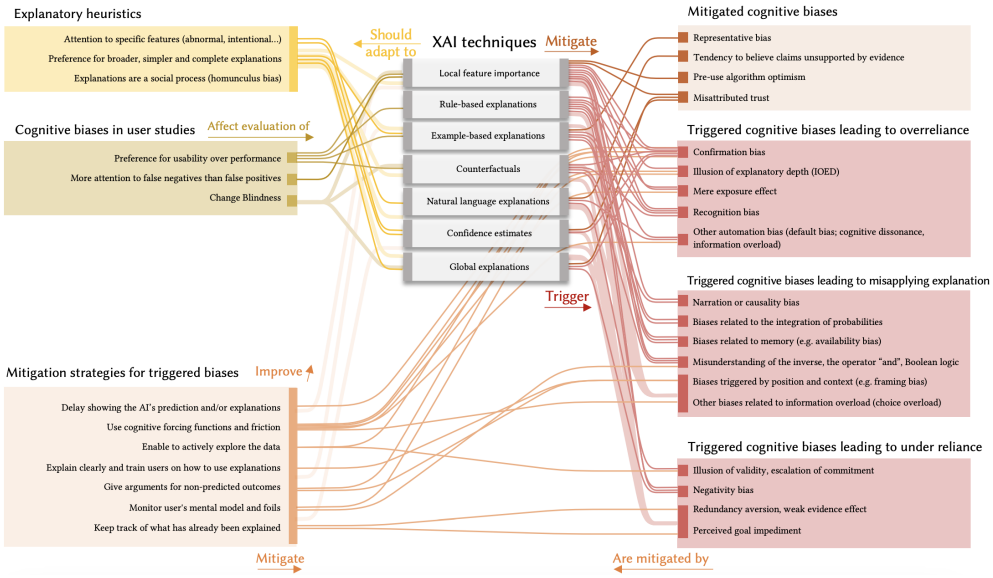

FIG. 1 – *Résumé des contraintes cognitives, des biais et des stratégies d'atténuation identifiés dans les articles inclus dans le corpus (n=37). Ce diagramme présente les différentes catégories de techniques d'explication observées dans le corpus (au centre). Chaque lien représente une connexion établie dans la littérature entre une technique d'explicabilité et un biais cognitif, ou entre un biais cognitif et une technique d'atténuation. Les légendes en couleur soulignées par des flèches indiquent comment et dans quelle direction les liens doivent être lus (par exemple, "Les techniques d'XAI devraient s'adapter aux heuristiques explicatives"). Les liens pâles et larges indiquent que le biais ou la contrainte cognitive s'applique de manière plus générale à toutes les méthodes d'XAI. Il a été identifié davantage de connexions entre les biais et les stratégies d'atténuation, mais seules les plus soutenues sont présentées pour des raisons de concision.*

Summary

This paper summarizes a literature review on cognitive biases influencing XAI-assisted decision-making. It goes beyond mere identification of cognitive biases in XAI, providing a heuristic map, guiding the future development of XAI systems that are more attuned to human cognitive processes. This map contributes to the evolution of the XAI field by emphasizing alignment with how individuals comprehend and use explanations provided by AI systems.