

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

Daniel Martin[†]

June 10, 2026

Abstract

Artificial intelligence (AI) systems increasingly assist human experts, but the consequences of AI assistance on productivity can be heterogeneous. Caplin, Deming, S. Li, Martin, Marx, Weidmann, and Ye (2025b) provide evidence that two characteristics, ability and belief calibration, help to determine the returns to AI assistance. This note 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 use the radiologist assessments from the repeated-case designs, which include 68 radiologists and 11,420 paired radiologist-patient-pathology observations. The results of this replication support the external validity of their core findings: lower baseline ability and higher calibration predict larger incremental value from AI.

*I thank Alex Moehring for helpful comments and 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 human users (Brynjolfsson, D. Li, and Raymond 2023; Noy and W. Zhang 2023; Autor 2024).

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

The task that CDLMMWY experimentally implement is an age guessing task (a “Bouncer” task¹ 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. 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. This experiment features professional experts assessing medically meaningful images and 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).

Moving from an age-guessing experiment to a professional radiology experiment allows for a high-expertise test of the CDLMMWY results. While the sample size of expert radiologists is naturally smaller, their assessments provide evidence consistent with the qualitative direction of the original analysis: 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.

¹I thank Jason Somerville for proposing this task name.

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 block of cases and the impact of AI on productivity to be estimated on another block of cases. CDLMMWY accomplish this by designing an experiment where all subjects complete one block of cases without AI and then a second block of cases either with or without AI.² To accomplish this with the radiologist data, I use the repeated-case designs of the AMRS experiment (Design 2 and Design 3). In Design 2, the same radiologist reads the same patient-pathology cases in all four randomly ordered treatment arms, and in Design 3, the same session-level radiologist-patient-pathology readings are repeated first without and then with AI assistance. Pooling these designs yields 68 radiologists and 11,420 paired radiologist-patient-pathology observations before aggregation to two outcome observations per radiologist. This paired structure makes the AI comparison less confounded by case difficulty, pathology mix, or radiologist composition, and the common cases make radiologist-level ability and calibration more comparable across radiologists.

For the pooled main sample and each design-specific appendix sample, I assign half of patients 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 constructed 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 are paired, and one patient from each pair is assigned to the outcome block. In the pooled main sample, I searched seeded random draws within these adjacent pairs and chose a split with 162 skill-block patients and 162 outcome-block patients. The selected split has zero monotonicity violations in the outcome-block AI calibration curve, a 0.07 percentage point difference in no-AI prior accuracy across blocks, and a 0.02 percentage point difference in

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

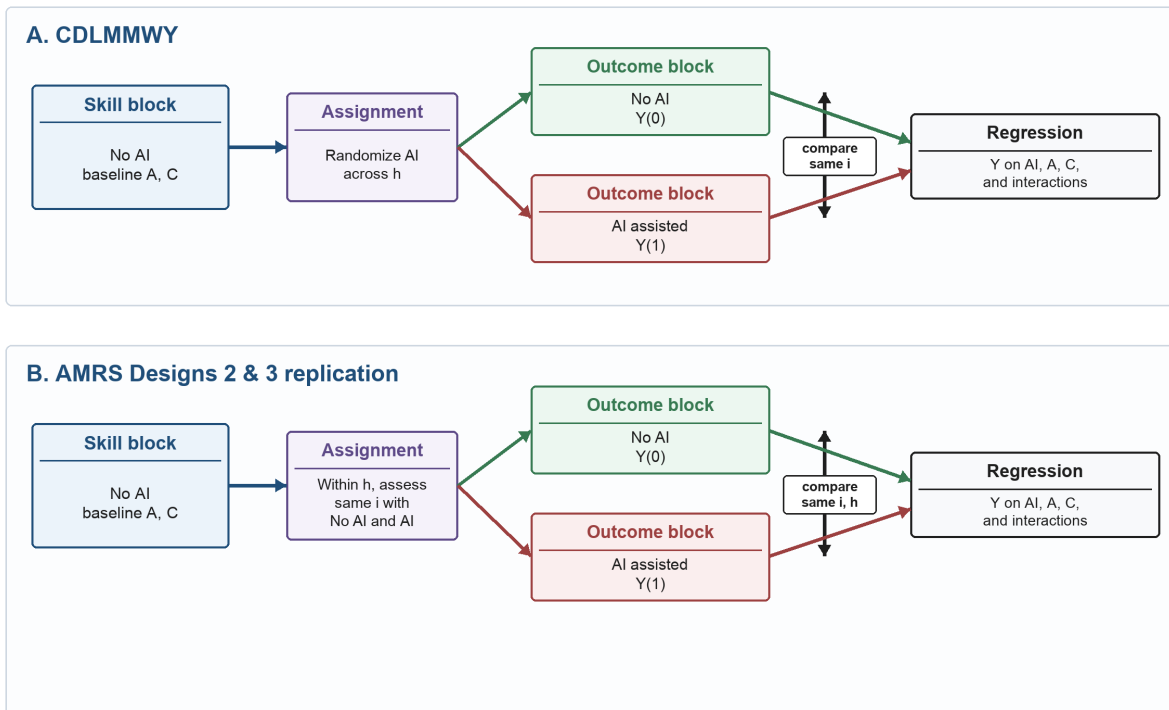


Figure 1: Comparison between the original CDLMMWY design and the replication. Panel A shows the original CDLMMWY structure, where baseline ability and calibration are measured in a no-AI skill block and AI assistance is randomized in the outcome block. Panel B shows the pooled AMRS Designs 2 and 3 replication 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.

no-AI prior confidence across blocks.

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

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

Accuracy A_h is then average correctness. For calibration, the confidence of beliefs over each case is $\max\{B_{hi}, 1 - B_{hi}\}$. Net confidence averages confidence minus correctness across cases, and calibration C_h is the negative absolute value of this average net confidence.

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

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

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 a finite-sample correction. 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.

3 Results

First I examine heterogeneity in accuracy and confidence. Figure 2 compares the accuracy distributions across CDLMMWY and AMRS, and Figure 3 compares the corresponding net-confidence distributions. The accuracy distributions are roughly comparable because CDLMMWY selected images to produce a similar performance distribution to that in AMRS.

Table 1 summarizes the key interaction coefficients from the fully specified model, comparing the pooled AMRS replication results directly to the corresponding estimates of CDLMMWY.³ As shown in the table, average AI access raises outcome block accuracy in the pooled AMRS sample by 1.79 percentage points, while the corresponding CDLMMWY treatment effect in this specification is 6.16 percentage points. Both interaction coefficients have the same sign as in CDLMMWY. The CDLMMWY ability-by-AI coefficient (β_1) is

³Appendix Table 2 reports the pooled estimates and separate estimates for Designs 2 and 3.

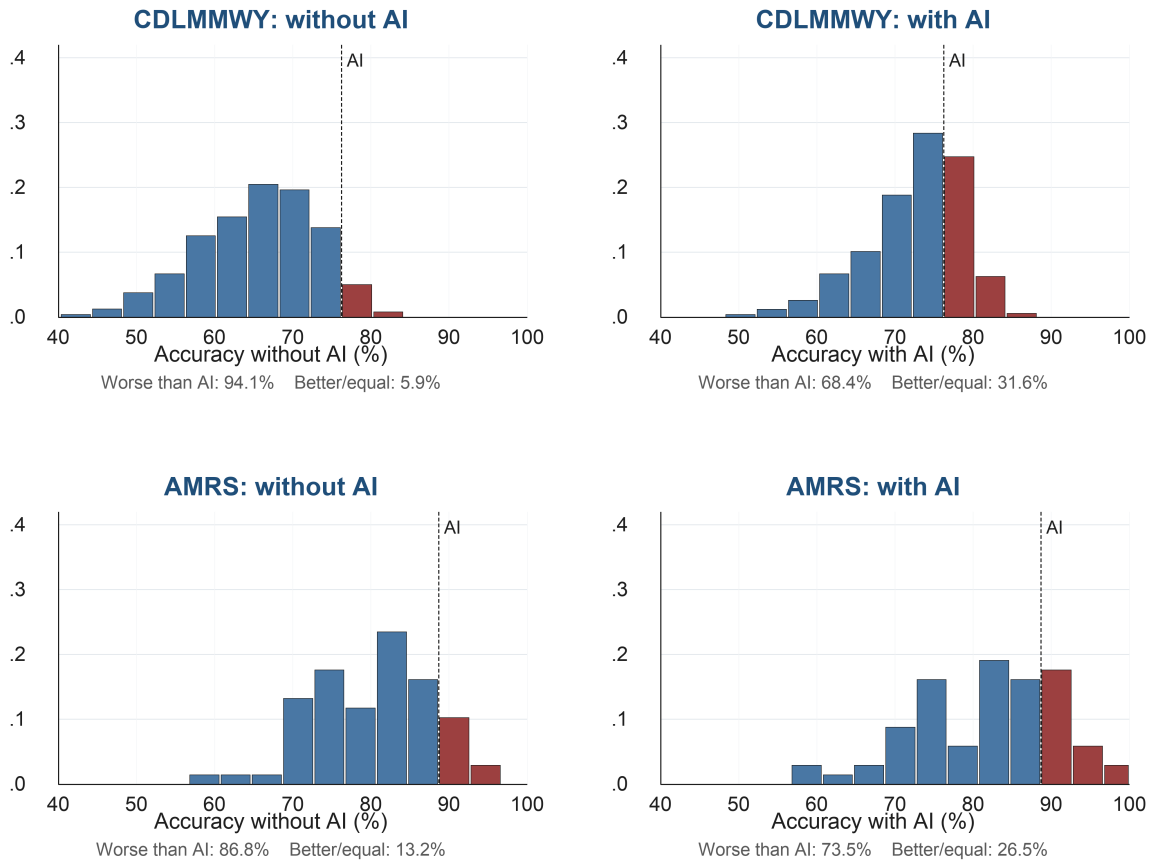


Figure 2: Accuracy distributions across studies. Each panel shows subject-level mean outcome-block accuracy, separately without and with AI. The dashed vertical line marks the AI-alone accuracy benchmark in the corresponding study.

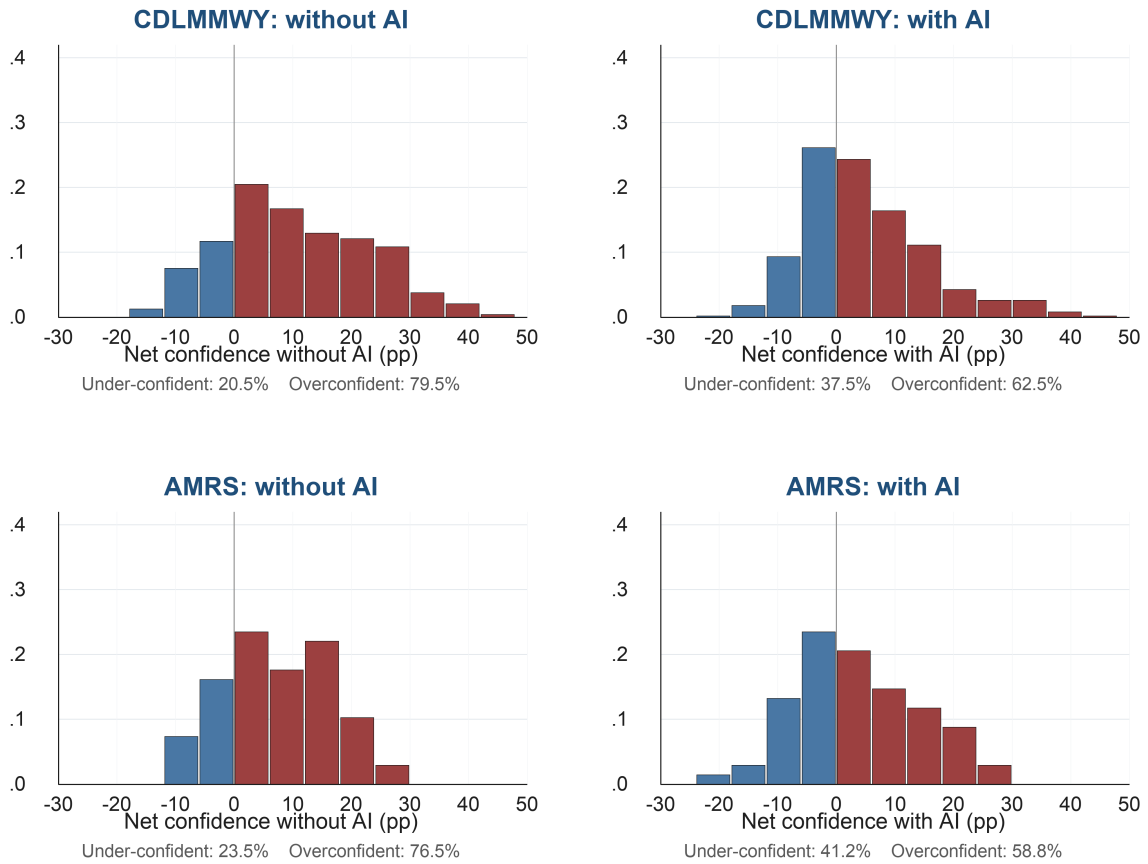


Figure 3: Net confidence distributions across studies. Net confidence is confidence minus correctness, so values closer to zero correspond to better calibration, and positive values indicate overconfidence and negative values indicate under-confidence.

-2.16, while the pooled AMRS analogue is -1.45. The CDLMMWY calibration-by-AI coefficient (β_2) is 1.35, while the pooled AMRS analogue is 1.86.⁴ Appendix Table 2 shows that the Design 2 and Design 3 estimates have the same signs as the pooled estimates, although the Design 3 magnitudes are larger.

Table 1: Table 2 analogue: key coefficients

| Term | CDLMMWY, col. 3 | Radiology replication, col. 3 |
|-------------------------|-----------------|-------------------------------|
| AI access | 6.16 (0.47) | 1.79 (0.67) |
| Ability \times AI | -2.16 (0.57) | -1.45 (0.93) |
| Calibration \times AI | 1.35 (0.54) | 1.86 (0.95) |
| Observations | 732 | 136 |

Figure 5 plots fitted values from a regression of $Y_h(1) - Y_h(0)$ on A_h and C_h to report AI gains at plus or minus one standard deviation of baseline ability and calibration. As shown in the figure, the four pooled AMRS fitted treatment effects are 5.10, 1.38, 2.20, and -1.53 percentage points for, respectively, low-ability/high-calibration, low-ability/low-calibration, high-ability/high-calibration, and high-ability/low-calibration users. As in CDLMMWY, the highest fitted gains are for low-ability, high-calibration users and the lowest fitted gains are for high-ability, low-calibration users.

4 Discussion

The pooled AMRS repeated-case sample qualitatively replicates the key CDLMMWY results in a professional diagnostic setting, as the ability and calibration patterns point in the same direction as the original results. This suggests that the CDLMMWY productivity 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.

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,

⁴Appendix Table 3 reports wild cluster bootstrap p-values. The calibration interaction is statistically significant using wild-cluster inference, while the ability interaction is estimated less precisely.

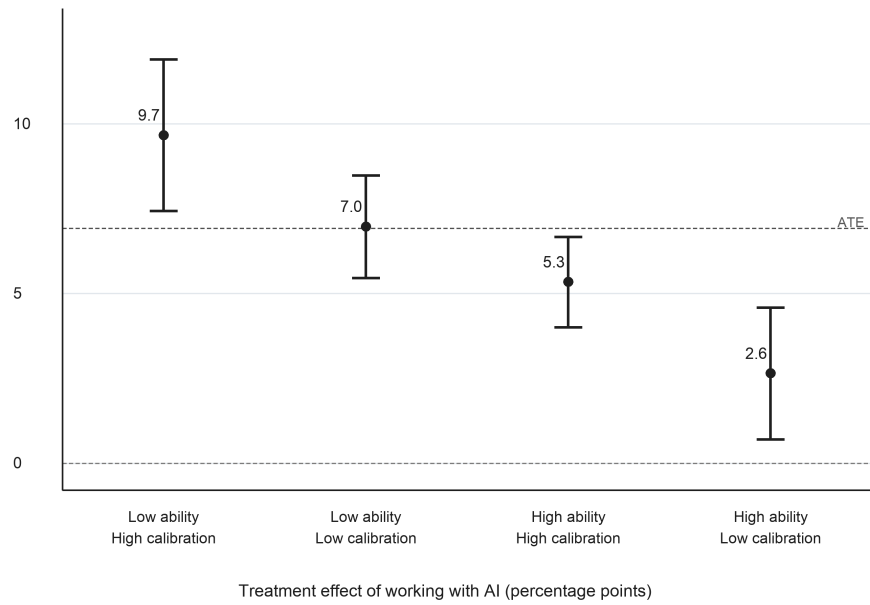


Figure 4: Heterogeneous value of AI by baseline ability and calibration in the CDLMMWY experiment.

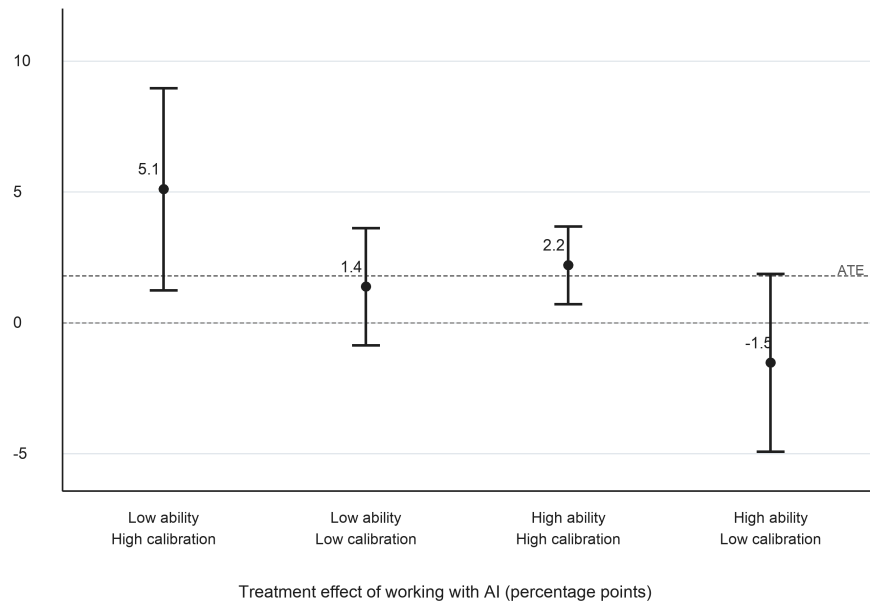


Figure 5: Heterogeneous value of AI by baseline ability and calibration in the pooled AMRS repeated-case radiology sample.

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](#)).

References

- Agarwal, Nikhil, Alex 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1038/s44387-025-00052-4). URL: <https://doi.org/10.1038/s44387-025-00052-4>.
- Moehring, Alex, 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](https://doi.org/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](https://doi.org/10.1038/s41746-025-01559-5). URL: <https://doi.org/10.1038/s41746-025-01559-5>.

A Robustness Checks

Table 2 reports the pooled and design-specific AMRS estimates. Table 3 reports checks on the interaction estimates using wild cluster bootstrap p-values, clustered by radiologist.

Table 2: AMRS estimates by design

| Term | Pooled Designs 2–3 | Design 2 | Design 3 |
|-------------------------|--------------------|--------------|--------------|
| AI access | 1.79 (0.67) | 2.19 (0.73) | 1.14 (1.08) |
| Ability \times AI | -1.45 (0.93) | -1.93 (0.94) | -5.57 (1.50) |
| Calibration \times AI | 1.86 (0.95) | 1.66 (0.79) | 3.22 (1.31) |
| Observations | 136 | 66 | 70 |
| Radiologists | 68 | 33 | 35 |

Table 3: 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.

| Sample | Term | Estimate | Cluster SE | Wild p | 80% MDE |
|----------|-------------------------|----------|------------|----------|---------|
| Pooled | Ability \times AI | -1.45 | 0.93 | 0.124 | 2.67 |
| Pooled | Calibration \times AI | 1.86 | 0.95 | 0.043 | 2.75 |
| Design 2 | Ability \times AI | -1.93 | 0.94 | 0.087 | 2.70 |
| Design 2 | Calibration \times AI | 1.66 | 0.79 | 0.070 | 2.28 |
| Design 3 | Ability \times AI | -5.57 | 1.50 | 0.005 | 4.31 |
| Design 3 | Calibration \times AI | 3.22 | 1.31 | 0.023 | 3.76 |