Explanations considered harmful: The Impact of misleading Explanations on Accuracy in hybrid human-AI decision making

Federico Cabitza[1,2][0000−0002−4065−3415], Caterina Fregosi[1][1111−2222−3333−4444], Andrea Campagner[2][0000−0002−0027−5157], and Chiara Natali[1][0000−0002−5171−5239]

[1] Università degli Studi di Milano-Bicocca, Milan, Italy
[2] IRCCS Ospedale Galeazzi-Sant'Ambrogio, Milan, Italy
federico.cabitza@unimib.it

Abstract. EXplainable AI (XAI) has the potential to enhance decision-making in human-AI collaborations, yet existing research indicates that explanations can also lead to undue reliance on AI recommendations, a dilemma often referred to as the 'white box paradox.' This paradox illustrates how persuasive explanations for incorrect advice might foster inappropriate trust in AI systems. Our study extends beyond the traditional scope of the white box paradox by proposing a framework for examining explanation inadequacy. We specifically investigate how accurate AI advice, when paired with misleading explanations, affects decision-making in logic puzzle tasks. Our findings introduce the concept of the 'XAI halo effect,' where participants were influenced by the misleading explanations to the extent that they did not verify the correctness of the advice, despite its accuracy. This effect reveals a nuanced challenge in XAI, where even correct advice can lead to misjudgment if the accompanying explanations are not coherent and contextually relevant. The study highlights the critical need for explanations to be both accurate and relevant, especially in contexts where decision accuracy is paramount. This calls into question the use of explanations in situations where their potential to mislead outweighs their transparency or educational value.

Keywords: Explainable artificial Intelligence (XAI) · Human-AI Interaction · Explainability paradox.

1 Introduction

Ideally, an effective explanation within a decision support context should empower its user — the human decision maker to whom this explanation is provided, alongside the machine's advice — to understand the rationale behind the system's suggestion. This understanding enables the user to judiciously determine whether the advice is appropriate and should be followed, or if it is flawed and should be disregarded. This capability, a form of controllability [19], underscores the belief that support from Explainable AI (XAI), through providing comprehensible explanations for its suggestions, enhances the user's experience across several dimensions of interaction with a decision support system. For instance, it can increase the user's confidence in the final decision, leading to greater satisfaction with the interaction [27]; foster appropriate reliance [2, 5] — enabling the user to trust the system when appropriate and be skeptical when the system might mislead — and thereby calibrate their trust [20, 1] in the system. The potential for explanations to positively impact the decision-making process, and the optimism that such an impact is achievable, justify the growing interest in this research area [21] and the development and evaluation of applications it facilitates.

However, issues arise with the acknowledgment of the very real possibility that not all explanations sufficiently clarify the advice to the user in a manner that is understandable [30] and, more critically, that inadequate explanations are not always necessarily indicative of incorrect advice.

To motivate the possibility of these issues arising, we note that the functional modules generating the advice and the associated explanations typically operate independently from each other, though they communicate [14]; they are often based on distinct data analysis technologies, especially given that the many XAI methods rely on linear models to generate their output [25]. Consequently, on the one hand, the mental model a user may construct from the explanations might not effectively mirror the advice-generating module's logic, while on the other hand there could be potentially misleading mismatches between the outputs of the two models [31]: crucially, both of these issues can potentially impede the development of a well-calibrated trust in the system [24].

Thus, two scenarios are of critical importance for understanding how to mitigate the above mentioned risks: first, when a plausible explanation wrongly convinces the user to follow incorrect advice, a phenomenon previously discussed in literature under the expression "white-box paradox," [10, 11]; and second, the less explored scenario where an implausible explanation leads to the dismissal of correct advice. These scenarios together form a conundrum we refer to as the "Explainability Paradox," where explanations can mislead the user and compromise their appropriate reliance on the system. This paper presents a study that explores the less investigated aspect of this varied and still little known phenomenon, that is the effect of misleading explanations coupled with otherwise correct advice. In particular, through a user study in the setting of AI-assisted logical-mathematical reasoning tests, we aimed at addressing the following research questions:

RQ1: Does a correct AI advice coupled with a misleading explanation affect user accuracy? In particular, would misleading explanations induce users in error? And is there any user strata that is more susceptible to this effect?

RQ2: Does a correct AI advice coupled with a misleading explanation affect user confidence in their response?

2 Background and Related Work

An explanation can be defined as an answer to a “why” question, incorporating both causal attribution and context [8]. Explainable Artificial Intelligence (XAI), a term introduced by Van Lent et al. [28], represents the hope to address the limitations of current AI systems, intended as “black boxes”, by enhancing trust and transparency through the provision of explanations [17, 22]. Research has indeed demonstrated the positive impacts of explanations on human-AI interaction, suggesting that XAI support can reduce the opacity of AI systems by making them more comprehensible [3, 29]. Comprehension and knowledge are precursors to trust; explanations not only facilitate the simulation of model predictions but also significantly increase user trust in these systems [2]. Evidence strongly confirms and supports the utility of XAI in decision-making processes, highlighting its contribution to improved accuracy [10, 2, 3]. However, recent developments have revealed potential challenges in scenarios involving hybrid (i.e., human-XAI) decision-making. In addition to causing the ’white box paradox’ [11, 9] mentioned above, explanations can result in reasoning errors such as backward reasoning and confirmation bias [3]. Moreover, contrary to initial beliefs that explanations might reduce system over-reliance, findings indicate they might actually increase it [4], leading users to rely too much on incorrect AI recommendations [26].

Human-XAI interactions are further complicated by the fact that current methods for explanation do not offer guarantees of either connection to the system advice nor causality, indicating that they can be wrong. This adds an additional potential error source for users [3]. To understand how to mitigate detrimental effects and develop functional interaction protocols, it is therefore crucial to thoroughly investigate both ’wrong’ and ’poor’ explanations. Some researchers have started to explore the effects of “poor” explanations. Eiband et al. conducted an experiment introducing explanations phrased to semantically insert a justification without delivering pertinent information. Their findings revealed that even these ‘placebic explanations’ induced a trust level comparable to real explanations [16]. Morrison et al. introduced the concept of imperfect explainable AI systems, defined as explanation techniques that can potentially generate explanations that do not fit with the AI’s predictions. Their research indicates that such imperfect explanations can foster inappropriate reliance on AI, affecting user performance, especially among non-expert. [23] Furthermore, Eberman et al. investigated the user impact of a contradiction between the decisions and the corresponding explanations, discovering that a lack of alignment to the advice leads users to experience negative mood and have a negative evaluation of the AI system’s support [15]. Despite efforts in the literature to investigate ‘poor’ explanations and their effects on users, a unified, comprehensive definition of this phenomenon remains absent. Our proposal in what follows aims to address this gap. Considering the independence of the functional modules that generate the advice and their associated explanations [14], the following sections will introduce a classification of poor explanations, hereinafter termed '*misleading explanations*', based on two dimensions: their coherence with the corresponding advice and their relevance to the task at hand.

3 How Explanations can be misleading

In the above literature, explanations are usually considered either good or bad without distinguishing what kind of shortcomings they present, for instance in terms of clarity, comprehensiveness, relevance and similar dimensions. We started considering these distinctions in [7], where we introduced two intuitive ways in which explanations can be "wrong": 1) in terms of alignment with respect to the advice given by the machine (what in the that article we called coherence); 2) in terms of relevance to understand how to correctly interpret the case at hand (what in that article we called pertinency). According to this purposely simple framework, explanations can be misleading either because they are incoherent with respect to the AI advice, or because they are not relevant with respect to the case, or for both these reasons.

We call *XAI halo effect* the effect exterted by explanations when they are wrong and make users wary of, and eventually discard, the AI advice even when this is correct. This name is inspired by the fact that the *halo effect* [18] is the cognitive bias whereby the perception of some traits of somebody or something, in our case the accuracy and trustworthiness of AI support, is influenced by the perception of one or more other traits of its, in this case explainability.

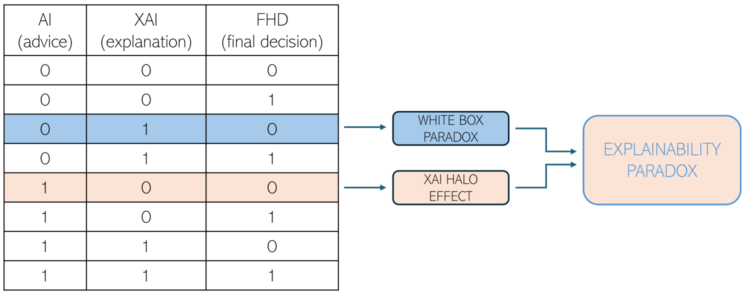
Focusing on this so-far neglected effect allows us to complement the existing literature on the opposite phenomenon: when good explanations can, paradoxically, affect appropriate reliance by making users accept the advice of the system even when this is wrong (the so called "white-box paradox" [10]).

These two detrimental effects can be grouped together by the more general term "explainability paradox", which *is* paradoxical in light of the promise that XAI poses on the potential to improve decision making performance, reliance appropriateness and user satisfaction (see Figure 1).

Although the above framework allows us to define three cases of misleading explanations (see Figure 2), the study that we are going to describe in the next sections focuses on one such case, where explanations are consistent with the given advice but not relevant to the classification task. This framework aims to facilitate a comprehensive understanding of the ways in which explanations can be misleading — not merely whether they can be, but also how — and to assess the impact of each type of deficiency. Ultimately, this knowledge could help to minimize the occurrence of the most consequential shortcomings.

4 Methods

We designed this study to investigate the effects of misleading explanations on user accuracy and confidence in human-XAI decision making. To this aim, we enrolled 22 Master's students from an Artificial Intelligence Master degree. Each of them was given 19 moderate-to-hard logic puzzles to solve similar to those that



AI (advice)	XAI (explanation)	FHD (final decision)
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Fig. 1: Reliance patterns for the case at hand (inspired by [6]). In this study we focus on the *XAI halo effect* (in red), that is when bad explanations induce human decision makers to distrust and reject good AI advice, thus making a mistake. In this figure, FHD refers to the final decision of the respondent

are administered in the psychometric assessment of the General Intelligence Factor. These puzzles (sourced from the website *Youmath*[3]) covered various types, including numerical and alphanumeric logic, deductive reasoning, graphic interpretation, and anagrams. The experiment consisted in a simulated interface of *ChatGPT Plus*, developed via *Figma* [13]. Participants engaged with the puzzles by either uploading an image of the pattern reasoning task or typing out the word logic puzzle. Following this, the simulated GPT provided the answer along with its explanation.

Given the research objective of assessing the performance of individuals in terms of accuracy and investigating the influence that misleading explanations can bring even in the case of right answers, we designed the 19 responses of the simulated AI system so that 13 of them were correct and 6 of them incorrect. This arrangement allows us to study the effect of incorrect explanations within a context where users can actually trust the machine (i.e., it is usually correct in other cases).

Among the 13 scenarios where the AI's advice was accurate, 6 presented misleading explanations, while the remaining 7 had correct (coherent and relevant, according to the framework presented in Section 3) explanations. All the 6 cases where AI advice was wrong had misleading explanations. In Figure 3, we present an example of how we generated plausible explanations that, while consistent with the correct advice, were irrelevant to the classification task and thus misleading. To prevent the onset of negative biases and a poorly calibrated trust in the system, we opportunistically placed 9 of the 13 correct answers in a row at the beginning of the series of responses to estabilish trust.

The experimental session was conducted in-person, within university spaces. Participants were invited to a collaborative session with the XAI system. During this session each logic puzzle was followed by a short text presenting both the system's advice and explanation followed by four answer options from which to

[3] https://www.youmath.it

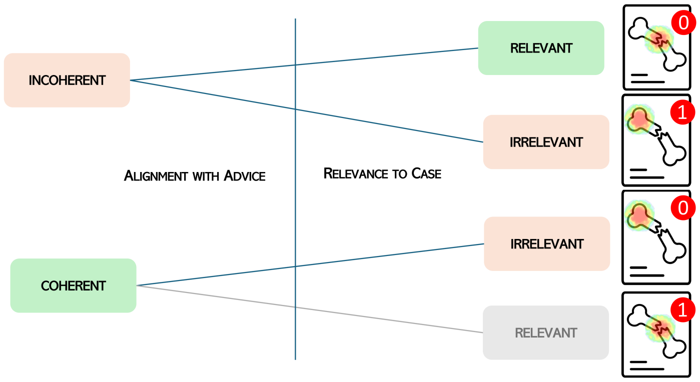
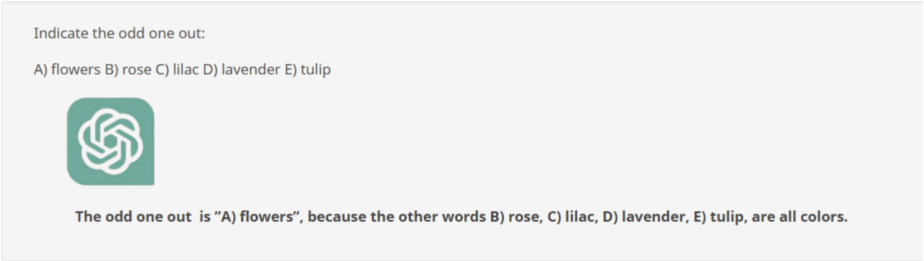


Fig. 2: The proposed framework distinguishes among misleading explanations these three types: incoherent&relevant; incoherent&irrelevant; coherent&irrelevant. In the present study we focus on this latter type of misleading explanations. In the radiological examples on the right, the ground truth is obvious from the icons (fracture / no fracture), the AI advice is the number in red circles (1: fracture; 0: no fracture); whole explanations are rendered in terms of salience maps pointing to the image elements backing up the advice.



Indicate the odd one out:

A) flowers B) rose C) lilac D) lavender E) tulip

The odd one out is "A) flowers", because the other words B) rose, C) lilac, D) lavender, E) tulip, are all colors.

Fig. 3: Example of a misleading explanation presented to participants via the simulated interface. In this case, the provided explanation is consistent with the correct advice (A), yet it misleads by suggesting the word 'flowers' should be discarded because the remaining words denote colors. However, discarding 'flowers' is correct because the other words are indeed all names of flowers.

choose the correct one. A LimeSurvey[4] questionnaire was used to collect their answers and confidence levels on the answers given on a 4-value ordinal scale, a semantic differential item where 1 corresponded to the minimum confidence (not sure at all) and 4 to the maximum confidence (almost certain), to mitigate central tendency bias.

[4] http://limesurvey.org

Statistical analysis were conducted through the hypothesis testing approach adopting a confidence level of .95 and a significance level (α) of .05. Mann–Whitney U tests were conducted on both average error rates and average confidence level due to the, respectively, non-normal and ordinal nature of the data.

5 Results

In what follows, we report the results coming from the quantitative analysis of the responses and scores collected during the experiment, grouped by the research questions presented in Section 1.

5.1 Impact on accuracy

In order to address RQ1, we compared the differences in accuracy on cases corresponding to the two configurations "correct AI and misleading XAI" and "correct AI and correct XAI". To this aim, we applied a non-parametric Mann-Whitney U test (with normal approximation) to compare the average accuracy of the participants' responses on the two groups of cases mentioned above. The p-value was $< .001$: thus, it is significant. Moreover, both the observed standardized effect size and the common language effect size (CLES) were large (0.51 and 0.79, respectively): in particular, the accuracy reported by participants on the "correct AI and correct XAI" cases was significantly higher than that on "correct AI and misleading XAI" ones. See also Figure 4 for a graphical representation of this result.

After finding a significant effect of misleading explanations on the participants' accuracy in "correct AI and misleading XAI" configuration, we decided to stratify the respondents into two categories: considering participants in the extreme (Q1 and Q4) quartiles based on the accuracies, we adopted a distribution-based criterion to define top vs low performers. In order to understand whether the detrimental effect of misleading explanations varied according to the participants' accuracy level, we applied a non-parametric Mann-Whitney U test (with normal approximation) to compare the differences in average accuracy of top-performer and low-performer in the two configurations. The p-value was .04, thus suggesting that the difference was significant. The observed standardized effect size was medium (0.4). This result, consistent with the findings of Morrison et al. [23] wherein the difference was observed between experts and non-experts, indicates that lower-performing individuals are more affected than top-performers by misleading explanations. The effect of the correct and misleading explanations on low- and high-performers is also depicted in Figures 5 and 6, in terms of, respectively, benefit diagrams and Gardner-Altman plots (also called paired plots).

Additionally, we evaluated the impact of the AI, for both correct and misleading explanations, on the reliance patterns of the humans, adopting the approach proposed in [6](see Figure 1): we represent this information in terms of Technology Impact (see Figure 7) and Conservatism Bias (see Figure 8) diagrams, which

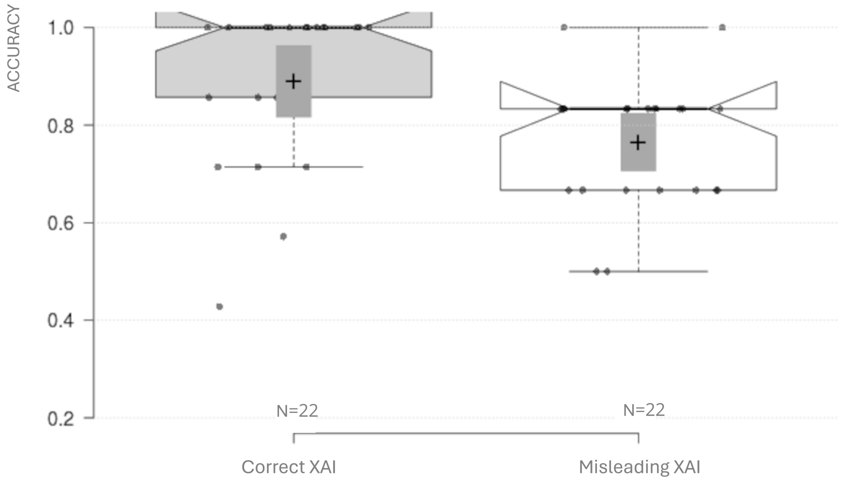


Fig. 4: Box-plots of the accuracy scores of the participants when they solved the good-advice&good-explanation tests (Correct XAI) and when they solved the good-advice&bad-explanation ones (Misleading XAI). Box notches indicate the 95% confidence intervals of the median, while crosses indicate the mean accuracy within its confidence interval (grey rectangle).Confidence scores were defined in a scale from 1 to 4.

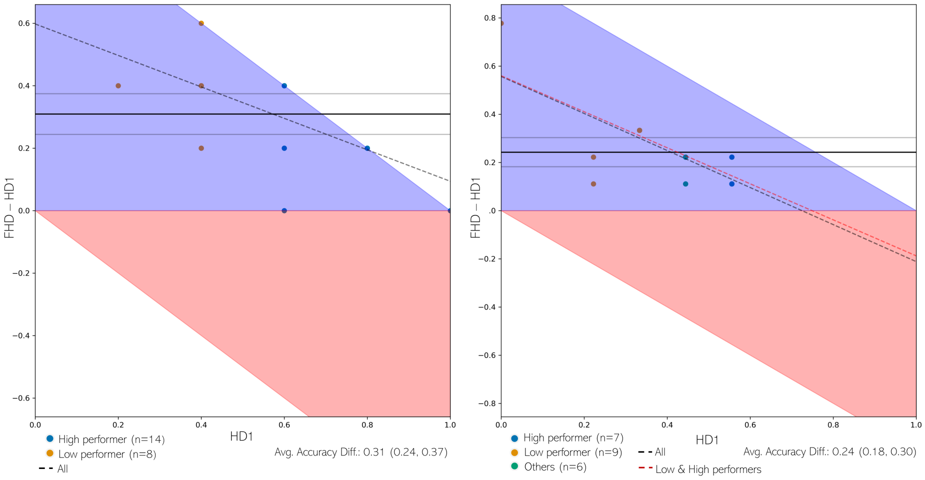


Fig. 5: Benefit diagrams for the "correct AI and correct XAI" (left) and "correct AI and misleading XAI" (right) cases. Generated with the tool available at https://mudilab.github.io/dss-quality-assessment/.

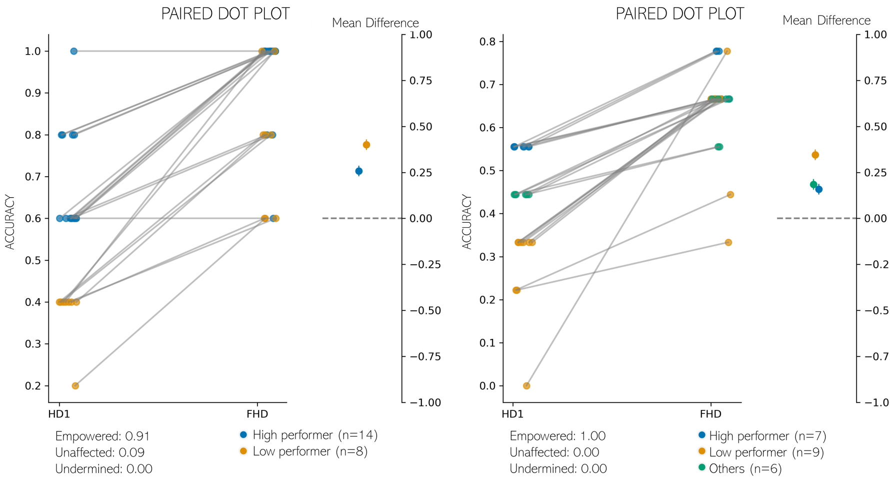


Fig. 6: Paired plots for the ”correct AI and correct XAI” (left) and ”correct AI and misleading XAI” (right) cases. Generated with the tool available at https://mudilab.github.io/dss-quality-assessment/.

depict the impact of the explanations in terms of odds ratios. We did not detect any significant difference in terms of technology impact: in particular, the AI had a significantly positive effect on decision-making, irrespective of the correctness of the explanations, even though the impact was on average more beneficial for the correct explanations. By contrast, the misleading explanations were associated with a significantly larger conservatism bias as compared with correct explanations. Moreover, Figure 8 indicates that receiving misleading explanations causes individuals to remain anchored to their initial decisions compared to receiving correct ones, resulting in a negative effect, though not significantly.

5.2 Impact on confidence

In order to address RQ2, we applied a non-parametric Mann-Whitney U test (with normal approximation) to compare the average confidence of the participants' response in the two configurations “correct AI and misleading XAI” and “correct AI and correct XAI”. The p-value was > .05 (.086): hence, the difference was not significant. Moreover, the observed standardized effect size was small (0.26). This finding indicates that misleading explanations do not seem to influence the confidence level expressed by the participants. This supports the hypothesis that the respondents do not realise that the explanations are misleading. See also Figure 9 for a graphical representation of this result.

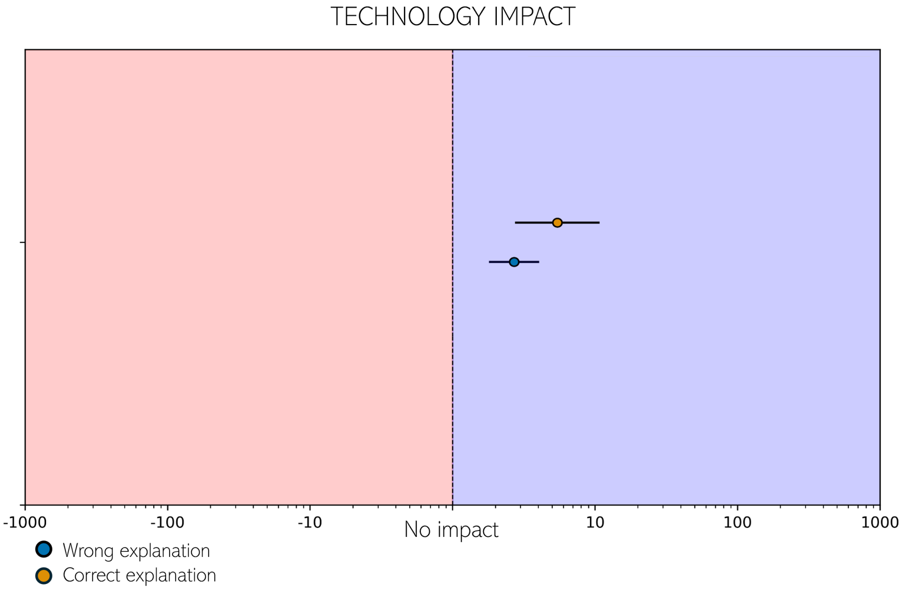


Fig. 7: Technology Impact diagram, the red region denotes an overall negative effect of the AI intervention, while the blue region denotes an overall positive effect. Generated with the tool available at https://mudilab.github.io/dss-quality-assessment/.

6 Discussion

In the above reported study we focused on the so called explainability paradox, and in particular on the pattern we denoted as XAI Halo effect: bad explanations corrupting the perceived quality of the advice. This pattern complements a likewise paradoxical pattern that has been investigated several times in the specialist literature [10, 11], that is often denoted as white-box paradox: good explanations making people mistakenly believe that bad advice is actually good.

Our findings reveal that there is a negative impact on decision-making accuracy when participants received right AI advice accompanied by misleading explanations, and significantly so, compared to when both the advice and explanations were good. This finding is less obvious than it appears: in fact, a correct advice in the type of cognitive effort regarding logic tests would easily allow the correctness of the advice to be checked. However, participants seemed to be satisfied with the explanation, with this latter generating a kind of temporary blindness about the correctness of the advice.

Moreover, misleading explanations seem to have no bad smell, so to say: the difference in confidence regarding the final decision, between the cases with either misleading explanations or coherent, relevant ones, was not significant. This means that misleading explanations did not significantly undermine the

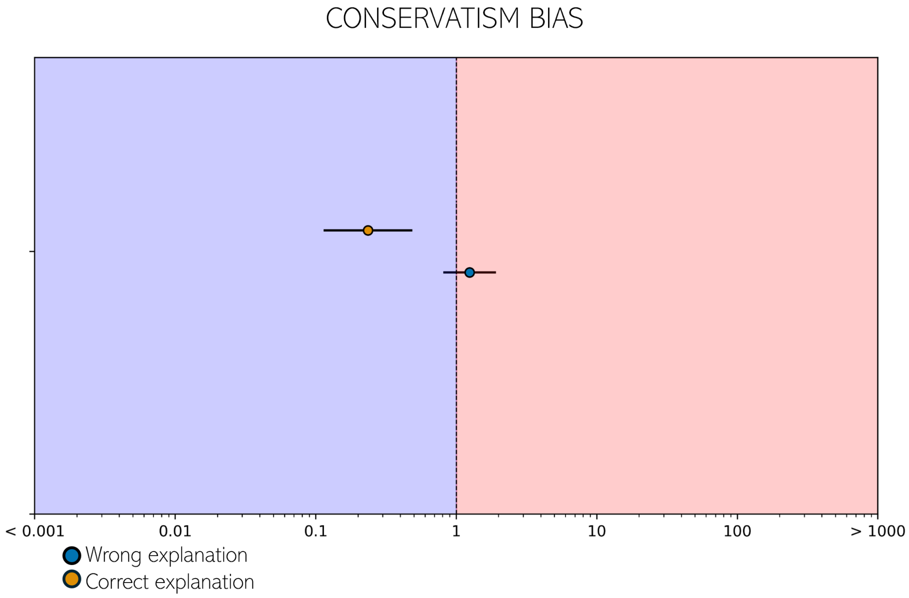


Fig. 8: Conservatism Bias diagram, the red region denotes an overall negative effect of the AI intervention, while the blue region denotes an overall positive effect. Generated with the tool available at https://mudilab.github.io/dss-quality-assessment/.

participants' confidence, who therefore were unsuspecting of the fact that they had followed a wrong suggestion.

Furthermore, we observed that misleading explanations are more harmful for low-performers than top-performers: this effect, although intuitive, was associated with a significant effect, notwithstanding the relatively low number of cases and participants; this therefore suggests that the mitigation of these effects should be considered seriously, especially so as not to harm those who would most need this kind of functionality.

This study, due to its small number of cases (19) and participants (22), inherently faces limitations and thus should be considered an exploratory examination of the impact of sub-optimal, and thereby potentially misleading, explanations on decision-making performance. For this reason, we emphasize the results related to the observed effect sizes rather than the observed significance levels, which, in any case, fall below the significance level for rejecting the null hypothesis regarding the impact on decision accuracy.

Moreover, our interpretation of the results relies on these simplifying assumptions: 1) that the task was perceived as difficult by the respondents, leading them naturally to consider the system's advice carefully; 2) the exposure to the cases, along with the advice and explanation, did not permit the respondents to form

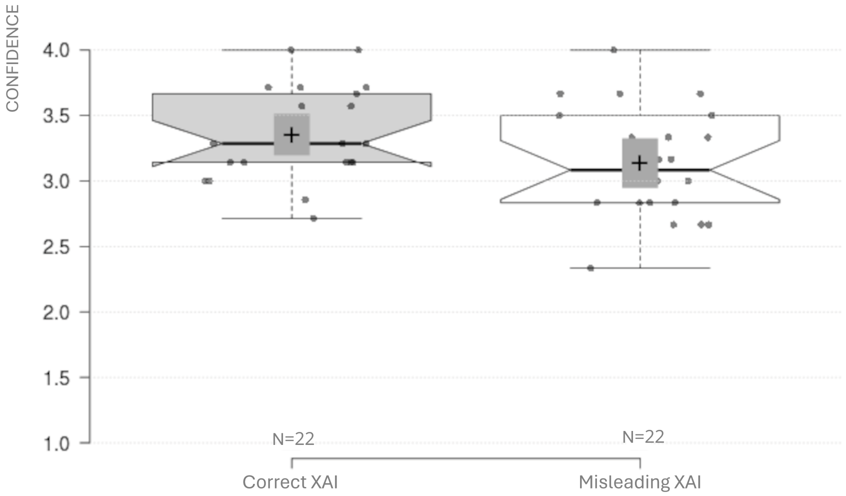


Fig. 9: Box-plots of the confidence scores reported by the participants after they have solved the good-advice&good-explanation tests (Correct XAI) and when they solved the good-advice&bad-explanation ones (Misleading XAI). Box notches indicate the 95% confidence intervals of the median confidence score, while crosses indicate the mean confidence score within its confidence interval (grey rectangle). Confidence scores were defined on a scale from 1 to 4.

too strong an opinion about the correct answer. A concurrent study we are conducting in the same setting supports the validity of these assumptions: notably, on average, participants who were not supported by the AI got slightly more than half of the answers wrong, whereas the AI was, on average, more accurate (70%) — a fact of which participants were informed. Moreover, the conjecture, contrary to our assumption, that respondents might display algorithmic aversion and stick to their own initial opinions, would render the results we observed conservative estimates of the actual effect if respondents were fully compliant with AI advice. Another factor contributing to the conservative nature of our results, and thus serving as a lower-bound estimate of the effects that might be observed under other conditions, regards the observed difficulty of the cases associated with misleading explanations, which was actually lower than for the cases associated with more accurate explanations. The fact that irrelevant explanations associated with correct advice led to a higher number of errors, especially in the simpler cases in the test, thus suggests that the explainability paradox is a real and significant effect that should not be underestimated and that all researchers and practitioners serious about XAI support should be aware of.

Therefore, our recommendation is to prioritize the quality of explanations — in terms of consistency with advice and relevance to the case — when designing XAI support and to demand high-quality explanations especially in situations where accuracy is more critical than transparency (e.g., in medicine [17]). However, researchers should also be aware of the effect associated with excellent explanations for poor advice, that is the other side of the explainability paradox (that some authors call white-box paradox). For this reason, our recommendation is designing XAI systems that do not generate explanations when the confidence of the advice module about its output is below a conservatively high threshold (see the concept of cautious learning and abstention [12]).

7 Conclusion

This study has investigated the phenomenon known as the "explainability paradox" within the domain of Explainable AI (XAI), focusing particularly on the adverse effects of misleading explanations on decision-making accuracy and confidence. Our findings underscore the crucial role of explanation quality in influencing user reliance on AI advice, shedding light on the so called *XAI halo effect*, wherein misleading explanations can significantly impair decision accuracy defiling accurate AI-generated advice. Notably, this impact is more pronounced among lower-performing individuals, underscoring the critical importance of tailoring explanations to support effective decision-making across all user groups.

Our exploration into the realm of misleading explanations reveals a complex interplay between explanation coherence (with respect to the advice given), relevancy (with respect to the case), and the consequent user trust in AI advice (in terms of either reliance or rejection). We observed that, despite the accuracy of AI advice, misleading explanations lead to a diminished capacity for users to critically assess the advice and understand its relevance for their final decision. This outcome emphasizes the necessity for XAI systems to not only generate accurate advice but also provide explanations that are contextually relevant and coherent, enhancing the overall quality and reliability of human-AI collaboration.

Although exploratory, our study highlights the need for ongoing research to delve deeper into the mechanisms through which explanations influence user perception and decision-making when XAI systems are deployed to support it. In our study we focused on cases where explanations were consistent with the given advice but not relevant to the classification task. We therefore advocate that future research should dedicated to investigating the remaining two configurations of our framework (see Figure 2), as well as to confirming the results we report in regard to the more common type, coherent&irrelevant explanations. Future research endeavors in this direction are not only warranted but crucial for the advancement of responsible and effective XAI implementations.

Acknowledgements

C. Fregosi and F. Cabitza acknowledge funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision making. CUP: H53D23008090001 funded by the European Union - Next Generation EU.

C. Natali gratefully acknowledges the PhD grant awarded by the Fondazione Fratelli Confalonieri, which has been instrumental in facilitating her research pursuits.

Disclosure of Interest

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
2. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D.: 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. pp. 1–16 (2021)
3. Bertrand, A., Belloum, R., Eagan, J.R., Maxwell, W.: 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 (2022)
4. Buçinca, Z., Malaya, M.B., Gajos, K.Z.: 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**(CSCW1), 1–21 (2021)
5. Bussone, A., Stumpf, S., O'Sullivan, D.: The role of explanations on trust and reliance in clinical decision support systems. In: 2015 international conference on healthcare informatics. pp. 160–169. IEEE (2015)
6. Cabitza, F., Campagner, A., Angius, R., Natali, C., Reverberi, C.: Ai shall have no dominion: on how to measure technology dominance in ai-supported human decision-making. In: Proceedings of the 2023 CHI conference on human factors in computing systems. pp. 1–20 (2023)
7. Cabitza, F., Campagner, A., Famiglini, L., Gallazzi, E., La Maida, G.A.: Color shadows (part i): Exploratory usability evaluation of activation maps in radiological machine learning. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 31–50. Springer (2022)
8. Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., Holzinger, A.: Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable ai. Expert systems with Applications **213**, 118888 (2023)
9. Cabitza, F., Campagner, A., Natali, C., Parimbelli, E., Ronzio, L., Cameli, M.: Painting the black box white: experimental findings from applying xai to an ecg reading setting. Machine Learning and Knowledge Extraction **5**(1), 269–286 (2023)
10. Cabitza, F., Campagner, A., Ronzio, L., Cameli, M., Mandoli, G.E., Pastore, M.C., Sconfienza, L.M., Folgado, D., Barandas, M., Gamboa, H.: Rams, hounds and white boxes: Investigating human–ai collaboration protocols in medical diagnosis. Artificial Intelligence in Medicine **138**, 102506 (2023)

11. Cabitza, F., Campagner, A., Simone, C.: The need to move away from agential-ai: Empirical investigations, useful concepts and open issues. International Journal of Human-Computer Studies **155**, 102696 (2021)
12. Campagner, A., Cabitza, F., Ciucci, D.: Three–way classification: Ambiguity and abstention in machine learning. In: Rough Sets: International Joint Conference, IJCRS 2019, Debrecen, Hungary, June 17–21, 2019, Proceedings. pp. 280–294. Springer (2019)
13. Dahlbäck, N., Jönsson, A., Ahrenberg, L.: Wizard of oz studies: why and how. In: Proceedings of the 1st international conference on Intelligent user interfaces. pp. 193–200 (1993)
14. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al.: Explainable ai (xai): Core ideas, techniques, and solutions. ACM Computing Surveys **55**(9), 1–33 (2023)
15. Ebermann, C., Selisky, M., Weibelzahl, S.: Explainable ai: The effect of contradictory decisions and explanations on users' acceptance of ai systems. International Journal of Human–Computer Interaction **39**(9), 1807–1826 (2023)
16. Eiband, M., Buschek, D., Kremer, A., Hussmann, H.: The impact of placebic explanations on trust in intelligent systems. In: Extended abstracts of the 2019 CHI conference on human factors in computing systems. pp. 1–6 (2019)
17. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**(11), e745–e750 (2021)
18. Huff, S.L., Higgins, C., Lin, J.T.M.: Computers and the halo effect. Journal of Systems Management **38**(1), 21 (1987)
19. Kieseberg, P., Weippl, E., Tjoa, A.M., Cabitza, F., Campagner, A., Holzinger, A.: Controllable ai-an alternative to trustworthiness in complex ai systems? In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 1–12. Springer (2023)
20. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human factors **46**(1), 50–80 (2004)
21. Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., et al.: Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion p. 102301 (2024)
22. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
23. Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., Perer, A.: The impact of imperfect xai on human-ai decision-making. arXiv preprint arXiv:2307.13566 (2023)
24. Papenmeier, A., Englebienne, G., Seifert, C.: How model accuracy and explanation fidelity influence user trust. arXiv preprint arXiv:1907.12652 (2019)
25. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
26. Schemmer, M., Kuehl, N., Benz, C., Bartos, A., Satzger, G.: Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. pp. 410–422 (2023)
27. Shin, D.: The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. International Journal of Human-Computer Studies **146**, 102551 (2021)

28. Van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior. In: Proceedings of the national conference on artificial intelligence. pp. 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2004)
29. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable ai. In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–15 (2019)
30. Wang, X., Yin, M.: Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In: 26th international conference on intelligent user interfaces. pp. 318–328 (2021)
31. Yeh, C.K., Hsieh, C.Y., Suggala, A., Inouye, D.I., Ravikumar, P.K.: On the (in) fidelity and sensitivity of explanations. Advances in neural information processing systems **32** (2019)