



Human learning is an understudied but promising lever for boosting human–AI synergy

Julian Berger^{a,b,c,d,1}, Jason W. Burton^{a,e,f}, Ralph Hertwig^a, Thomas Kosch^g, Ralf H. J. M. Kurvers^a, Benito Kurzenberger^h, Christopher Lazik^g, Linda Onnasch^h, Tobias Rieger^h, Anna I. Thoma^a, Dirk U. Wulff^{a,j}, and Stefan M. Herzog^{a,1}

Edited by David L. Donoho, Stanford University, Stanford, CA; received January 5, 2026; accepted August 19, 2026

Humans collaborating with AI hold the promise of achieving superior outcomes compared to either acting alone (i.e., human–AI synergy). However, the conditions that facilitate such synergy when humans are advised by AI are not well understood. A recent meta-analysis showed that, on average, human–AI combinations do not outperform the better individual agent. We argue that this pessimistic conclusion arises from insufficient attention to human learning in experimental designs. To substantiate this claim, we reanalyzed all 74 studies included in the original meta-analysis and found that most previous research overlooked design features that foster human learning (e.g., outcome feedback to participants). Our reanalysis further revealed that studies providing outcome feedback show tentatively higher synergy than those without outcome feedback. Crucially, feedback paired with AI explanations was associated with positive synergy, while explanations without feedback were associated with negative synergy—suggesting that explanations improve synergy mainly when humans can learn to verify the AI's reliability through feedback. Our reanalysis suggests that the current literature underestimates the potential of human–AI collaboration because it predominantly relies on paradigms that do not facilitate human learning, thus hindering humans from effectively adapting their collaboration strategies. However, experiments directly varying learning opportunities are needed for stronger, causal conclusions. We advocate for a paradigm shift in human–AI interaction research that explicitly addresses human learning and thus enhances our understanding of and support for successful human–AI collaboration.

human–AI collaboration | synergy | learning | feedback | AI explanations

Humans collaborating with AI have the potential to achieve better outcomes than either independently, a phenomenon called human–AI synergy (Fig. 1A). Synergy is well documented whenever independent human and AI judgments are combined statistically (1). However, when and why this synergy occurs when humans interact with AI output is currently not well understood (2), although decision-theoretic analyses have characterized the bounds of possible synergy (3). Vaccaro et al. (4) recently conducted a systematic review and meta-analysis of 74 studies comparing the performance of humans, AI, and human–AI combinations where humans needed to interact with AI systems. The studies span a wide range of tasks, such as classifying images, predicting house prices, or writing code. On average, human–AI combinations performed worse than the best of humans or AI alone (i.e., negative human–AI synergy). Positive synergy emerged when humans outperformed AI alone, but synergy was negative when AI outperformed humans alone. Other moderators could not explain differences across studies. For instance, neither AI confidence nor AI explanations (helping humans understand how a model arrived at its output) led to human–AI synergy. We argue that human learning is a crucial, but overlooked facilitator of synergy (Fig. 1B and C). Reanalyzing all studies covered in Vaccaro et al. (4), we show that the few studies that gave humans feedback tended to report positive synergy—especially when AI explanations were also provided. We argue that neglecting human learning leads to underestimating the potential for human–AI synergy and therefore call for a greater focus on human learning to better understand and foster successful human–AI collaboration.

Conclusions about human performance and rationality drastically depend on how a phenomenon is studied (5, 6). For example, when people can experience the probabilistic structure of a task environment through repeated interactions, their decisions are close to what probability theory would suggest whereas this is not so when participants are merely given textual descriptions of the task (6). Could Vaccaro et al.'s (4) findings of negative synergy be partly due to the way human–AI collaboration is typically studied?

Author affiliations: ^aCenter for Adaptive Rationality, Max Planck Institute for Human Development, Berlin 14195, Germany; ^bDepartment of Business and Management, University of Southern Denmark, Odense 5230, Denmark; ^cDepartment of Economics and Business, Universitat Pompeu Fabra, Barcelona 08005, Spain; ^dBarcelona School of Economics, Barcelona 08005, Spain; ^eDepartment of Psychology, University of Copenhagen, Copenhagen 1353, Denmark; ^fCopenhagen Center for Social Data Science, University of Copenhagen, Copenhagen 1353, Denmark; ^gDepartment of Computer Science, Humboldt-Universität zu Berlin, Berlin 12489, Germany; ^hDepartment of Psychology & Ergonomics, Technische Universität Berlin, Berlin 10587, Germany; ⁱDepartment of Psychology, University of Basel, Basel 4055, Switzerland; and ^jDepartment of Business Analytics and Decision Sciences, Vienna University of Economics and Business, Vienna 1020, Austria

Author contributions: J.B., J.W.B., R.H., T.K., R.H.J.M.K., B.K., C.L., L.O., T.R., A.I.T., D.U.W., and S.M.H. designed research; J.B., J.W.B., R.H., T.K., R.H.J.M.K., B.K., C.L., L.O., T.R., A.I.T., D.U.W., and S.M.H. performed research; J.B., J.W.B., R.H., T.K., R.H.J.M.K., B.K., C.L., L.O., T.R., A.I.T., D.U.W., and S.M.H. contributed new reagents/analytic tools; J.B., J.W.B., R.H., T.K., R.H.J.M.K., B.K., C.L., L.O., T.R., A.I.T., D.U.W., and S.M.H. analyzed data; and J.B., J.W.B., R.H., R.H.J.M.K., L.O., T.R., A.I.T., D.U.W., and S.M.H. wrote the paper.

The authors declare no competing interest.

Copyright © 2026 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

¹To whom correspondence may be addressed. Email: berger@mpib-berlin.mpg.de or herzog@mpib-berlin.mpg.de.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2536100123/-DCSupplemental>.

Published September 21, 2026.

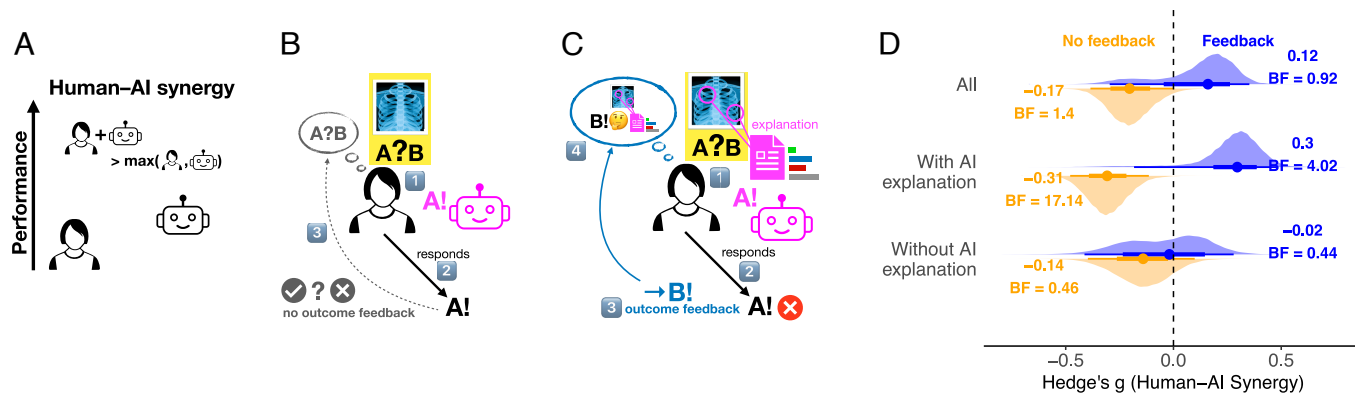


Fig. 1. (A) Positive human–AI synergy occurs when the joint system outperforms either agent alone. (B) Human–AI interaction without learning opportunities: The human knows the AI's choice but receives no outcome feedback. (C) With learning opportunities: The human sees the AI's choice and its explanation, receives outcome feedback, and across trials learns about their own and the AI's relative skill. (D) Robust Bayesian model-averaged meta-regression analyses (12) quantifying evidence that outcome feedback moderates human–AI synergy in the studies reviewed by Vaccaro et al. (4). X-axis shows conditional marginal posterior estimates of Hedges' g for studies with and without feedback. Dots and whiskers show the posterior median and 66%/95% credible intervals. Bayes factors (BF) quantify evidence that the estimate is nonzero versus zero. Results are presented for all studies, and separately for studies with and without AI explanations.

Learning, experience, and feedback are crucial for successful human decision-making (6, 7). For example, observing outcome feedback (i.e., the correct response after each case) helps people identify reliable sources of high-quality information (8). It can also help humans assess the accuracy of their own decisions and of AI advice, and how best to integrate them into their final judgments (9). Also, guiding humans' attention to fewer, relevant task features (e.g., a subset of medical diagnostic cues) and indicating how to use those features can further improve human judgments (7). Explainable AI may likewise help humans focus on relevant task features—and even indicate how the AI uses those features—and thus facilitate human learning (10). Such guidance should be particularly useful when humans and AI excel at different kinds of cases and humans need to learn to adaptively rely on their own judgment or AI's assistance (11). AI explanations, expressions of AI confidence and other meta-information can thus serve as additional task features on top of humans' own perceived confidence that can help humans develop adaptive strategies for when (not) to rely on AI (2).

To investigate how much human learning is recognized in the literature on synergy, we reanalyzed the studies reviewed by Vaccaro et al. (4), quantifying design factors likely to benefit or hinder learning (e.g., provision of feedback or practice trials; coded separately for each experimental condition, see *SI Appendix*). The typical study involved a single session with no practice trials, a median of 30 task trials and no outcome feedback—lack features known to facilitate learning. Focusing on the 10 (14%) of 74 studies that included experimental conditions with outcome feedback (providing 98 [26%] of 370 effect sizes, see *SI Appendix, Extended Methods* for details), only three studied how humans' use of AI assistance evolved across trials; the rest reported average performance across trials, potentially masking how humans adapt to the task based on feedback. None of the studies experimentally varied whether outcome feedback was provided or not.

To test whether outcome feedback—especially with AI explanations—fosters human–AI synergy, we ran robust Bayesian meta-regressions (12) modeling synergy based on whether outcome feedback and AI explanations were provided (*SI Appendix*). We quantify synergy using Hedges' g effect size to measure the standardized difference in performance between the human–AI combination and whichever performed better alone. That

is, positive (negative) estimates indicate that the human–AI combination performed better (worse) than either humans or AI alone. Comparing studies with (98 effect sizes) and without feedback (272; Fig. 1D), provided tentative evidence that studies with feedback reported higher synergy ($BF_{inclusion} = 2.36$). Without feedback, synergy tended to be negative (Hedges' $g = -0.17$, 95% credible intervals [CI] -0.37 to 0.00 , $BF_{inclusion} = 1.4$) while positive with feedback ($g = 0.12$, 95% CI -0.28 to 0.34 , $BF_{inclusion} = 0.92$). While neither estimate was credibly different from zero, there was tentative evidence that synergy was higher in studies with than without feedback (contrast: $g = 0.34$; 95% CI -0.01 to 0.69 , posterior probability of a positive direction: 84%).

Next, we investigated whether outcome feedback increased synergy when combined with AI explanations. For studies with explanations (Fig. 1D), feedback was a strong positive moderator of synergy ($BF_{inclusion} = 21$). When people had access to both explanations and feedback (24 effect sizes, 14%), synergy tended to be positive ($g = 0.3$, 95% CI -0.18 to 0.48 , $BF_{inclusion} = 4.02$); when they had explanations but not feedback (148), synergy was clearly negative ($g = -0.31$, 95% CI -0.48 to -0.11 , $BF_{inclusion} = 17.14$). With explanations, providing feedback was associated with increased synergy (contrast: Hedges' $g = 0.60$, 95% CI 0 to 0.94 , PD = 97%); without explanations, feedback did not moderate synergy ($BF_{inclusion} = 0.59$) (Fig. 1D). This result indicates that AI explanations increase synergy mainly when humans can learn to verify the AI's reliability. To assess whether study characteristics confounded our analyses, we ran additional meta-regressions that included each moderator considered by Vaccaro et al. alongside feedback (see <https://osf.io/mqw4y>). The effect directions remained consistent across all 12 models.

In sum, our reanalysis of Vaccaro et al. (4) showed that the studies largely neglect human learning. This is consistent with Lai et al. (13) who surveyed 124 studies on human–AI decision-making (after 2018 focused on human–computer interaction; HCI) and found that only 6 (5%) of studies offered outcome feedback. To better assess the role of human learning in human–AI collaboration through evidence synthesis, we make three recommendations: First, we echo Vaccaro et al.'s (4) call to report standardized performance metrics, even when it is not the study's primary outcome. Second, we call for sharing participant- and trial-level data, which would allow more directly studying how

humans learn to use AI; we note a dearth of publicly available datasets. Third, broader search terms and including studies prior to 2018 would better represent further relevant fields such as human factors (11) and judgment and decision-making (14).

Three caveats: First, the usefulness of feedback is constrained by the task structure and learning environment (15, 16): When feedback is accurate, timely, and predictive task features are available, learning is facilitated; when feedback is noisy, delayed, or predictive task features are missing or opaque, learning is obstructed (15). As researchers design the task environment—for example, by (not) providing trial-by-trial feedback and AI explanations that guide attention to relevant features—the extent to which one observes human–AI synergy will depend on those choices. Second, most studies only report average performance across trials, thus our between-study analyses cannot capture potential within-study improvements through feedback. Meta-analyzing joint performance in later trials—after participants have had a chance to learn—would better reflect the achieved learning. Thus, our results may actually represent a conservative estimate of how learning may foster synergy. We investigated such within-study learning in three studies where trial-level data was available (see <https://osf.io/mqw4y>) and found that participants learned to align with better-performing AI over trials when informative signals were available, but not otherwise. Finally, our results are correlational and experiments directly varying learning opportunities are needed for stronger, causal conclusions.

Within the design space of possible human–AI decision-making experiments, providing feedback or other learning opportunities is just one of many design choices. Task selection, timing, and format of AI advice, incentives, participant populations, and evaluation metrics all jointly define the region of the design space from which data is collected and conclusions are drawn (5). Hullman et al. (16) argue that conclusions about human bias in AI advice taking are overly pessimistic and unwarranted because HCI studies—including some covered in Vaccaro et al. (4)—do not allow participants, or even idealized rational decision makers, to identify the correct response. The pessimistic view of human–AI synergy (4) may thus reflect sampling from a learning-handicapping part of the design space rather than

the frontier of possible human–AI synergy. Going forward, we therefore advocate for a systematic and theory-informed exploration of the design space (5, 16), including experimentally varying the presence and type of feedback (17) and the task structure (7, 15).

In sum, our reanalysis suggests that supporting human learning—particularly by providing outcome feedback and AI explanations—is an underappreciated and understudied but promising lever for increasing human–AI synergy. Considering the broader literature on human learning (6–8), boosting competences (10, 18), and the importance of study design (5, 6, 16), these meta-analytic results thus paint a more optimistic picture and highlight the need to explicitly and experimentally study when and how fostering human learning boosts human–AI synergy.

Materials and Methods

Data. We use open data from ref. 4 for coding of effect sizes and AI explanations of 370 experimental conditions in the 74 studies analyzed. Other features (e.g., feedback) were coded by independent systematic review of every study by two of six authors (J.B., J.W.B., T.K., B.K., C.L., and T.R.), with disagreements arbitrated by J.B. (see *SI Appendix* for coding scheme).

Meta-Regression Analyses. All meta-regression models were estimated using RoBMA version 3.5. (12). We quantify human–AI synergy using Hedges' *g* to measure the standardized difference in performance between the human–AI combination and whichever performed better alone. Positive (negative) estimates indicate that the human–AI combination performed better (worse) than either humans or AI alone. For more details on RoBMA models see *SI Appendix*.

Data, Materials, and Software Availability. Data and code data have been deposited in OSF (19).

ACKNOWLEDGMENTS. This work was supported by the European Union's Horizon Europe program (project Hybrid Human Artificial Collective Intelligence in Open-Ended Decision Making (HACID); GA 101070588) and by the Deutsche Forschungsgemeinschaft (German Research Foundation)—grant HE 2768/11-1. We thank Susannah Goss for editing the manuscript. Open access funding provided by the Max Planck Society.

1. M. Steyvers, H. Tejada, G. Kerrigan, P. Smyth, Bayesian modeling of human-AI complementarity. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2111547119 (2022).
2. M. Steyvers, A. Kumar, Three challenges for AI-assisted decision-making. *Perspect. Psychol. Sci.* **19**, 722–734 (2024).
3. Z. Guo, Y. Wu, J. D. Hartline, J. Hullman, "A decision theoretic framework for measuring AI reliance" in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, New York, NY, USA, 2024), pp. 221–236.
4. M. Vaccaro, A. Almaatouq, T. Malone, When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat. Hum. Behav.* **8**, 2293–2303 (2024).
5. A. Almaatouq et al., Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.* **47**, e33 (2024).
6. T. Lejarraga, R. Hertwig, How experimental methods shaped views on human competence and rationality. *Psychol. Bull.* **147**, 535–564 (2021).
7. N. Karelaia, R. M. Hogarth, Determinants of linear judgment: A meta-analysis of lens model studies. *Psychol. Bull.* **134**, 404–426 (2008).
8. A. Novaes Tump, M. Wolf, J. Krause, R. H. J. M. Kurvers, Individuals fail to reap the collective benefits of diversity because of over-reliance on personal information. *J. R. Soc. Interface* **15**, 20180155 (2018).
9. Z. Z. Jiang, L. Argote, P. V. Singh, Learning in human-AI collaboration (2025).
10. S. M. Herzog, M. Franklin, Boosting human competences with interpretable and explainable artificial intelligence. *Decision* **11**, 493–510 (2024).
11. T. Rieger, H. Schindler, L. Onnasch, E. Roesler, Explaining AI weaknesses improves human-AI performance in a dynamic control task. *Int. J. Hum. Comput. Stud.* **199**, 103505 (2025).
12. F. Bartoň, M. Maier, T. D. Stanley, E. J. Wagenmakers, Robust Bayesian meta-regression: Model-averaged moderation analysis in the presence of publication bias. *Psychol. Methods*, <https://doi.org/10.1037/met0000737> (2025).
13. V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, C. Tan, "Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies" in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23* (Association for Computing Machinery, New York, NY, USA, 2023), pp. 1369–1385.
14. J. W. Burton, M. Stein, T. B. Jensen, A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Mak.* **33**, 220–239 (2020).
15. R. M. Hogarth, T. Lejarraga, E. Soyer, The two settings of kind and wicked learning environments. *Curr. Dir. Psychol. Sci.* **24**, 379–385 (2015).
16. J. Hullman, A. Kale, J. Hartline, "Underspecified human decision experiments considered harmful" in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, B. Lee, M. Chetty, P. O. T. Dugas, Eds. (Association for Computing Machinery, 2025).
17. D. Ahn, A. Almaatouq, M. Gulabani, K. Hosanagar, "Impact of model interpretability and outcome feedback on trust in AI" in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, M. L. Wilson, P. T. Dugas, I. Shklovski, Eds. (ACM, Honolulu, HI, USA, 2024), pp. 1–25.
18. S. M. Herzog, R. Hertwig, Boosting: Empowering citizens with behavioral science. *Annu. Rev. Psychol.* **76**, 851–881 (2025).
19. J. Berger et al., Fostering human learning is crucial for boosting human-AI synergy. OSF. https://osf.io/mqw4y/overview?view_only=c0daf88ca7444d9abf7a79e2fe0a33b. Deposited 7 August 2025.