Revisiting the ABCs of Working with AI: A Replication with Radiologists*

Daniel Martin†

May 14, 2026

Abstract

Artificial intelligence (AI) systems increasingly assist human experts, but the consequences of AI assistance on productivity can be uneven. Caplin, Deming, S. Li, Martin, Marx, Weidmann, and Ye (2025b) show that two characteristics, ability and belief calibration, jointly determine the returns to AI assistance. This paper shows that their results replicate to a setting where professional radiologists analyze chest X-rays with access to state-of-the-art machine learning predictions. I leverage the public Collab-CXR data repository described by Moehring, Kutwal, Huang, Banerjee, Jacobi, Eber, Mendoza, Chung, Dayan, Gupta, Bui, Truong, Pareek, Langlotz, Lungren, Agarwal, Rajpurkar, and Salz (2025) and first analyzed for human-AI collaboration by Agarwal, Moehring, Rajpurkar, and Salz (2023). To faithfully reproduce the analysis in Caplin, Deming, S. Li, Martin, Marx, Weidmann, and Ye (2025b), I restrict attention to the paired-image sessions, which include 35 radiologists and 3,500 radiologist-patient-pathology readings observed both with and without AI assistance. The results of this replication support the external validity of their core findings: lower baseline ability predicts higher incremental value from AI, higher calibration predicts more incremental value from AI, and AI strongly reduces productive inequality when calibration is high.

*I thank the Sloan Foundation for support under the "Cognitive Economics at Work" grant.
†Department of Economics, University of California, Santa Barbara.

1 Introduction

Artificial intelligence (AI) systems are being considered for a wide array of professional settings. A natural first question is whether AI raises decision quality *on average*. For economic and policy purposes, however, the distribution of gains is equally important. For instance, recent work has shown that AI can narrow performance gaps by boosting lower-performing users (Brynjolfsson, D. Li, and Raymond 2023; Noy and W. Zhang 2023; Autor 2024).

However, AI may leave disparities unchanged if users cannot recognize when to follow AI advice (and when not to). In a controlled image-classification experiment, Caplin, Deming, S. Li, Martin, Marx, Weidmann, and Ye (2025b) (CDLMMWY hereafter) find that AI assistance is most valuable for users with low baseline ability and well-calibrated beliefs about their own performance. Calibration is likely to matter because a user must decide when to trust their own judgment and when to defer to the algorithm.

The simplicity of this logic suggests that these results about ability and belief calibration (the "ABCs" of working with AI) might hold more generally. However, the task in CDLMMWY is an age guessing task (a "Bouncer" task[1] because the goal is to guess whether a person is over or under 21) and participants are Prolific workers, so there might be concerns about whether these results extend to other settings, particularly those that feature traditional work tasks and workers who are experienced at the task.

To help address this concern, I examine whether ability and belief calibration also matter in the high-expertise field of radiology. I leverage data from the radiology experiment of Agarwal, Moehring, Rajpurkar, and Salz (2023) (AMRS hereafter), which asks professional radiologists to report diagnostic probabilities for chest X-rays with and without AI support. In addition to having professional experts assessing medically meaningful images, this setting offers a strong domain-specific AI benchmark. The resulting dataset contains probabilistic reports, AI predictions, timing, clickstream, and diagnostic standard variables (Moehring, Kutwal, Huang, Banerjee, Jacobi, Eber, Mendoza, Chung, Dayan, Gupta, Bui, Truong, Pareek, Langlotz, Lungren, Agarwal, Rajpurkar, and Salz 2025).

By moving from an age-guessing experiment to a professional radiology experiment, I am able to provide a high-expertise test of the CDLMMWY results. While the sample size of expert radiologists is naturally smaller, the findings provide evidence consistent with the qualitative direction of the original framework: lower baseline ability predicts higher incremental value from AI assistance, and higher belief calibration predicts greater gains from AI. These results suggest that the behavioral mechanisms identified by CDLMMWY may extend beyond the original context to professional diagnostic work.

[1] I thank Jason Somerville for proposing this task name.

This paper responds to recent calls for more external-validity evidence on human-AI collaboration beyond controlled laboratory tasks. Recent work calls for in-context studies as human-AI teaming moves from controlled settings toward applied, higher-stakes domains, and medical-AI studies similarly emphasize that benefits depend on expertise, interaction context, task performance, and user experience (Gonzalez, Donahue, Goldstein, Heidari, Jalali, Schelble, Singh, and Woolley 2026; Kargarnovin, Hernandez, Reiners, Cruz-Neira, Bochenek, and Karwowski 2026; Liu, J. Zhang, Shuaiqi Chen, and Shanguang Chen 2025; Wekenborg, Gilbert, and Kather 2025).

2 Data and Measures

In constructing the replication sample, I start from the AMRS top-level diagnostic analysis sample, which contains visible submitted reports for the two top-level diagnostic pathologies with AI predictions and US diagnostic-standard labels. Following AMRS, I drop warmup rows and rows from the radiologists who generated the diagnostic-standard labels, and I exclude aggregate/administrative categories.

Next, to faithfully reproduce the CDLMMWY analysis, it was necessary to construct a data set that allows ability and calibration to be estimated on one fixed block of cases and the impact of AI on productivity to be estimated on another fixed block of cases.[2] CDLMMWY accomplish this by designing an experiment where all subjects complete one fixed block of cases without AI and then randomizing subjects to complete a second fixed block of cases with or without AI. To accomplish this with the radiologist data, I use the paired-image sessions of the AMRS experiment, which correspond to their Design 3. The feature of Design 3 that makes it especially suitable for the analysis of CDLMMWY is that it repeats the exact same radiologist-patient-pathology cases with and without AI, whereas Designs 1 and 2 assign different cases randomly across AI blocks. I assign half of these cases to a "skill block" and the other half to an "outcome block". I then estimate ability and calibration on the skill block cases without AI and compare performance with and without AI in the outcome block. Figure 1 summarizes the mapping between the two data sets and the CDLMMWY analysis.

To mirror the CDLMMWY image assignment procedure (for assigning images to the skill or outcome blocks), I construct a patient-level split rather than using a purely random split. The split is global across radiologists and is made at the patient-image level, so all focal pathologies and all radiologists' readings for a patient remain in the same block. Patients are sorted lexicographically by the focal pathology true labels and AI scores, adjacent patients

[2]The data and code for CDLMMWY are distributed with their *Management Science* supplemental materials (Caplin, Deming, S. Li, Martin, Marx, Weidmann, and Ye 2025a).

Figure 1: Comparison between the original CDLMMWY design and the replication. Panel A shows the original CDLMMWY structure, where baseline ability and calibration are measured in a no-AI skill block and AI assistance is randomized in the outcome block. Panel B shows the replication structure, where the same radiologist-case-pathology readings are observed without and with AI, and there is a patient-level split into skill and outcome blocks.

are paired, and one patient from each pair is assigned to the outcome block. I search seeded random draws within these adjacent pairs and choose a split with 161 skill-block patients and 161 outcome-block patients. The selected split has zero monotonicity violations in the outcome-block AI calibration curve, a 0.01 percentage point difference in no-AI prior accuracy across blocks, and a 0.14 percentage point difference in no-AI prior confidence across blocks.

The resulting Design 3 data set is organized at the radiologist-patient-pathology level. Let B_{hi} denote radiologist h's probability report for case i, S_i the diagnostic-standard truth, and M_i the AI prediction. Reported beliefs are "correct" if they would have induced correct choices at a decision threshold of 50%:

$$R_{hi} = \begin{cases} 1 & \text{if } |B_{hi} - S_i| < 0.5 \\ 0.5 & \text{if } B_{hi} = 0.5 \\ 0 & \text{else,} \end{cases}$$

Accuracy is then the average correctness. For calibration, the confidence of beliefs over each case is $K_{hi} = \max\{B_{hi}, 1 - B_{hi}\}$, and net confidence is the expected difference between their confidence and correctness. Finally, calibration is the negative absolute value of their net confidence. Figure 2 compares the accuracy distributions across CDLMMWY and AMRS, and Figure 3 compares the corresponding net-confidence distributions.

3 Empirical Specification

Accuracy and calibration in the no-AI cases in the skill block provide the estimated baseline traits: A_h and C_h, respectively. Accuracy in the outcome block then supplies no-AI performance $Y_h(0)$ and AI-assisted performance $Y_h(1)$. Following CDLMMWY, the regression specification is:

$$Y_{ht} = \alpha_0 + \alpha_1 A_h + \alpha_2 C_h + \beta_0 AI_t + \beta_1 (A_h \times AI_t) + \beta_2 (C_h \times AI_t) + \varepsilon_{ht} \tag{1}$$

where $AI_t = 0$ corresponds to no AI, $AI_t = 1$ corresponds to AI-assisted, and A_h and C_h are standardized within the sample before estimation. Standard errors are clustered by radiologist, using the finite-sample correction implemented in `statsmodels`. Because the AMRS data do not include an IQ analogue, my primary specification corresponds to the model in column 3 of Table 2 in CDLMMWY.

Figure 5 uses the equivalent radiologist-level improvement, $Y_h(1) - Y_h(0)$ on A_h and C_h, to report fitted AI gains at plus or minus one standard deviation of baseline ability and calibration. Figure 6 uses the fitted model to compare the interquartile range of predicted accuracy without AI, with AI, and under the counterfactual that baseline miscalibration is set to zero.

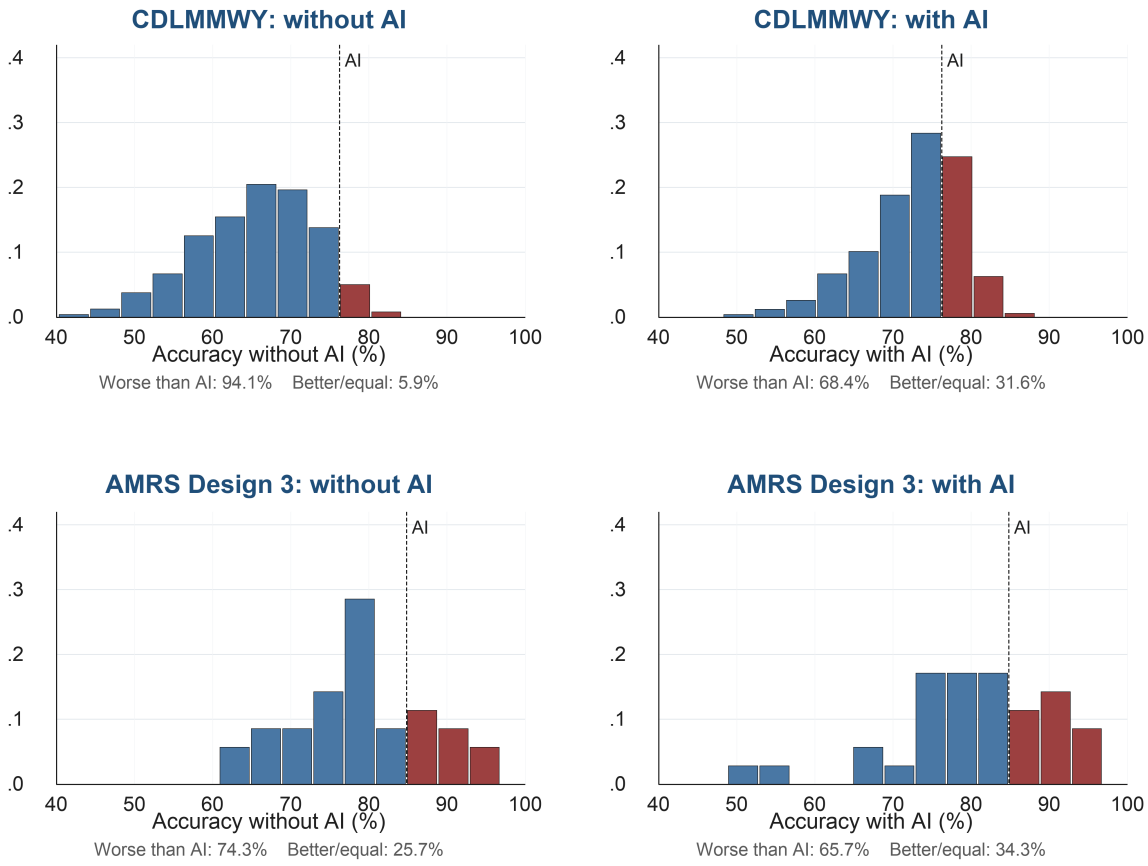

Figure 2: Accuracy distributions across studies. Each panel shows subject-level mean outcome-block accuracy, separately without and with AI. The dashed vertical line marks the AI-alone accuracy benchmark in the corresponding study.

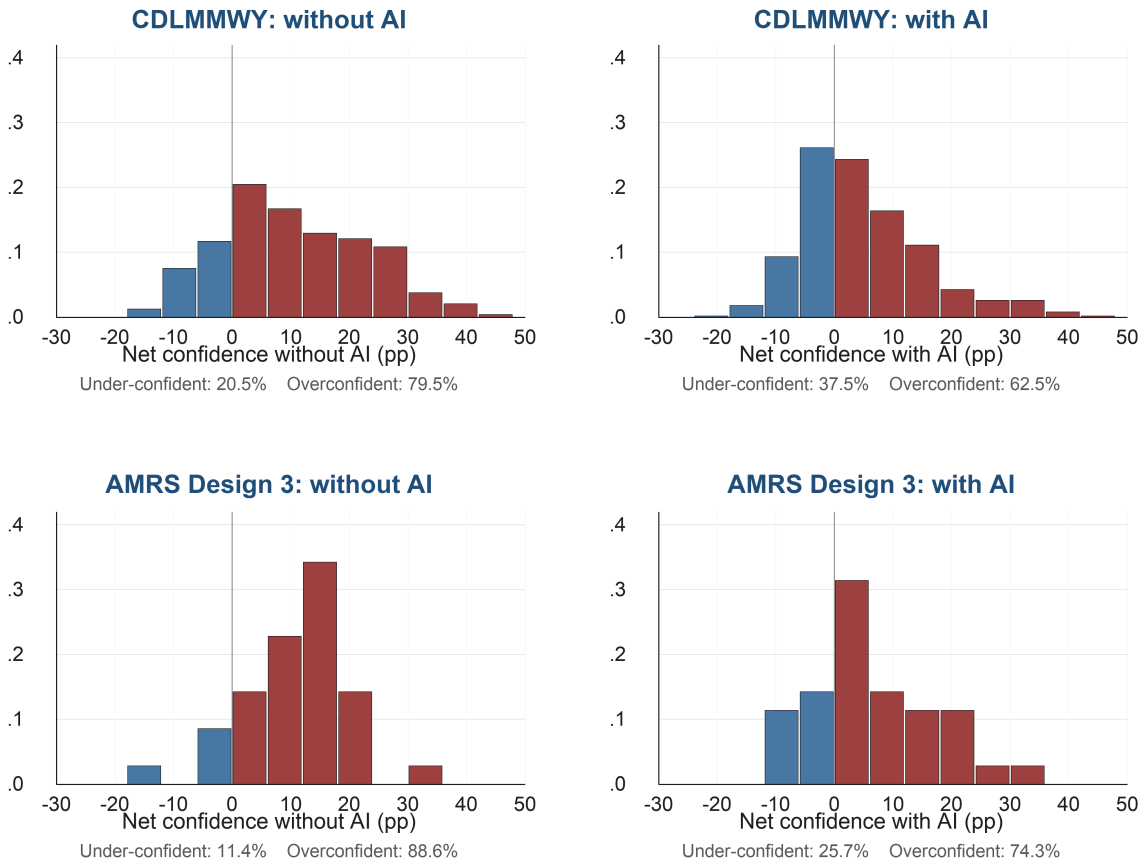

Figure 3: Net confidence distributions across studies. Net confidence is confidence minus correctness, so values closer to zero correspond to better calibration, and positive values indicate overconfidence and negative values indicate under-confidence.

4 Results

Table 1 summarizes the key interaction coefficients from the fully specified model, comparing the AMRS replication results directly to the corresponding estimates of CDLMMWY. As shown in the table, average AI access raises outcome block accuracy in the paired-image AMRS sample by 1.14 percentage points. The corresponding CDLMMWY treatment effect in this specification is 6.16 percentage points. The CDLMMWY ability-by-AI coefficient (β_1) is -2.16, while the AMRS analogue is -5.57. The CDLMMWY calibration-by-AI coefficient (β_2) is 1.35, while the AMRS analogue is 3.22. Both interaction coefficients have the same sign as in CDLMMWY and are statistically significant in this specification.[3]

Table 1: Table 2 analogue: key coefficients

Term	CDLMMWY, col. 3	Radiology replication, col. 3
AI access	6.16 (0.47)	1.14 (1.08)
Ability × AI	-2.16 (0.57)	-5.57 (1.50)
Calibration × AI	1.35 (0.54)	3.22 (1.31)
Observations	732	70

The four AMRS fitted treatment effects have the same ordering as CDLMMWY. The estimates are 9.93, 3.49, -1.20, and -7.65 percentage points for, respectively, low-ability/high-calibration, low-ability/low-calibration, high-ability/high-calibration, and high-ability/low-calibration users. The Table 2 calibration-by-AI coefficient is positive as in CDLMMWY.

AI also compresses model-predicted performance gaps. In AMRS, the IQR of predicted accuracy falls from 7.39 percentage points without AI to 3.32 percentage points with AI. The calibration counterfactual reproduces the stronger CDLMMWY pattern: under perfect calibration, the with-AI IQR is 1.73 percentage points, which is smaller than the actual with-AI IQR.

5 Discussion

The paired-image AMRS sample reproduces the key CDLMMWY results qualitatively in a professional diagnostic setting, as the ability and calibration patterns point in the same direction as the original results. The sample choice is deliberate: Design 3 uses the same images with and without AI, allowing for a clean replication of the CDLMMWY analysis. This aspect of experimental design also reduces case-mix noise in treatment effect measurement, which is valuable given that the number of cases per radiologist is smaller. However,

[3] Appendix Table 2 reports wild cluster bootstrap p-values.

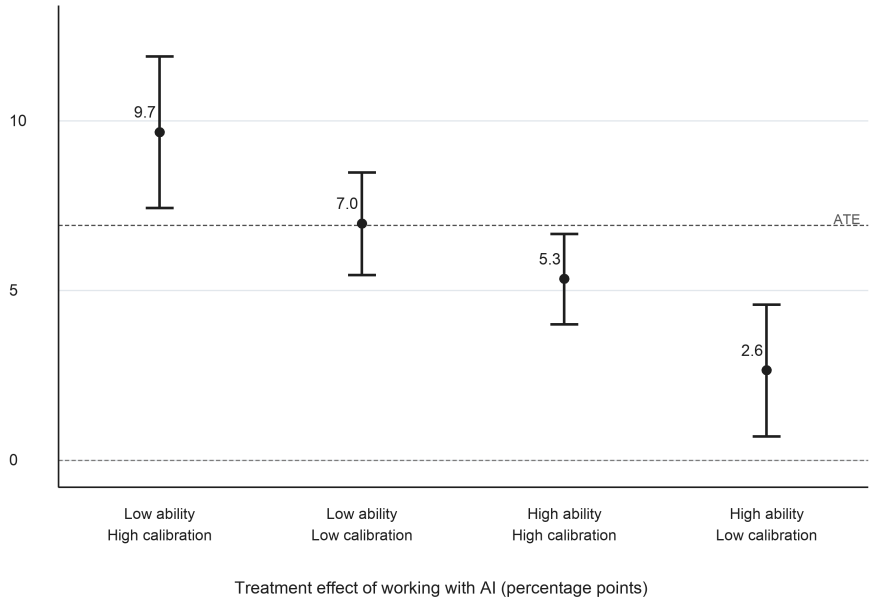

Figure 4: Heterogeneous value of AI by baseline ability and calibration in the CDLMMWY experiment.

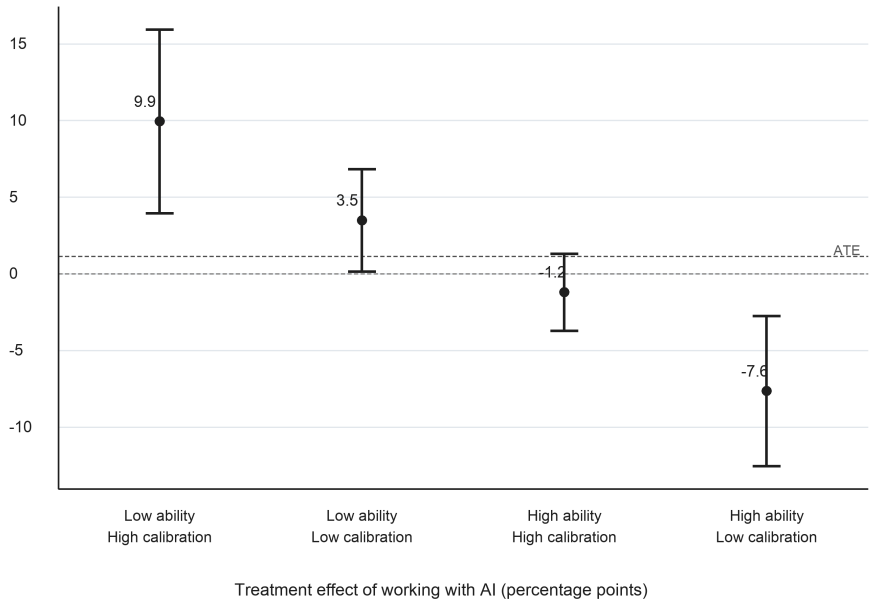

Figure 5: Heterogeneous value of AI by baseline ability and calibration in the AMRS paired-image radiology sample.

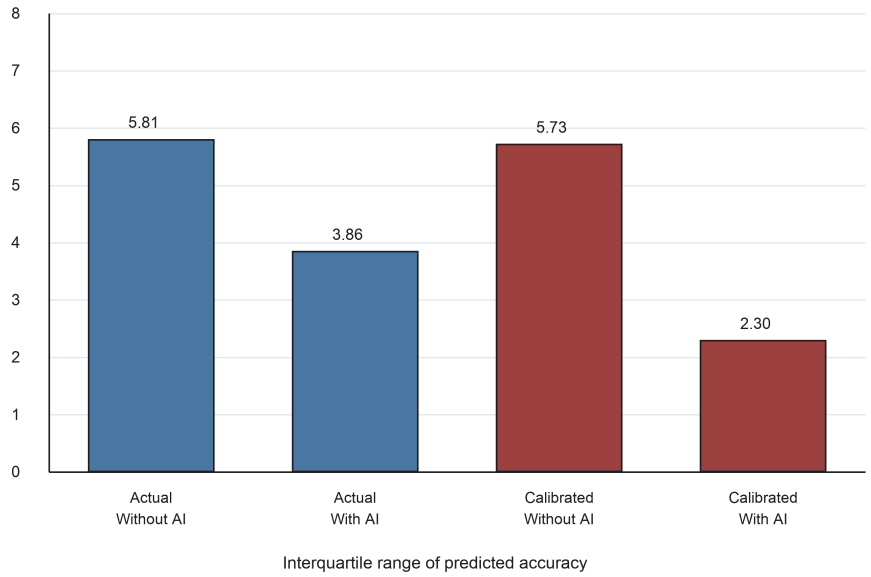

Figure 6: Counterfactual productivity gap in the CDLMMWY experiment.

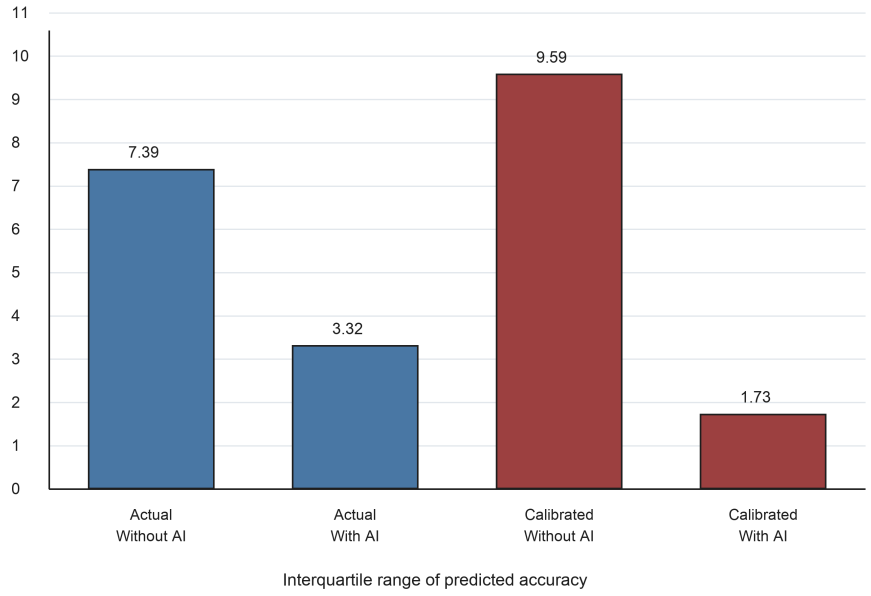

Figure 6: Counterfactual productivity gap in the AMRS paired-image radiology sample.

using just the paired-image sample comes at the cost of a smaller radiologist sample, so the estimates of heterogeneous effects are not highly powered. Given the current SEs, the approximate 80% MDEs are about 4.3 percentage points for Ability $\times$ AI and 3.8 percentage points for Calibration $\times$ AI.

The replication is nevertheless informative. It shows that the CDLMMWY productivity and compression results are not limited to the original age-classification task. This distinction is substantively important: if these ability- and calibration-based patterns are general features of AI-assisted work, then training, interface design, and delegation policies should target not only ability but also users' knowledge of their own accuracy. The paired-image evidence is consistent with that mechanism, while the smaller sample warrants caution in comparing the magnitudes of the heterogeneous effects across settings.

References

Agarwal, Nikhil, Anne Moehring, Pranav Rajpurkar, and Tobias Salz (2023). *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology*. NBER Working Paper 31422. National Bureau of Economic Research. DOI: 10.3386/w31422. URL: https://doi.org/10.3386/w31422.

Autor, David (2024). "Applying AI to Rebuild Middle Class Jobs". In: *National Bureau of Economic Research Working Paper*.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond (2023). *Generative AI at Work*. NBER Working Paper. National Bureau of Economic Research.

Caplin, Andrew, David J. Deming, S. Li, Daniel Martin, Philip Marx, Ben Weidmann, and K. J. Ye (2025a). *Online Appendix and Data Files for The ABCs of Who Benefits from Working with AI: Ability, Beliefs, and Calibration*. INFORMS. DOI: 10.1287/mnsc.2024.08994. URL: https://doi.org/10.1287/mnsc.2024.08994.

— (2025b). "The ABCs of Who Benefits from Working with AI: Ability, Beliefs, and Calibration". In: *Management Science*. DOI: 10.1287/mnsc.2024.08994. URL: https://doi.org/10.1287/mnsc.2024.08994.

Gonzalez, Cleotilde, Kate Donahue, Daniel G. Goldstein, Hoda Heidari, Mohammad S. Jalali, Beau Schelble, Aarti Singh, and Anita Williams Woolley (2026). "Toward a science of human–AI teaming for decision making: A complementarity framework". In: *PNAS Nexus* 5.3. Published online February 19, 2026, pgag030. DOI: 10.1093/pnasnexus/pgag030. URL: https://doi.org/10.1093/pnasnexus/pgag030.

Kargarnovin, Shaida, Christopher Ivan Hernandez, Dirk Reiners, Carolina Cruz-Neira, Grace Bochenek, and Waldemar Karwowski (2026). "From testbeds to high-stakes work: a review of human–AI teaming domains and teaming factors". In: *Frontiers in Robotics and AI* 13. Published May 7, 2026, p. 1733942. DOI: 10.3389/frobt.2026.1733942. URL: https://doi.org/10.3389/frobt.2026.1733942.

Liu, Peng, Jiaxin Zhang, Shuaiqi Chen, and Shanguang Chen (2025). "Human–AI teaming in healthcare: $1 + 1 > 2$?" In: *npj Artificial Intelligence* 1. Published December 2, 2025, p. 47. DOI: 10.1038/s44387-025-00052-4. URL: https://doi.org/10.1038/s44387-025-00052-4.

Moehring, Anne, M. Kutwal, R. Huang, O. Banerjee, A. Jacobi, C. Eber, D. Mendoza, M. Chung, E. Dayan, Y. Gupta, T. D. T. Bui, S. Q. H. Truong, A. Pareek, C. P. Langlotz, M. P. Lungren, Nikhil Agarwal, Pranav Rajpurkar, and Tobias Salz (2025). "A Dataset for Understanding Radiologist-Artificial Intelligence Collaboration". In: *Scientific Data* 12, p. 739. DOI: 10.1038/s41597-025-05054-0. URL: https://doi.org/10.1038/s41597-025-05054-0.

Noy, Shakked and Whitney Zhang (2023). "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence". In: *Science* 381.6654, pp. 187–192.

Wekenborg, Magdalena Katharina, Stephen Gilbert, and Jakob Nikolas Kather (2025). "Examining human–AI interaction in real-world healthcare beyond the laboratory". In: *npj Digital Medicine* 8. Published March 19, 2025, p. 169. DOI: 10.1038/s41746-025-01559-5. URL: https://doi.org/10.1038/s41746-025-01559-5.

A Robustness Checks

Table 2 reports checks on the paired-image estimates using wild cluster bootstrap p-values for the two interaction coefficients, clustered by radiologist.

Table 2: Robustness checks for interaction estimates. 999 wild cluster bootstrap replications. MDEs are in percentage points and use a two-sided 5% test with 80% power and a 35-radiologist small-cluster reference.

Term	Estimate	Cluster SE	Cluster p	Wild p	80% MDE
Ability × AI	-5.57	1.50	0.0007	0.005	4.31
Calibration × AI	3.22	1.31	0.0187	0.023	3.76