Accounting for Over-Reliance in AI-Assisted Performance Studies

Weina Jin[0000−0001−9253−4779] and Ghassan Hamarneh[0000−0001−5040−7448]

School of Computing Science, Simon Fraser University, Burnaby, Canada
{weinaj, hamarneh}@sfu.ca

Abstract. To evaluate the utility of artificial intelligence (AI) models in assisting humans, human-AI collaboration studies usually measure whether AI-assisted users improved their task performance compared to unassisted performance. However, such a performance gain can be explained not only by a synergy of human and AI strengths that enables appropriate reliance on AI, but also by users over-relying on a superior AI regardless of AI prediction quality. We suggest changes to human-AI collaboration studies to account for over-reliance, including replacing performance gain with complementary performance as the study endpoint; controlling, measuring, and reporting over-reliance; and disclosing the negative influence of over-reliance when discussing study results. Our critical analysis aims to improve the scientific rigor of human-AI collaboration studies and to encourage prioritizing collaborative requirements in AI development that enable users' appropriate reliance on AI.

Keywords: Human-AI collaboration · Human-centered AI · Complementary performance.

1 Introduction

Providing decision assistance and collaborating with end users is the common way to involve artificial intelligence (AI) systems in high-stakes domains, such as AI-assisted clinical decision support. In human-AI collaboration, the utility of AI to assist users is usually evaluated through human subject experiments. The study design, conduct, data analysis, and result interpretation of such evaluations should undergo critical scrutiny to ensure scientific rigor. In this work, we conduct a critical analysis of the current paradigm of evaluation studies to improve their scientific rigor. We argue that the mediator of over-reliance should be taken into account when interpreting results from AI-assisted performance gains. Our work focuses on AI assistants that perform the same task as the users, and we adopt the definition of task performance that compares the agreement of decision or prediction with some "correct" answer, using metrics such as accuracy, F1, and AUC for classification tasks, and mean squared error and coefficient of determination for scoring tasks.

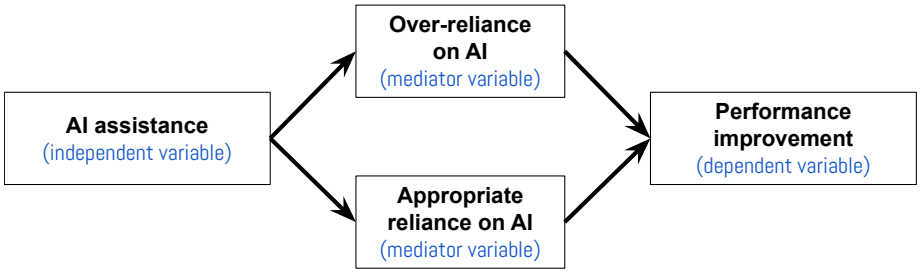

Fig. 1. Causal graph (directed acyclic graph) of the independent, dependent, and mediator variables in experiments that use AI-assisted performance improvement as the study objective. Arrows "show the direction of hypothesized causal effects." [3]

2 Over-reliance can explain away the positive effect of performance gain

To evaluate the utility of AI assistance, a typical study design paradigm is to compare the performance of the user-AI team with the baseline when users perform the task alone. If there is a statistically significant improvement in AI-assisted performance compared to the performance of unassisted users, then the AI assistance is considered to show a positive effect by helping users improve their performance [1,2]. We argue that this paradigm is flawed.

As shown in Fig. 1, such a performance gain of the human-AI team can be explained by two possible mechanisms: 1) over-reliance and 2) appropriate reliance on AI. These mechanisms are mediators that "explain the mechanism or process that underlies the relationship between an independent variable and a dependent variable" [4]. Over-reliance is defined as humans accepting both correct and incorrect predictions from AI, whereas appropriate reliance on AI is defined as humans relying on AI's predictions when AI is correct and not relying on AI's predictions when AI is incorrect [5]. Over-reliance and under-reliance hamper human-AI collaboration and need to be mitigated. In particular, the negative consequences of over-reliance include reduced human autonomy and deskilling [6,5,7].

When users are assisted by a superior AI (an AI model that has shown better performance than users on the given task), over-reliance can explain away the positive effect of the performance gain. By explaining away, we mean that while we assume the AI-assisted performance gain is from the positive effect of users' appropriate reliance on AI (i.e., the bottom path in Fig. 1), the performance gain can also be partially or fully attributed to the negative effect of over-relying on a superior AI assistant regardless of its correctness (i.e., the upper path in Fig. 1). For example, in an exam, if students are equipped with a superior AI assistant, they can simply copy whatever the AI suggests and get higher scores than their own.

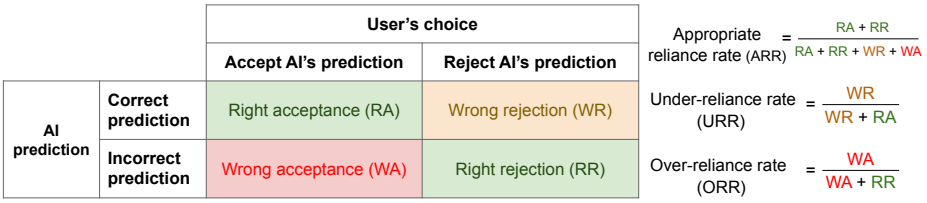

Fig. 2. A confusion matrix to show the calculation of over-reliance, under-reliance, and appropriate reliance rates. Color indicators: Red (Bad), Yellow (Warning), Green (Good).

Table 1. Metrics of appropriate reliance, over-reliance, and under-reliance rates. All the metrics are in the range $[0, 1]$. Higher number indicates a higher rate. The metrics are applicable for classification problems. ARR, ORR, and URR are aligned with the definitions of accuracy, false positive rate (FPR), and false negative rate (FNR), respectively. We use a similar definition as FPR rather than positive predictive value for ORR to make it independent of AI performance and comparable among different studies.

Metric	Definition	Meaning
Appropriate reliance rate (ARR)	$\frac{RA+RR}{RA+RR+WR+WA}$	Among AI's correctly or incorrectly predicted instances, the total ratio of instances that are accepted or rejected by users.
Over-reliance rate (ORR)	$\frac{WA}{WA+RR}$	Among AI's incorrectly predicted instances, the ratio of instances that are accepted by users.
Under-reliance rate (URR)	$\frac{WR}{WR+RA}$	Among AI's correctly predicted instances, the ratio of instances that are rejected by users.

This phenomenon is evidenced in human-AI collaboration experiments. For example, in a user study with neurosurgeons on the brain tumor classification task, assisted by an AI model with 88.0% accuracy, doctors improved their performance from their unassisted accuracy of 82.5% to 87.7% with statistical significance [8]. A fine-grained data analysis showed that with AI assistance, physicians' decision patterns converged to be more similar to AI's suggestions even when the AI was incorrect, indicating the sign of over-reliance [8]. Using our defined appropriate, over-, and under-reliance rates (ARR, ORR, and URR) in Table 1, we can calculate that both ARR and ORR increased, and URR decreased[1], showing that both appropriate reliance and over-reliance contributed to AI-assisted performance gain.

[1] The calculation is based on the numbers of instances in Fig. 4 of [8] as follows: doctor (DR) alone baseline: $ORR = 39/(39 + 50) = 39/89$; $ARR = (576 + 50)/759 = 626/759$; $URR = 94/(94+576) = 94/670$. DR+AI: $ORR = 47/(47+42) = 47/89$ (increased from baseline); $ARR = (612+42)/759 = 654/759$ (increased from baseline); $URR = 58/(58 + 612) = 58/670$ (decreased from baseline).

3 Suggested changes in human-AI collaboration studies to account for over-reliance

Due to the existence of over-reliance, AI-assisted performance gains cannot be interpreted as evidence of a purely positive effect of AI assistance. Instead, such studies should follow rigorous scientific criteria to control for alternative explanations when establishing evidence of AI utility in human-AI collaboration. We suggest the following changes to the existing paradigm of human-AI collaboration studies to improve their scientific rigor.

1. Using complementary performance, instead of performance gain, as the study endpoint. Studies using performance gain as the endpoint or objective without addressing the influence of over-reliance can mislead readers and the public to default the performance gain to AI-assisted appropriate reliance. Since the endpoint of performance gain is flawed, the study design should consider using alternative endpoints that can effectively rule out the negative effect of over-reliance. One such endpoint is complementary performance, which is defined as the user-AI team achieving better performance than either user or AI alone [9]. Achieving complementary performance is a strong indicator of appropriate reliance to mitigate over-reliance. It has been proven that the ratio of right acceptance (RA) to wrong acceptance (WA) of AI predictions needs to achieve a certain threshold to attain complementary performance [10][2].

2. Controlling, measuring, and reporting over-reliance. Since over-reliance may not be completely eliminated but can be mitigated, it is the responsibility of the researchers to control, reduce, measure, and report over-reliance in human-AI collaboration studies. For example, the design of AI systems can encourage users' skepticism; the AI system should be able to reliably indicate the limitations, scope, weaknesses, uncertainty, and evidence for its suggestions to be a good collaborator and assistant in human-AI collaboration [11]. Researchers may also utilize psychological evidence to mitigate over-reliance, such as increasing cognitive engagement [12]. But the effectiveness of these approaches may depend on specific contexts and cannot be guaranteed.

The study should include the fine-grained data analysis and reporting of over-reliance. Once the data of participants' agreement or disagreement with AI suggestions is recorded, it is straightforward to calculate our proposed metrics of ARR, ORR, and URR (Table 1), as we have done in footnote 1. Fig. 4 in [8] also provides an example to visualize participants' (dis)agreement changes due to AI assistance and over-reliance.

3. Disclosing and discussing the influence of over-reliance when interpreting study results. Lastly, if performance gain is used as the study endpoint, the discussion should explicitly address the role of over-reliance in explaining away the positive mechanism of performance gain, and disclose that

[2] Theorem 2 in [10] shows the condition with respect to $\mathbb{E}[f^r]$ and $\mathbb{E}[f^w]$ to achieve complementary performance, where $\mathbb{E}[f^r]$ and $\mathbb{E}[f^w]$ are expectations of the probability of human acceptance of AI's right and wrong predictions, respectively. The definitions of $\mathbb{E}[f^r]$ and $\mathbb{E}[f^w]$ align with RA and WA in this paper.

the performance gain cannot be attributed solely to the positive aspects of using AI, but can also reflect the negative effect of over-relying on AI.

4 Limitations and future work

The scope of this work is limited to scenarios in which the AI assistant performs the same task as the users, and users have a clear binary decision to accept or reject an AI prediction. Our scope does not cover a variety of other more complex human-AI collaborative settings in the real world, such as when the capacity of AI does not fully overlap with users' skills, or when users partially accept or reject AI predictions. Our empirical case analysis in footnote 1 is limited to one case of a binary classification task, and we did not analyze more cases from other types of tasks. Future studies can extend this critical analysis to different AI-based decision support settings. Our scope is also limited by our adoption of a narrow definition of task performance to compare the decision or prediction with some "correct" answer. We did not consider a broader definition of task performance [13], which may include metrics such as performance on fairness or task completion time.

We hope our work can inspire future works to critically examine and improve scientific rigor in human-AI collaborative study design, data analysis, and reporting. Future works can iterate on insights from these critical analyses, and incorporate them as extensions of existing study guidelines, such as SPIRIT-AI [14] and CONSORT-AI [15].

5 Conclusion and implication for the technical development of collaborative AI

In this work, we emphasize that over-reliance can explain away the positive effect of AI assistance in performance gain. We propose metrics to measure appropriate reliance, over-reliance, and under-reliance, and demonstrate the existence of over-reliance as a mediator in a case study of AI-assisted performance gain. We suggest changes in the study design, data analysis, reporting, and interpretation of human-AI collaboration studies to improve their scientific rigor. Shifting the study endpoint from performance gain to complementary human-AI performance fundamentally redefines AI development objectives. Instead of prioritizing outperforming humans—where collaboration is merely an afterthought—prioritizing complementary performance requires embedding collaborative requirements directly into the blueprint of AI models. This means to develop AI as decision support in high-stakes domains, the AI technical development priorities could include the following:

1. The AI development needs to prioritize the objective to reliably indicate the weaknesses and uncertainty of AI models' predictions. This objective, which may be achieved through the development of explainable AI and uncertainty estimation techniques, provides the prerequisite that enables users' appropriate reliance and complementary performance [16,10,11].

2. Being able to effectively collaborate with users and improve complementary performance means that the AI development needs to incorporate understandings of technical limitations [17], human factors [18,19,20,21], and sociotechnical factors [22,23,24,25] into technical specifications from the outset in AI development [26]. To give an example of this line of effort, Jin et al. [27] incorporated clinical users' inputs and requirements for explainable AI techniques as the five evaluation criteria for the development of potentially clinical viable explainable AI techniques.

Acknowledgments. We thank Kumar Abhishek and Mark Nielsen for their valuable discussions. We thank the anonymous peer reviewers for their constructive suggestions. This project is supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant RGPIN-2020-06752.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Zi-Hao Bo, Hui Qiao, Chong Tian, Yuchen Guo, Wuchao Li, Tiantian Liang, Dongxue Li, Dan Liao, Xianchun Zeng, Leilei Mei, Tianliang Shi, Bo Wu, Chao Huang, Lu Liu, Can Jin, Qiping Guo, Jun-Hai Yong, Feng Xu, Tijiang Zhang, Rongpin Wang, and Qionghai Dai. Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network. *Patterns*, 2(2):100197, February 2021.
2. Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael 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*, AIES '22, page 617–626, New York, NY, USA, 2022. Association for Computing Machinery.
3. Ari M. Lipsky and Sander Greenland. Causal directed acyclic graphs. *JAMA*, 327(11):1083–1084, 03 2022.
4. Wikipedia. Mediation (statistics). https://en.wikipedia.org/wiki/Mediation_(statistics), June 2026. [Online; accessed 22-June-2026].
5. Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), April 2023.
6. Krzysztof Budzyń, Marcin Romańczyk, Diana Kitala, Paweł Kołodziej, Marek Bugajski, Hans O Adami, Johannes Blom, Marek Buszkiewicz, Natalie Halvorsen, Cesare Hassan, Tomasz Romańczyk, Øyvind Holme, Krzysztof Jarus, Shona Fielding, Melina Kunar, Maria Pellise, Nastazja Pilonis, Michał Filip Kamiński, Mette Kalager, Michael Bretthauer, and Yuichi Mori. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903, 2025.
7. Arvind Narayanan and Sayash Kapoor. Why an overreliance on AI-driven modelling is bad for science. *Nature*, 640(8058):312–314, April 2025.

8. Weina Jin, Mostafa Fatehi, Ru Guo, and Ghassan Hamarneh. Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. *Artificial Intelligence in Medicine*, 148:102751, 2024.
9. Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, October 2024.
10. Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Why is plausibility surprisingly problematic as an XAI criterion? arXiv:2303.17707, 2025.
11. Amin Shirangi, Todd Pittinsky, and Judy Gichoya. Rethinking Human-AI Collaboration in Radiology. *Radiology*, 318(2):e252760, 2026. PMID: 41701018.
12. Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.
13. Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA, 2022. Association for Computing Machinery.
14. Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, Melanie J. Calvert, Ara Darzi, Christopher Holmes, Christopher Yau, David Moher, Hutan Ashrafian, Jonathan J. Deeks, Lavinia Ferrante di Ruffano, Livia Faes, Pearse A. Keane, Sebastian J. Vollmer, Aaron Y. Lee, Adrian Jonas, Andre Esteva, Andrew L. Beam, Maria Beatrice Panico, Cecilia S. Lee, Charlotte Haug, Christophe J. Kelly, Christopher Yau, Cynthia Mulrow, Cyrus Espinoza, John Fletcher, David Moher, Dina Paltoo, Elaine Manna, Gary Price, Gary S. Collins, Hugh Harvey, James Matcham, Joao Monteiro, M. Khair ElZarrad, Lavinia Ferrante di Ruffano, Luke Oakden-Rayner, Melissa McCradden, Pearse A. Keane, Richard Savage, Robert Golub, Rupa Sarkar, and Samuel Rowley. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine*, 26(9):1351–1363, 2020.
15. Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston, An-Wen Chan, Ara Darzi, Christopher Holmes, Christopher Yau, Hutan Ashrafian, Jonathan J. Deeks, Lavinia Ferrante di Ruffano, Livia Faes, Pearse A. Keane, Sebastian J. Vollmer, Aaron Y. Lee, Adrian Jonas, Andre Esteva, Andrew L. Beam, An-Wen Chan, Maria Beatrice Panico, Cecilia S. Lee, Charlotte Haug, Christopher J. Kelly, Christopher Yau, Cynthia Mulrow, Cyrus Espinoza, John Fletcher, Dina Paltoo, Elaine Manna, Gary Price, Gary S. Collins, Hugh Harvey, James Matcham, Joao Monteiro, M. Khair ElZarrad, Lavinia Ferrante di Ruffano, Luke Oakden-Rayner, Melissa McCradden, Pearse A. Keane, Richard Savage, Robert Golub, Rupa Sarkar, and Samuel Rowley. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, 26(9):1364–1374, 2020.
16. Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.
17. Weina Jin, Ashish Sinha, Kumar Abhishek, and Ghassan Hamarneh. Ethical Medical Image Synthesis. arXiv:2508.09293, 2025.
18. Joshua Strong, Harry Rogers, Emma Sun, Anna Louise Todsen, Jody Ede, Cherry Lumley, Nick Yeung, Helen Higham, and J. Alison Noble. Human-AI Collaboration in Healthcare: A Scoping Review. *npj Digital Medicine*, 2026.

19. Thomas Dratsch, Xue Chen, Mohammad Rezazade Mehrizi, Roman Kloeckner, Aline Mähringer-Kunz, Michael Püsken, Bettina Baeßler, Stephanie Sauer, David Maintz, and Daniel Pinto dos Santos. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*, 307(4):e222176, 2023. PMID: 37129490.
20. Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, 169:102941, 2023.
21. Emely Rosbach, Jonathan Ganz, Jonas Ammeling, Andreas Riener, and Marc Aubreville. Automation Bias in AI-assisted Medical Decision-making under Time Pressure in Computational Pathology. In Christoph Palm, Katharina Breininger, Thomas Deserno, Heinz Handels, Andreas Maier, Klaus H. Maier-Hein, and Thomas M. Tolxdorff, editors, *Bildverarbeitung für die Medizin 2025*, pages 129–134, Wiesbaden, 2025. Springer Fachmedien Wiesbaden.
22. Weina Jin, Elise Li Zheng, and Ghassan Hamarneh. Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice. arXiv:2510.10588, 2025.
23. Christian Herzog and Jason Branford. Relational Ethics and Structural Epistemic Injustice of AI in Medicine. *Philosophy and Technology*, 38(4), November 2025.
24. Margarita Boenig-Liptsin, Anissa Tanweer, and Ari Edmundson. Data science ethos lifecycle: Interplay of ethical thinking and data science practice. *Journal of Statistics and Data Science Education*, 30(3):228–240, July 2022.
25. Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.
26. Weina Jin, Nicholas Vincent, and Ghassan Hamarneh. AI for Just Work: Constructing Diverse Imaginations of AI beyond "Replacing Humans". arXiv:2503.08720, 2025.
27. Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical Image Analysis*, 84:102684, 2023.