Bayesian reasoning for overcoming over-reliance in AI-assisted decision making

Daria Mikhaylova[1,*], Tommaso Turchi[1], Gustavo Cevolani[2] and Alessio Malizia[1]

[1]Department of Computer science, University of Pisa, Largo B. Pontecorvo, 3, 56127, Pisa, Italy
[2]IMT School for Advanced Studies Lucca, Piazza S. Francesco, 19, 55100, Lucca, Italy

Abstract

The tendency of users to blindly follow the suggestions of AI-powered decision support systems is a troubling phenomenon that may lead to unjustifiable errors and unreasonable decisions. Known as over-reliance, it is especially problematic when decisions involve high uncertainty and high costs for all parties. In this paper we discuss over-reliance as a factor that worsens the quality of the decision-making process and introduces unwanted and uncontrollable noise, like other cognitive biases. We propose to approach over-reliance within a Bayesian account of rationality and to model it as a form of *base rate neglect*, a well-known bias violating sound probabilistic reasoning. Finally, based on existing studies both on over-reliance in human-computer interaction (HCI) and on cognitive biases in human reasoning, we suggest a form of interaction that may be useful for mitigating over-reliance.

Keywords

over-reliance, automation bias, AI-assisted decision making, Bayesian reasoning, human-computer interaction

1. Introduction

Today machine learning models are extensively used as a support for tasks in various domains, including manufacturing, creative industries, public services, education, and medicine. Known for their rapid processing time and variety of reliable algorithmic methods, they can help to reduce cognitive overload of the specialists, increase decision efficiency, and perform preliminary procedures before expert intervention.

The type of decisions supported by AI systems, varies greatly in complexity and stakes. In some situations, AI systems provide an extension or even a replacement for the limited experience or expertise of single human decision-makers in prediction tasks. Since large volumes of images or tabular data are used to train machine learning models, they offer powerful tools for statistical inference and, in many tasks, are provably more accurate than humans. In other situations, these models are employed in *evaluative* judgements [1, pp. 46-48], such as in criminal re-offence evaluations, employees or student choices or social welfare allocation. Associated with high uncertainty and high stakes for both decision makers and the affected side, these tasks require consistency and reproducibility. For this kind of evaluative judgements, AI support systems are often perceived as a valuable tool for improving the quality of the decision process by overcoming *biases* of human decision makers. The intuition is that algorithms may mitigate the influence of contextual factors (subjective opinions, inclinations, cognitive overload, stress, distraction, etc.) on the decisions of humans and make them less variant across time and different judges.

However, it is now well-known that AI systems may also influence the decision making process negatively and often in unexpected ways. On the one hand, the big data used to train the models not only leads to high predictive power but also brings into play implicit biases from past decisions and

HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy

*Corresponding author.
daria.mikhaylova@phd.unipi.it (D. Mikhaylova); tommaso.turchi@unipi.it (T. Turchi); gustavo.cevolani@imtlucca.it (G. Cevolani); alessio.malizia@unipi.it (A. Malizia)
0000-0002-8853-4669 (D. Mikhaylova); 0000-0001-6826-9688 (T. Turchi); 0000-0001-8770-8877 (G. Cevolani); 0000-0002-2601-7009 (A. Malizia)

data collection procedures [2]. On the other hand, the use of the AI tool itself may trigger cognitive biases that do not depend on the training data, nor on the performance of the model. These types of distortions are less acknowledged and studied, but nevertheless have important ethical and social implications, in particular in the light of the meaningful human oversight for high stakes decisions required by recently enforced "AI Act" [3, Art.14].

In this paper, we focus on one such phenomenon arising from the interaction with intelligent decision support systems and known to produce an unjustified variance in the decisions of individuals, i.e., "over-reliance". Over-reliance refers to the tendency of users of AI support systems to base their decisions and actions solely on the suggestions coming from the system, ignoring other relevant information, in particular the clues inaccessible to the system itself. Over-reliance may lead decision makers to unreasonable decisions, that they probably wouldn't have reached if unassisted.

In Sec. 2, we briefly discuss existing theoretical approaches to explain and mitigate over-reliance and we re-conceptualize over-reliance as a cognitive bias, thus addressing it as a deviation from rational expert judgement. In particular, we model over-reliance as a form of *base rate neglect*, a well-known cognitive bias which departs from sound Bayesian reasoning under uncertainty (Sec. 3). Finally, we suggest an interaction that we believe may aid decision makers in better evaluating the contribution of an AI system and their own observations to the final decision (Sec. 4). were

2. Over-reliance

Over-reliance refers to a tendency of users to accept suggestions and predictions from an intelligent decision support systems with insufficient or no further evaluation, sometimes leading to surprising and unjustifiable errors in decisions (for a comprehensive review and examples see, e.g., [4, 5]). It affects decision makers in different fields where AI support systems are employed, including clinical diagnosis, forensic science, credit scoring and human resources management.

Since over-reliance was first acknowledged in the aviation industry in the late eighties, the Human-Computer Interaction (HCI) community has developed different approaches to explain the phenomenon and design mitigation strategies. One of the first explanations of over-reliance was proposed in connection with operators monitoring automated control systems and referred to *complacency*, a psychological state of self-satisfaction resulting in non-vigilance. [6, 7, 8, 9]. A normative level of attention required to correctly control an automated system was calculated borrowing from the theory of signal processing and expressed as a sufficient sampling rate according to the Nyquist–Shannon theorem.

Today, particularly in relation to AI-assisted decision making, the most prevalent explanation of over-reliance is in terms of an attitude of trust towards automated systems [10, 11, 12]. Accordingly, to understand and mitigate over-reliance, HCI studies have controlled various factors that usually contribute to reciprocal intrapersonal trust, such as self-confidence [e.g. 13] for what concerns humans, and interpretability [e.g. 14, 15, 16] and performance of the model [e.g. 17] as far as the system itself is concerned.

Within this approach, a widely accepted normative framework for AI-assisted decision making regards *the appropriate level of reliance* that should correspond to the AI performance. If the AI system is accurate, say, 80% of time, then the human expert should rely on its suggestion in around 80% of cases. The intuition is that if the reliance level is right, we can expect more accurate predictions from an AI-assisted human user, than from only the user or only the machine.

This approach has however its own limitations. First, the overall calibrated reliance may not be sufficient to guarantee better outcomes. *Complementary performance* in AI assisted decision making may be achieved only if human experts not only rely sufficiently and in a limited way on the AI suggestions but also understand *when exactly* to accept the suggestion. To achieve this the users need to correctly evaluate their own expertise and the expertise of the AI. This limitation has been recently acknowledged and is widely discussed in current studies. Guo et al. [18] for instance, formalised a decision framework that distinguishes between performance loss caused by mis-reliance (over-reliance or under-reliance),

and that caused by erroneous evaluation of accuracy of the human or the AI in prediction for each particular case.

As of today, to achieve complementarity in performance some researches in the HCI community propose to use the personalized interaction with AI-assistance. The general idea is to vary access to AI suggestion, interface or interaction pattern based on the user's past interactions with the system. The personalisation may be a function of either past reliance [19, 20] or the accuracy of the system and of the user for similar tasks [21]. These approaches require building a dedicated supplementary machine learning model that predicts the user behaviour, a solution previously explored also in the context of fair machine learning [22].

Another line of research on complementarity in decisions leave to the human user the evaluation of the expertise of the AI and that of possible errors in a particular case. This evaluation is commonly incorporated as a step in the interaction with the decision aid and is requested as a confidence of the user in the decision. The overall idea is that making the user aware of their self-confidence before and after seeing the AI suggestion may lead to a more calibrated reliance in cases where the model is not accurate, and the user is confident in their own decisions. However, the results of experimental studies showed that although confidence decreases when the user experiences unreliable system and when the prediction is expressed as a less certain[1], the reliance on the proposed suggestion remains the same [23, 24, 13, 25]. Moreover, Li et al. [26] demonstrated that case-by-case confidence in the decision and reported self-confidence is altogether aligned with the confidence of the AI model for that decision.

The concern about accuracy in particular cases also led to the use of more granular definitions and measures of over-reliance, which distinguish between correct and incorrect suggestions of AI and, in some cases, of the human user. In fact, currently the most common experimental measure is the proportion of cases in which the human decision maker accepted incorrect AI suggestions over all cases, and under-reliance as a proportion of rejected suggestions that were correct.

The above approaches partially address the question of complementary expertise, but do not address a second limitation of the calibrated reliance approach, that is its substantial dependence on accuracy as the only metrics for evaluating both over-reliance and the overall joint decision-making process. While perfectly suitable for immediately verifiable predictions, this approach cannot account for evaluative judgements and non-verifiable predictions since their *correctness* is either unknown or can be evaluated only in the long run on a series of similar cases. In such cases, a better strategy is to address human-AI interaction focusing on the reasoning process that leads to the final, joint judgement, in the light of what we know from the cognitive science of human reasoning and decision making [1]. This allows us to interpret and assess over-reliance, even when it is impossible to evaluate the correctness of the outcome decision, in terms of how the decision process deviates significantly from the prescriptions of normative frameworks for sound reasoning, such as logic, probability theory, or rational choice theory.

This approach follows the "heuristics and biases" research program [27, 28, 1], which explores the informal reasoning and decision-making strategies spontaneously employed by both laypersons and experts (the so-called heuristics) to study the possible resulting deviations from the norms of rationality (the so-called biases). It was first applied to over-reliance in early studies of cases where the excessive reliance of pilots on flight management systems led to accidents. Mosier et al [29, 30, 31] coined the term *automation bias* to describe systematic errors that occur when human operators use automation suggestions as cognitive heuristics, a mental shortcut for the decision. The heuristic use, they observed, substitutes for "vigilant information seeking and processing" expected from a rational and responsible decision maker. The decision process, affected by automation bias, would not satisfy the expectation of the limited variance between judgements [1], because the users would disagree with the very decision, they would have taken on the basis of the same relevant information but without the aid of the decision support.

[1]The so-called confidence of the model is the output of the softmax function from the last layer of the neural network, commonly interpreted as a probability of the input data to belong to the predicted class

3. Over-reliance as base rate neglect

To date, several studies in Human-AI interaction approach over-reliance as a cognitive bias and drew parallels with more well-studied cognitive errors, such as confirmation bias [32] or anchoring effect [33, 34]. However, they do not address directly one of the aspects of over-reliance — namely how it prevents the decision maker from seeking or weighing additional information relevant to the case and incorporating the recommendation of the AI system in their own flow of reasoning. We argue that over-reliance may be seen as a particular instance of another well-known and extensively studied cognitive error — i.e., *base rate neglect*, first described by Kahneman and Tversky [35] and reproduced at least to some extent in multiple studies (for examples and reviews see [36, 37, 38]). In the following, we briefly outline the general idea, which is discussed in more details by Mikhaylova & Cevolani [39], with applications in the context of forensic decision making.

A simple example of base rate neglect discussed by Kahneman [1] will illustrate the general point. A typical case of real-world evaluative judgement is that of a hiring specialist assessing a candidate for a managerial position, with the aim of predicting whether the successful candidate will remain in the position for more than three years. According to Kahneman, hiring specialists tend to ground their estimates entirely on the information presented in the resumes of the candidates. While perfectly intuitive, this process fails to take into account a crucial piece of *external* information: the attrition rate for managers in that specific field, i.e., how likely is that any randomly chosen candidate will still hold the job in three years. Failing to assess such initial probability, especially when it is quite low, may lead to over-estimating the final probability of success, in particular if the CV is compelling and we think that CV is indeed a good indicator of the future performance. The ignored initial probability is the so-called base rate, and the fallacy of ignoring or under-weighting it is referred to as "base rate neglect".

Here, the underlying normative theory is provided by Bayesian rationality [40, 41]. In general, to evaluate the probability in favour or against some hypothesis H (in the example, that the candidate will stay for at least three years) in the light of some evidence E (the resume in this case), one should follow Bayes' rule:

$$p(H|E) = \frac{p(E|H)p(H)}{p(E|H)p(H) + p(E|\neg H)p(\neg H)} \tag{1}$$

Here, H and $\neg H$ are, respectively, the relevant hypothesis and its negation, e.g., "the candidate will stay" or "the candidate will leave". (In the general case, we may consider a full set of mutually incompatible and jointly exhaustive hypotheses, instead of a single binary hypothesis.) The so-called likelihood of H, i.e., $p(E|H)$, is the probability of the evidence being observed provided the hypothesis is true. The likelihood $p(E|\neg H)$ of its negation is, instead, the probability of the evidence being observed, provided the hypothesis is false. Bayes' rule prescribes that $p(H|E)$ — the final or "posterior" probability of H — has to be proportional to both the initial probability of H and to the likelihood of H (i.e., the probability of observing E assuming H is true). Intuitively, H is more likely in light of E the more H was likely before observing E, and the more likely is E assuming that H is true instead of false.

The phenomenon of base rate neglect occurs when the decision maker, for whatever reason, fail to properly take into account prior probabilities. This typically happens in two ways. One is more or less implicitly setting all relevant prior probabilities equal to $1/n$ (where n is the number of hypothesis). In the case of a binary decision task, this means assuming $p(H) = p(\neg H) = 0.5$, which leads to strongly overestimate the prior probability of the hypothesis being true in many cases (like in the hiring example above, where the probability of a manager staying for three years may well be very low in general).

Another way of triggering the base-neglect bias is directly conflating the required posterior probability $p(H|E)$ with the known likelihood $p(E|H)$ of the hypothesis, a well-documented confusion known as the "probabilistic inverse fallacy". In both cases, the final decision is entirely driven by the evaluation of the likelihood of the hypotheses at issue, ignoring the essential information about their prior probability.

As we suggest, this bias is precisely the kind of cognitive mechanism that may be at play in many instances of over-reliance, when the user accepts an AI suggestion without evaluating further relevant information. The general idea is the following. The decision maker interacting with an AI-based

support system aims at evaluating some hypothesis H. The system provides a suggestion that works as a relevant piece of evidence E. The decision maker takes into account such evidence in order to rationally estimate the final probability of H on the basis of both of E and of its prior probability, as prescribed by Bayes rule. If the agent only relies on E to make the evaluation, base rate neglect and over-reliance will follow.

Note that, in the case of machine-learning models, the likelihood $p(E|H)$ is measured in terms of "recall" or "true positive rate", whereas the likelihood $p(E|\neg H)$ corresponds to the "false positive rate" (FPR) [42]. This leads to the following reformulation of Bayes' rule as applied to estimating the posterior probability of a hypothesis given the evidence provided by the AI system's suggestion:

$$p(H|E) = \frac{Recall \cdot p(H)}{Recall \cdot p(H) + FPR \cdot p(\neg H)} \tag{2}$$

Over-reliance occurs when the user ignores the information about $p(H)$, as explained above.

In this section we briefly introduced the analysis of over-reliance as a deviation from Bayesian rationality, i.e., as a case of base rate neglect. We suggest that human decision-makers may systematically ignore relevant background information (as represented by the prior probability of the hypothesis at issue) and use the reliability of the system (represented by the Recall measure) as a proxy to assess the desired posterior probability. This tendency is well documented in the cognitive science literature, even if, as far as we know, has not been discussed in the context of HCI studies on over-reliance. In the next section we describe a proposal for the interaction and interface that may mitigate the over-reliance as an under-weighting of external information.

4. Mitigating over-reliance with probability estimation

In the previous sections we briefly described a theoretical approach to over-reliance as a cognitive bias known as *base rate neglect*. This analysis opens up an opportunity to take advantage of many decades of studies on base rate neglect and to suggest mitigation strategies based on the cognitive science literature. Most of the existing proposals focus on facilitating Bayesian reasoning in unassisted decision making — a non-trivial task that is addressed through training and finding a more natural way to represent probabilities, for example with the aid of "frequency trees" [43][2].

For AI-assisted decision-making, it seems a good starting point to see if emphasising the request for outcome prediction and exposing the users to the calculation of posterior probability would prevent over-reliance. As we mentioned in section 2, currently experimental results show that to overcome the over-reliance it's necessary in the interaction with AI decision support to ask the user's evaluation of the case and the associated uncertainty, but it is commonly requested as a confidence in the decision and not as an evaluation of the posterior probability. These two measures are, however, distinctly different. The studies on base rate neglect in cognitive psychology demonstrated that decision makers in intuitive judgements apparently substitute the question "how likely is the outcome?" with "how good is the provided evidence?" [35, 44]. If the evidence is not reliable, the predictions of the final probability decrease, but do not shift towards base rates, as prescribed by the Bayesian reasoning, and remains highly correlated with the representativeness of the evidence. In other words, the less accurate evidence apparently only decreases confidence in the prediction instead of soliciting a completely different one, based on the outside information. For this reason, we stress the necessity to solicit the prediction of the outcome, instead of the confidence, and give support to the user to evaluate it.

As a contribution to ongoing research on over-reliance, we propose to develop and test a prototype of AI-assisted decision-making system with an "AI-follow" interaction pattern. Such interaction "begins with the user forming an independent preliminary prediction given the decision-making problem. Following this initial judgment, the AI's predicted outcome is presented and may be accompanied by support information" [45]. This type of interaction is commonly opposed to "AI-first assistance", in which the system displays simultaneously the decision-making problem and the AI-predicted outcome.

[2]A method that was explored for communicating AI model confidence by Cao et al [25]

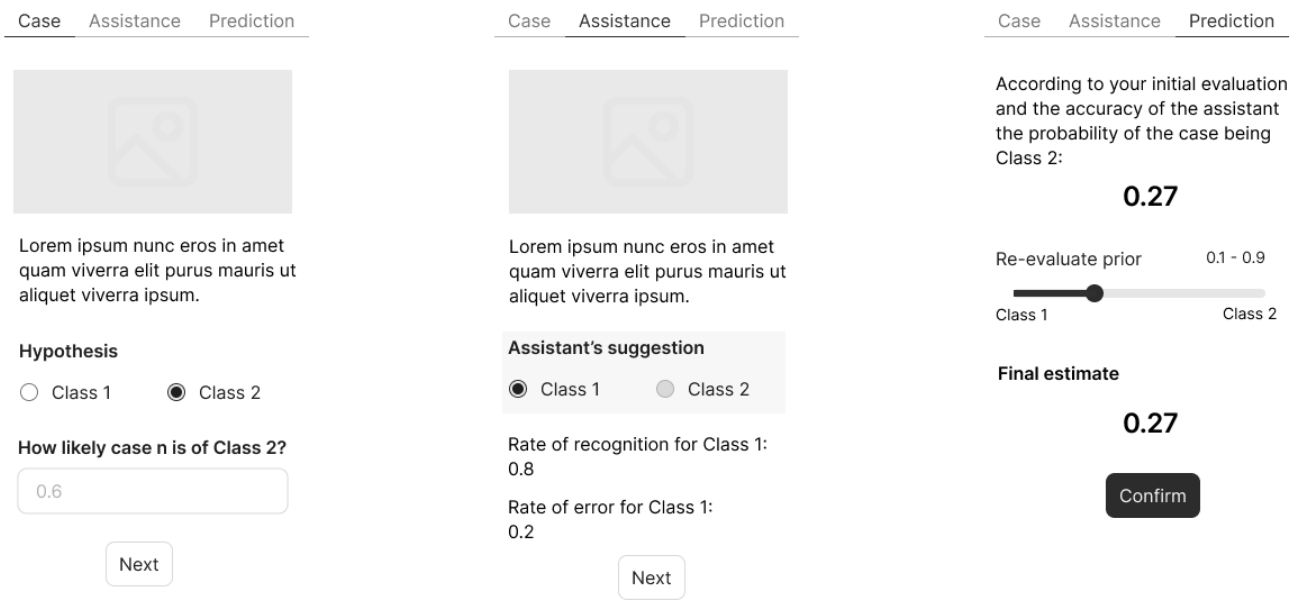

Figure 1: A prototype of the interface for the system that provides a calculation of the posterior probability. Left: providing a case information and requesting the hypothesis and the estimate of its probability. Middle: Showing the system suggestion and reliability information. Right: Showing computed posterior probability and an interactive slider to evaluate the influence of the prior.

The change from the latter to former is seen as a cognitive forcing technique that may in itself decrease over-reliance [46], however alone it is known also to produce under-reliance.

To our knowledge, only Agudo et al [47] used an interface that requested a probability estimation of the hypothesis (whether the defendant was guilty or not of a criminal offence, given a series of testimonies), however, their aim was not to manipulate aggregation of probabilities, but to test if the change in interaction pattern (from AI-first to AI-follow assistance) would affect over-reliance. Building on their work, we propose several changes. First of all, we propose to apply the model of Bayesian reasoning to the interaction with the AI system, not only requesting the user's probability estimate for their hypothesis, but also providing the output of an algorithm for opinion aggregation with AI prediction.

Furthermore, we suggest taking into account different ways in which humans and AI-systems express probability estimates. While humans may express probabilities on a scale from 0 to 100, the output of the classification system is primarily a predicted class, eventually associated with "confidence" on a scale from 50 to 100 for a binary prediction (if the confidence is less than 50, the other class is reported as the output). In order to enable aggregation, either the human input should be aligned with the way the AI system operates, or the system's output must align with the human way of communicating probabilities. For now, our proposal is to first request from the user the prediction of the class and then the associated probability on a 50–100 scale.

Figure 1 presents a prototype interface that implements this kind of interaction. In the first step (shown in the prototype as the tab "Case") the system displays the background information relevant to the case and the input to a model (either tabular or image which depends on the decision task), it also asks the evaluation of the user for the hypothesis and an estimate of its prior probability. In the second step (tab "Assistance") the system continues to display the information for the case and gives the AI prediction, it also makes the reliability information clear, providing the Recall and the false positive rate. In the last step (tab "Prediction" on Fig.1) of the interaction the system computes the posterior probability based on the prior, provided by the user and displays the result. We suggest also to give the user the possibility to interactively explore how the prior probability contributes to the prediction of the outcome, given the reliability of the system. We expect that this kind of interaction may have two

relevant consequences. First, it would make salient to the user the role of prior probabilities in assessing the final probability of the decision, thereby mitigating base rate neglect and hence over-reliance on the suggestions of the system. Second, it would provide a sort of "robustness" check, making clear to the user how reliable the system needs to be in order to safely disregard prior probabilities and simply accept the provided suggestion as a final prediction.

5. Conclusion

In this paper, we proposed to approach the over-reliance on AI-powered decision support systems as a well-studied cognitive bias, base rate neglect.

Our analysis of the over-reliance as a cognitive bias provides a way to interpret and to assess it, even when it is impossible to evaluate the correctness of the outcome decision, in terms of how the decision process deviates significantly from the prescriptions of normative frameworks for sound reasoning.

As a way to mitigate over-reliance, we propose to implement in the interface and interaction of the AI-powered decision support system the calculation of the posterior probability of the outcome and invite the user to evaluate the relationship between the prior probability, usually coming from the external sources, and the prediction of the system, according to Bayes' rule.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

Acknowledgments

Authors thank the reviewers of the draft version for their useful remarks and suggestions. D.M. expresses her gratitude to Prof. Alan Dix and Dr. Ben Wilson for thorough discussion of this work and their detailed comments for the workshop and the final versions.

References

[1] D. Kahneman, O. Sibony, C. R. Sunstein, Noise: A Flaw in Human Judgment, second ed., William Collins, 2021.

[2] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, Bias in data-driven artificial intelligence systems—An introductory survey, WIREs Data Mining and Knowledge Discovery 10 (2020). doi:10.1002/widm.1356.

[3] European Parliament and The Council, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/eu, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj.

[4] 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 : JAMIA 19 (2012) 121–127. doi:10.1136/amiajnl-2011-000089.

[5] D. Lyell, E. Coiera, Automation bias and verification complexity: a systematic review, Journal of the American Medical Informatics Association 24 (2017) 423–431. doi:10.1093/jamia/ocw105.

[6] R. Parasuraman, R. Molloy, I. L. Singh, Performance consequences of automation-induced "complacency.", The International Journal of Aviation Psychology 3 (1993) 1–23. doi:10.1207/s15327108ijap0301_1.

[7] V. Riley, Issues Associated with Operator Use of Automation, Technical Report, SAE International, Warrendale, PA, 1995. doi:10.4271/951985.

[8] R. Parasuraman, V. Riley, Humans and Automation: Use, Misuse, Disuse, Abuse, Human Factors 39 (1997) 230–253. doi:10.1518/001872097778543886.

[9] R. Parasuraman, D. H. Manzey, Complacency and Bias in Human Use of Automation: An Attentional Integration, Human Factors: The Journal of the Human Factors and Ergonomics Society 52 (2010) 381–410. doi:10.1177/0018720810376055.

[10] B. M. Muir, Trust between humans and machines, and the design of decision aids, International Journal of Man-Machine Studies 27 (1987) 527–539. doi:10.1016/S0020-7373(87)80013-5.

[11] P. Madhavan, D. A. Wiegmann, Similarities and differences between human–human and human–automation trust: an integrative review, Theoretical Issues in Ergonomics Science 8 (2007) 277–301. doi:10.1080/14639220500337708.

[12] A. Ferrario, M. Loi, E. Viganò, In ai we trust incrementally: a multi-layer model of trust to analyze human-artificial intelligence interactions, Philosophy & Technology 33 (2019) 523–539. doi:10.1007/s13347-019-00378-3.

[13] S. Ma, X. Wang, Y. Lei, C. Shi, M. Yin, X. Ma, "Are you really sure?" Understanding the effects of human self-confidence calibration in ai-assisted decision making, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024. doi:10.1145/3613904.3642671.

[14] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. Bernstein, R. Krishna, Explanations Can Reduce Overreliance on AI Systems During Decision-Making, 2023. doi:10.48550/arXiv.2212.06823, arXiv:2212.06823 [cs] type: article.

[15] M. Vered, T. Livni, P. D. L. Howe, T. Miller, L. Sonenberg, The effects of explanations on automation bias, Artificial Intelligence 322 (2023) 103952. doi:10.1016/j.artint.2023.103952.

[16] M. Schemmer, N. Kühl, C. Benz, G. Satzger, On the Influence of Explainable AI on Automation Bias, 2022. doi:10.48550/arXiv.2204.08859, arXiv:2204.08859 [cs] type: article.

[17] Z. Ashktorab, M. Desmond, J. Andres, M. Muller, N. N. Joshi, M. Brachman, A. Sharma, K. Brimijoin, Q. Pan, C. T. Wolf, E. Duesterwald, C. Dugan, W. Geyer, D. Reimer, Ai-assisted human labeling: Batching for efficiency without overreliance, in: Proceedings of the ACM on Human-Computer Interaction, volume 5, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–27. doi:10.1145/3449163.

[18] Z. Guo, Y. Wu, J. Hartline, J. Hullman, A decision theoretic framework for measuring ai reliance, 2024. doi:10.48550/ARXIV.2401.15356.

[19] Y. Fukuchi, S. Yamada, Dynamic selection of reliance calibration cues with ai reliance model, IEEE access 11 (2023) 138870–138881.

[20] S. Swaroop, Z. Buçinca, K. Z. Gajos, F. Doshi-Velez, Personalising ai assistance based on overreliance rate in ai-assisted decision making, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25, ACM, 2025, pp. 1107–1122. doi:10.1145/3708359.3712128.

[21] S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, X. Ma, Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3544548.3581058.

[22] F. Mazzoni, R. Guidotti, A. Malizia, A Frank System for Co-Evolutionary Hybrid Decision-Making, in: I. Miliou, N. Piatkowski, P. Papapetrou (Eds.), Advances in Intelligent Data Analysis XXII, volume 14642, Springer Nature Switzerland, Cham, 2024, pp. 236–248. doi:10.1007/978-3-031-58553-1_19.

[23] J. Cecil, E. Lermer, M. F. C. Hudecek, J. Sauer, S. Gaube, Explainability does not mitigate the negative impact of incorrect ai advice in a personnel selection task., Scientific reports 14 (2024) 9736.

[24] R. Wiczorek, J. Meyer, Effects of Trust, Self-Confidence, and Feedback on the Use of Decision Automation, Frontiers in Psychology Volume 10 - 2019 (2019). doi:10.3389/fpsyg.2019.00519.
[25] S. Cao, A. Liu, C.-M. Huang, Designing for appropriate reliance: The roles of ai uncertainty presentation, initial user decision, and user demographics in ai-assisted decision-making, Proc. ACM Hum.-Comput. Interact. 8 (2024). doi:10.1145/3637318.
[26] J. Li, Y. Yang, Q. V. Liao, J. Zhang, Y.-C. Lee, As confidence aligns: Exploring the effect of ai confidence on human self-confidence in human-ai decision making, 2025. doi:10.48550/ARXIV.2501.12868.
[27] D. Kahneman, Thinking, fast and slow, Psychology/economics, Allen Lane, London, 2011. Literaturangaben.
[28] D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgment under Uncertainty: Heuristics and Biases, i ed., Cambridge University Press, 1982.
[29] K. L. Mosier, L. J. Skitka, M. D. Burdick, S. T. Heers, Automation Bias, Accountability, and Verification Behaviors, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 40 (1996) 204–208. doi:10.1177/154193129604000413.
[30] K. L. Mosier, L. J. Skitka, Human Decision Makers and Automated Decision Aids: Made for Each Other?, Routledge, 1996, pp. 201–220. doi:10.1201/9781315137957.
[31] L. J. Skitka, K. Moiser, M. Burdick, Does automation bias decision-making?, International Journal of Human-Computer Studies 51 (1999) 991–1006. doi:10.1006/ijhc.1999.0252.
[32] K. L. Mosier, L. Skitka, S. Heers, M. Burdick, Automation bias: decision making and performance in high-tech cockpits, The International Journal of Aviation Psychology 8 (1997) 47–63. doi:10.1207/s15327108ijap0801_3.
[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, IUI '21, ACM, 2021. doi:10.1145/3397481.3450639.
[34] C. Rastogi, Y. Zhang, D. Wei, K. R. Varshney, A. Dhurandhar, R. Tomsett, Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making, Proc. ACM Hum.-Comput. Interact. 6 (2022) 83:1–83:22. doi:10.1145/3512930.
[35] D. Kahneman, A. Tversky, On the psychology of prediction., Psychological Review 80 (1973) 237–251. doi:10.1037/h0034747.
[36] M. Bar-Hillel, The base-rate fallacy in probability judgments, Acta Psychologica 44 (1980) 211–233. doi:10.1016/0001-6918(80)90046-3.
[37] G. Pennycook, C. Newton, V. Thompson, Base-rate neglect, second edition. ed., Routledge„ Abingdon, Oxon, 2017. Includes bibliographical references and indexes.
[38] J. J. Koehler, The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges, Behavioral and Brain Sciences 19 (1996) 1–17. doi:10.1017/s0140525x00041157.
[39] D. Mikhaylova, G. Cevolani, Automation bias and Bayesian reasoning in AI-assisted decision making, [Manuscript submitted for publication] (2025).
[40] J. Sprenger, S. Hartmann, Bayesian Philosophy of Science, Oxford University Press, 2019. doi:10.1093/oso/9780199672110.001.0001.
[41] L. Bovens, S. Hartmann, Bayesian Epistemology, Oxford University Press, Oxford, New York, 2004.
[42] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.
[43] G. Gigerenzer, U. Hoffrage, How to improve Bayesian reasoning without instruction: Frequency formats., Psychological Review 102 (1995) 684–704. doi:10.1037/0033-295x.102.4.684.
[44] D. Kahneman, Maps of Bounded Rationality: Psychology for Behavioral Economics, The American Economic Review 93 (2003) 1449–1475. doi:10.1257/000282803322655392.
[45] C. Gomez, S. M. Cho, S. Ke, C.-M. Huang, M. Unberath, Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review, 2024. doi:10.48550/arXiv.2310.19778, arXiv:2310.19778 [cs].
[46] 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) 188:1–188:21. doi:10.1145/3449287.
[47] U. Agudo, K. G. Liberal, M. Arrese, H. Matute, The impact of AI errors in a human-in-the-loop process, Cognitive research: principles and implications 9 (2024) 1–1. doi:10.1186/s41235-023-00529-3.