NBER WORKING PAPER SERIES

THE ABC'S OF WHO BENEFITS FROM WORKING WITH AI:
ABILITY, BELIEFS, AND CALIBRATION

Andrew Caplin
David J. Deming
Shangwen Li
Daniel J. Martin
Philip Marx
Ben Weidmann
Kadachi Jiada Ye

Working Paper 33021
http://www.nber.org/papers/w33021

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2024

We thank the Sloan Foundation for supporting this research under the "Cognitive Economics at Work" grant. IRB approval from the University of California, Santa Barbara. Preregistration: https://aspredicted.org/pm63-gvdv.pdf. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The ABC's of Who Benefits from Working with AI: Ability, Beliefs, and Calibration
Andrew Caplin, David J. Deming, Shangwen Li, Daniel J. Martin, Philip Marx, Ben Weidmann, and Kadachi Jiada Ye
NBER Working Paper No. 33021
October 2024
JEL No. D81, J24

ABSTRACT

We use a controlled experiment to show that ability and belief calibration jointly determine the benefits of working with Artificial Intelligence (AI). AI improves performance more for people with low baseline ability. However, holding ability constant, AI assistance is more valuable for people who are calibrated, meaning they have accurate beliefs about their own ability. People who know they have low ability gain the most from working with AI. In a counterfactual analysis, we show that eliminating miscalibration would cause AI to reduce performance inequality nearly twice as much as it already does.

Andrew Caplin
Department of Economics
New York University
19 W. 4th Street, 6th Floor
New York, NY 10012
and NBER
andrew.caplin@nyu.edu

David J. Deming
Harvard Kennedy School
Malcolm Wiener Center for Social Policy
79 JFK St
Cambridge, MA 02138
and NBER
david_deming@harvard.edu

Shangwen Li
Department of Economics
New York University
sl8968@nyu.edu

Daniel J. Martin
Department of Economics
University of California, Santa Barbara
2127 North Hall
Santa Barbara, CA 93106-9210
danielmartin@ucsb.edu

Philip Marx
Department of Economics
Louisiana State University
501 South Quad Dr.
Baton Rouge, LA 70803
pmarx@lsu.edu

Ben Weidmann
Harvard Kennedy School
Malcolm Wiener Center for Social Policy
79 JFK St
Cambridge, MA 02138
benweidmann@hks.harvard.edu

Kadachi Jiada Ye
Department of Economics
New York University
jy1603@nyu.edu

A data appendix is available at http://www.nber.org/data-appendix/w33021

The ABC's of Who Benefits from Working with AI: Ability, Beliefs, and Calibration

Andrew Caplin, David Deming, Shangwen Li, Daniel Martin, Philip Marx, Ben Weidmann, and Kadachi Jiada Ye*

October 1, 2024

Abstract

We use a controlled experiment to show that ability and belief calibration jointly determine the benefits of working with Artificial Intelligence (AI). AI improves performance more for people with low baseline ability. However, holding ability constant, AI assistance is more valuable for people who are calibrated, meaning they have accurate beliefs about their own ability. People who know they have low ability gain the most from working with AI. In a counterfactual analysis, we show that eliminating miscalibration would cause AI to reduce performance inequality nearly twice as much as it already does.

1 Introduction

In the future, many workers will use AI tools. How will these tools affect productivity and inequality in the labor market? Broadly speaking, AI is trained on a vast corpus of human examples and can perform a variety of tasks at, or near, the level of human experts. But AI rarely outperforms the best humans, and workers can complement AI both with contextual knowledge and the ability to deal with atypical examples. In this paper we propose that complementarities between workers and AI tools will be greatest when workers have an accurate appraisal of their own abilities.

We show that workers with well-calibrated beliefs make better use of AI assistance. Calibrated beliefs are those that are aligned with objective likelihoods. For example, when

*Caplin, Li, and Ye: New York University; Deming and Weidmann: Harvard University; Martin: University of California, Santa Barbara; and Marx: Louisiana State University. We thank the Sloan Foundation for supporting this research under the "Cognitive Economics at Work" grant. IRB approval from the University of California, Santa Barbara. Preregistration: https://aspredicted.org/pm63-gvdv.pdf.

90% sure, a calibrated person is correct 90% of the time.[1] Belief calibration is important regardless of the AI's capabilities. Consider the case of an AI that provides assistance to workers who are making some prediction about a future event, such as whether a financial transaction is fraudulent or whether a person will reoffend if released on bail. If the AI's prediction is very poor, under-confident workers will defer too much even if they have better information. If the AI's prediction is extremely accurate, overconfident workers will override it when they should defer, limiting the scope of their gains from AI assistance (Hoffman, Kahn and Li, 2018).

In this paper, we introduce an experimental framework to measure both ability and calibration and their impact on the gains from AI assistance. To measure these characteristics properly we must elicit probabilistic assessments from a large sample of participants in a task with clear ground truth and with an appropriate level of task difficulty relative to a calibrated AI.

We ask 732 experimental participants to report the likelihood that 160 different people in photographs are over 21 years of age. Some participants are randomly assigned an AI assistant that provides them with an algorithmic confidence score. The AI assistant is more accurate than the average person but less accurate than the best person, a fact that is known to participants. Treatment group participants receive AI assistance for only some images, which allows us to estimate baseline ability at the task. We also measure general cognitive ability by administering a short form of the Ravens matrices test.

We find that AI assistance improves prediction accuracy on average and with heterogeneous impacts across individuals. We present three main findings addressing the question of 'who benefits?'. First, the impact of AI assistance depends on whether participants are calibrated about their own abilities. Holding baseline ability fixed, a one standard deviation increase in calibration increases the average treatment effect of AI assistance by 20 percent.

Second, we find that participants with high calibration *and* low ability gain the most from working with AI. This is consistent with finding that AI closes the performance gap

[1] For a review of literature on belief calibration and systematic departures, see Benjamin (2019).

between lower performers and their more expert peers (Brynjolfsson, Li and Raymond, 2023; Noy and Zhang, 2023; Autor, 2024).

Third, we find that miscalibration limits the extent to which AI assistance can reduce performance inequality. In a counterfactual analysis where miscalibration is eliminated, the productivity gap (captured by interquartile ratio, a measure of spread) would shrink nearly twice as much as it does already with the introduction of AI assistance. The reason is that low-ability participants also tend to be more miscalibrated, which limits the benefits from working with AI.

Our results build on several findings from the literature on AI and labor market inequality. First, we contribute to the growing evidence in labor economics that, in many cases, AI tools increase average productivity and reduce disparities in output between high-skilled and low-skilled workers (Gruber et al., 2020; Brynjolfsson, Li and Raymond, 2023; Noy and Zhang, 2023; Merali, 2024; Choi, Monahan and Schwarcz, 2023). For example, Gruber et al. (2020) find that giving health insurance brokers AI-based decision support improves retiree choices by $278 on average and decreases performance heterogeneity among brokers. Our findings replicate these empirical patterns and highlight a potential mechanism – belief calibration – that may be important in understanding who benefits from AI adoption.

Second, we extend the experimental literature examining the way in which AI advice influences human decision-making (Dietvorst, Simmons and Massey, 2015; Green and Chen, 2019; Tejeda et al., 2022; Dargnies, Hakimov and Kübler, 2024). Green and Chen (2019) provide an early example of work exploring calibration and AI assistance. They show that in the context of making predictions about loan defaults and court appearances, Mechanical Turk participants working with AI were not able to effectively evaluate either the accuracy of algorithmic risk assessments or their own predictions.[2] In an experimental study of the effects of gender on hiring decisions, Dargnies, Hakimov and Kübler (2024) find that more confident participants (as measured by how often they thought they hired the best performer) were less likely to completely override their choices with AI-generated

[2] Broader motivation for studying belief calibration is provided by an extensive literature on overconfidence as a general trait (e.g., Moore and Healy (2008)).

ones than under-confident participants. Closer to our task, Tejeda et al. (2022) ask participants to classify fuzzy versions of images from the ImageNet database with the aid of AI recommendations. They present a hierarchical Bayesian framework that incorporates multiple factors, including decision-makers' confidence in their classifications, and use this framework to infer that participants are more likely to accept AI recommendations when they are not confident about their decision but the AI is highly confident.

Last, our findings build directly on Agarwal et al. (2023), who randomly offer AI assistance to radiologists after asking them to report their confidence in an estimate of how likely chest radiographs were to exhibit different pathologies.[3] They find that AI assistance does not improve performance, even though the AI is more accurate than about 75 percent of radiologists in their sample. They provide evidence that this is not because the radiologists ignore the AI assistance and that this occurs because they incorrectly combine information across predictions. Importantly, they find evidence of significant heterogeneity in participants' use of AI, which is consistent with our results.

Overall, mounting evidence suggests that AI adoption has heterogeneous impacts across individuals. The primary contribution of our paper is to examine whether this heterogeneity is a function of measurable, individual differences in skills. Using a carefully controlled experiment, we examine the extent to which differences in AI value-add are driven by ability and belief calibration (the 'ABCs of who benefits').

The paper is structured as follows. Section 2 introduces a minimal conceptual framework to define our main measures of ability, beliefs, and calibration, and to discuss the measurement challenges contained therein. Section 3 discusses the experimental design and participant pool, while Section 4 discusses our empirical strategy. Section 5 presents our experimental analysis and results. Section 6 concludes with a discussion.

[3] To accurately assess individual differences in ability and calibration, we amend the experimental design in Agarwal et al. (2023) in two ways. First, we ask people to perform a very simple task that does not require specialized training, so that we can collect data on many individuals. Second, we gather enough assessments from each person to estimate individual ability (e.g. percent of faces identified correctly) as distinct from calibration (e.g. knowing how accurate one is at identifying faces). In combination with the randomization of AI assistance, these individual-level measurements allow us to estimate heterogeneous treatment effects as a function of heterogeneous baseline skills.

2 Ability, Beliefs, and Calibration

2.1 Measures

Each individual decision-maker i faces a number of binary classification task instances $\omega \in \Omega = [0,1]$ for which we elicit probabilistic beliefs $B_i(\omega) \in [0,1]$, henceforth encoded in random variable B_i. True states across task instances are summarized by a random variable $S \in \{0,1\}$ with balanced prior $E[S] = 1/2$. We use the relationship between an individual's beliefs B_i and the true states S to define a simple measure of both the individual's *ability* – what they know – and *calibration* – how well they know what they know.

For ability, define a random variable summarizing whether beliefs would have induced correct choices at a decision threshold of 50%:

$$R_i = \begin{cases} 1 & \text{if } |B_i - S| < 0.5 \\ 0.5 & \text{if } B_i = 0.5 \\ 0 & \text{else,} \end{cases}$$

and define the individual's ability as the expected accuracy of such choices:

$$A_i = E[R_i] \tag{1}$$

For calibration, define a random variable summarizing the confidence of beliefs over each image:

$$K_i = \max\{B_i, 1 - B_i\},$$

and define the individual's net confidence as the expected difference between their confidence and correctness:

$$N_i = E[K_i - R_i]$$

If correctness exceeds confidence $N_i < 0$, we say that an individual is under-confident; if confidence exceeds correctness $N_i > 0$, we say that an individual is over-confident. In either case, the larger the difference between perceived and actual performance, the less an individual's beliefs are calibrated. We capture this notion by defining the individual's calibration as the negative absolute value of their net confidence:

$$C_i = -\big|E[R_i - K_i]\big| \tag{2}$$

Empirically, an individual is calibrated if, say, expressing 75% confidence in their classifications, they are also correct 75% of the time. Alternatively, calibration could be further disaggregated across bins of an individual's confidence (DeGroot and Fienberg, 1983) or estimated parametrically (Grether, 1980); in Appendix A.4 we show that our results are robust to such alternative definitions of calibration and to alternative definitions of ability, such as the Area Under the Receiver Operator Characteristic Curve, or AUC (Bamber, 1975; Hanley and McNeil, 1982).

Our main question is how ability and calibration relate to the ability to incorporate AI advice. In contrast to human decision makers, we ensure that the AI model's probabilistic predictions, which we denote by M, are calibrated. Working with the AI, individual i observes and integrates the AI predictions M before reporting their AI-augmented subjective beliefs $\bar{B}_i$, which determine AI-augmented ability $\bar{A}_i$ as in (1) and calibration $\bar{C}_i$ as in (2).[4] If the human benefits from the AI but also does not repeat its predictions M wholesale, we expect that both $\bar{B}_i \neq B_i$ and $\bar{B}_i \neq M$. Our main prediction is that conditional on their individual ability A_i, individuals who are more calibrated will benefit more from working with AI in terms of their AI-augmented ability $\bar{A}_i$.

2.2 Causal Framework

We now embed our main prediction into a standard causal framework. In a random subset of task instances we call the control block, all individuals work without AI assistance. We estimate individual skills A_i and C_i for each individual i using plug-in estimators of the respective definitions (1) and (2) in the control block of task instances. In the remaining subset of task instances we call the treated block, we randomly assign as treatment $T_i \in \{0, 1\}$ whether human i worked with or without AI assistance. To distinguish outcomes from skills, we adopt standard potential outcomes notation and let $Y_i(0) = A_i$ denote individual i's performance if (potentially) assigned to work alone, let $Y_i(1) = \bar{A}_i$ denote individual i's performance if assigned to work with AI, and let $Y_i = Y_i(T_i)$ denote the observed outcome given treatment status. We estimate individual performance Y_i for

[4] We adopt the convention that a bar indicates an analogous term as defined before, but when working with AI.

each individual using a plug-in estimator of (1) in the treated block of tasks.

The treatment effect $Y_i(1) - Y_i(0)$ is the change in performance for human i from working with AI. Our research question is predicated on the possibility that the effects of working with AI are heterogeneous across individual skill types, i.e., that $E[Y_i(1) - Y_i(0)|A_i = a, C_i = c]$ varies in (a, c). Specifically, our main prediction is that the effect is increasing in calibration c given ability a. The conditioning on ability is important, since the net confidence underlying calibration (2) is a linear function of accuracy. We aggregate such effects in a linear regression model:

$$Y_i = \alpha_0 + \alpha_1 A_i + \alpha_2 C_i + \beta_0 T_i + \beta_1 A_i T_i + \beta_2 C_i T_i + \varepsilon_i, \tag{3}$$

Our main prediction then is that the coefficient β_2 on the interaction of calibration and treatment is significant and positive.

2.3 Measurement Challenges

Our experiment will allow us to estimate individual ability and calibration from observed beliefs, working with or without AI. Nevertheless, estimation and use of these measures raise several challenges. For each participant we need to elicit a large number of probabilistic beliefs to estimate ability and calibration skills and to reduce measurement error therein. We need to do so in a task with (i) a clear, objective ground truth, and (ii) an appropriate perceived and actual difficulty, so that the distribution of probability reports and latent beliefs is not skewed toward extremes. Our focus on treatment effect heterogeneity requires this for a large sample of participants and is furthermore predicated on the presence of heterogeneity in subjects' calibration, conditional on their ability. Additionally, for working with an AI, we need AI predictions to be well-calibrated over the selected images and to be of a suitable diagnostic ability. We next describe our experimental design for surmounting these challenges, beginning with the experimental task.

3 Experimental Design

3.1 The Task

In each round of our experiment, participants were presented with an image of a human face and were asked to report the probability that the individual was over 21 years old at the time the image was taken. Participants were informed that all images were taken between 2010 and 2014, with half of the images depicting individuals who were under 21 years old and the other half depicting individuals over 21 at the time the image was taken.[5]

In half of the rounds, participants in the treatment group were also shown an "AI Assistant guess," which was an AI prediction of whether the individual was over 21 (a percentage between 0 and 100). Participants were told that the AI Assistant guesses were more accurate than the average human, but worse than the most skilled human. See the Appendix for screenshots of choices with and without AI predictions.

3.2 Implementation

We recruited participants from Prolific, a popular platform for conducting online experiments. Prolific requires participants to be at least 18 years old, and we further restricted participants to be U.S. residents. In the recruitment material, participants were told that they would be given $5 as a participation fee if they finished the experiment and that they could receive additional bonus payments based on their choices in the experiment.

After being shown a consent form that outlined the basic conditions of the experiment, each participant was asked whether they would like to participate. Our experiment was deemed exempt from the Federal Regulations at 45 CFR 46.101(b) by the Human Subjects Committee at the University of California, Santa Barbara (protocol number 11-22-0691). Upon agreeing to participate, participants answered two unincentivized questions as an attention check. Participants had only one chance to answer these questions correctly to proceed with the rest of the experiment.

[5] To avoid confusion, we excluded images in which the individual could have been exactly 21 years old.

The rest of the experiment began with several pages of instructions, which provided an estimate of the session length (30-35 minutes), the possible bonus payments ($5 and $1), an explanation of the task they would complete in each round, the total number of rounds (160), and how the bonus payments would be awarded. To incentivize the truthful reporting of beliefs in the task, we used the same incentives provided by Agarwal et al. (2023). A random round was selected (all were equally likely to be selected), and based on the reported probability and true age, the binarized scoring rule (Hossain and Okui, 2013) was used to determine the probability of getting the $5 bonus payment. Following Danz, Vesterlund and Wilson (2022), we provided few details about the scoring rule but indicated that the likelihood of receiving the bonus was maximized by truthful reporting of their belief in each round.

After the experimental instructions, participants completed four unincentivized practice rounds. Participants were not provided with any feedback about their choices or the true age of the individual during the practice rounds. After the practice rounds, participants faced 160 incentivized rounds, divided into eight blocks of 20 rounds each. There were opportunities to pause between blocks, which was intended to mitigate fatigue effects. Participants were also not provided feedback during the incentivized rounds.

In these rounds, participants were provided with a slider bar that ranged from 0 to 100 for reporting the probability that the person in the image was over 21 years old. To minimize anchoring effects, the slider bar did not have a starting location. Participants could click anywhere on the bar to select a probability and adjust it before clicking the submit button to record their choice and proceed to the next round. They could not go back to previous rounds once the submit button was pressed, and they were given a time limit of 60 seconds for each round. Subjects were informed in the instructions that they would have no chance of winning the bonus payment if, in the round that was randomly selected for payment, they had not submitted a response within the time limit.

3.3 AI Assistant

Participants in the control group were not shown any AI Assistant predictions during the 160 incentivized rounds. For participants in the treatment group, AI predictions were provided in blocks 2, 3, 5, and 8, which correspond to the treated blocks. The arrangement of treated and control blocks was chosen to minimize differences in fatigue between treated and control images and to minimize order effects. In each session, images were assigned to be in a control block or treated block, so we refer to images as either being treated or control images. While images were fixed across participants in a session to be either treated or control, within participant the images were randomly assigned to a random round within the block.

Participants in the treatment group were not told in advance that they would be given AI predictions. Instead, after completing the first block of rounds, participants in the treatment group were told that they would be shown an AI Assistant's guess for the upcoming block of 20 rounds (the first treated block). In addition to being told that the AI Assistant performs better than the average human but not as well as the best human, they were also told that the AI predictions were calibrated, and they were given the option to learn what it means for the AI to be calibrated.

For our AI predictions, we used the "Caffe" model trained by Rothe, Timofte and Van Gool (2018), which predicts human ages based on facial images. To generate a confidence score that the person in an image was over 21 years old, we summed the confidence scores for all ages above 21.

3.4 Extra Tasks

Following the 160 rounds, participants completed two additional tasks. The first task was to complete 14 questions from the Raven's Progressive Matrices test. This task was incentivized by paying the participant a $1 bonus if a randomly selected question was answered correctly. The second task was a brief survey in which participants were asked to describe their experience in the experiment, indicate the point at which they felt fatigue during the experiment, and provide any other feedback.

The final part of the experiment consisted of result pages for the incentivized tasks. As in Caplin et al. (2020), we attempted to make the random outcomes used in payment more credible by using the machine time at the moment the participant pressed a button.

3.5 Design Choices

In this section, we discuss several of the main decisions we faced while designing the experiment.

Task Choice: We chose this age classification task for several reasons. First, it is a task that can be completed quickly and with minimal effort, allowing us to collect a large sample of participants and a large number of reports per participant. Second, we could easily modify the task's difficulty by varying the ages included in the experiment. For example, ages closer to 21 are generally more difficult to distinguish than those further way. We set the level of difficulty so that the best participant could outperform the algorithm, while the average human performance remained below it, as in Agarwal et al. (2023).[6] Third, this task requires no specialized training, as it originates from a real-world setting. Common examples of guessing whether someone is under or over 21 include scenarios involving liquor sales, such as stores, bars, restaurants, and entertainment venues. Additionally, people unconsciously judge ages in their daily lives. Thus, age classification does not require specialized expertise and is well-suited for an online experiment to collect a substantial amount of data. Fourth, this task has a clear ground truth, the age of the person in the photograph. Our coarse classification scheme provides further protection from small amounts of measurement error in the recorded age. This differs from Agarwal et al. (2023), where the subjective assessments of experts are aggregated to determine the ground truth.

Image Selection: We utilized the IMDB-WIKI dataset to source images for our experiment.[7] This dataset is the largest publicly available collection of face images with

[6] To this end, we selected images based on ages: 16 to 19 years old for the under-21 category and 23 to 26 years old for the over-21 category. We only know the year when the photo was taken, so to avoid the possibility of including subjects who were exactly 21, we excluded images of individuals who could have been 20, 21, or 22 years old.

[7] See https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/ for details.

age labels, and its size and diversity are useful for training machine learning algorithms. Our aim was to select 160 images from this dataset so that the distribution of probability reports would be as even as possible, which would provide room for both humans and the AI to contribute in helping to identify the category of an image. Additionally, we wanted the AI predictions to be well calibrated over the selected images. To accomplish these objectives in a systematic way, we ran two pilots of 100 participants each. See the Appendix for more details on these pilots and how we used the results of the pilots to select 160 images.

Number of Rounds: In our two pilot sessions, we asked subjects to complete 100 rounds, and on average participants finished the experiment within 15 minutes. Since performance did not decline over rounds, we increased the number of rounds to 160, allowing participants to complete the experiment within approximately 25 minutes on average.

Time Limit: There were two reasons why we introduced a fixed time limit of 60 seconds for participants to report a probability for each image. First, we wanted participants to remain focused on the experiment and not set aside the experiment to work on other tasks. Second, our pilots indicated that 60 seconds was sufficiently long to assess each image. In line with our pilots, this time limit was only reached in 0.23% of subject-round pairs in our experiment (275 out of 117,120 rounds). We exclude these responses from subsequent analyses, leaving 116,845 responses.

4 Results

4.1 Summary Statistics

We recruited 732 participants over two sessions, with 376 participants in the first session and 356 in the second session. Of those, 493 were assigned to the treatment of working with AI, and 239 were assigned to the control of never working with the AI. Summary statistics across both sessions are presented in Table 2 of Appendix A.1. This table shows that there was balance across control and treated participant characteristics.

Figure 1 plots the empirical distributions of ability and net confidence in the treated image block, split by subjects working with and without AI. The dotted lines indicate the performance of the AI. This illustrates several motivating facts. First, the AI has ability in the top 6% of human performers working alone, and working with the AI significantly improves average human performance in our task. However, there remains dispersion in human performance when working with AI, with 68% of treated subjects still performing worse than if they had followed the AI recommendations wholesale.[8] Finally, there is sizeable dispersion in humans' net confidence when working alone, with 79.5% of humans over-confident in their own ability, and with individual confidence exceeding accuracy by 10.6 percentage points on average. We now turn to the estimation of our main regression (3) to determine how much the heterogeneity in the gains from working with AI can be explained by individual-level ability and calibration.

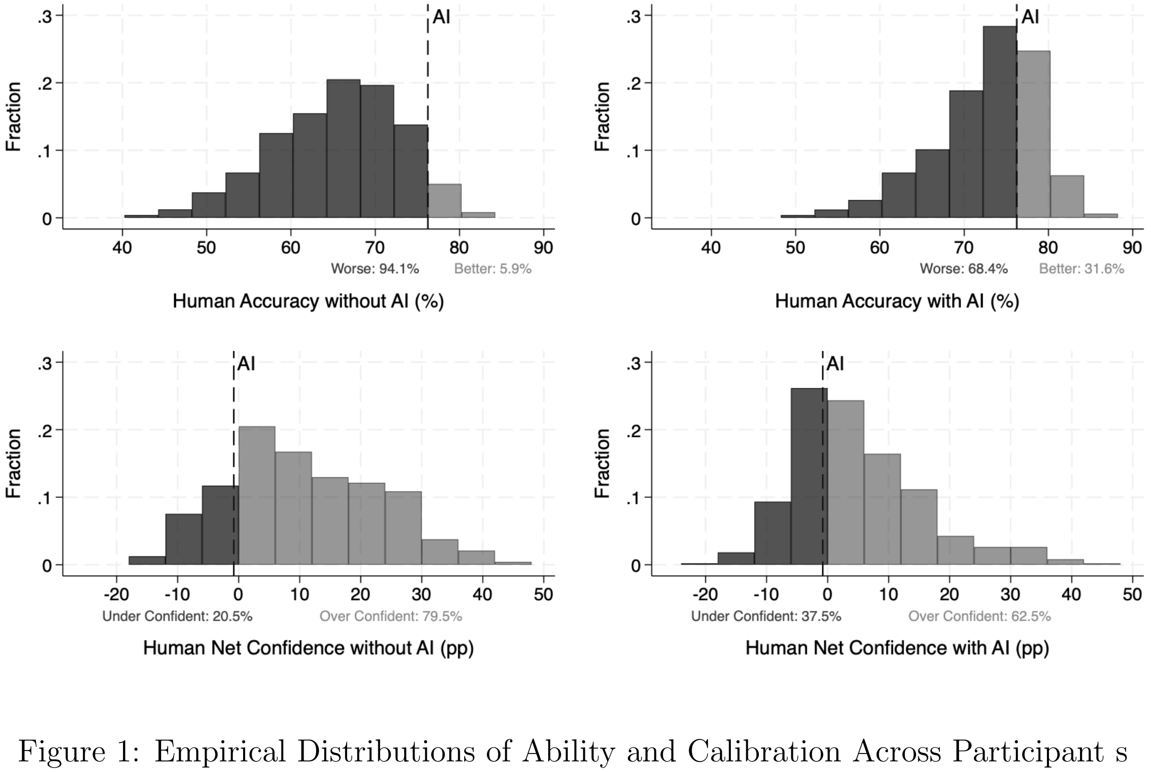

Figure 1: Empirical Distributions of Ability and Calibration Across Participant s

[8] Subjects repeat the AI recommendation exactly (i.e., within a percentage point) 6% of the time with AI and 2% of the time without AI.

4.2 Calibration and the Effects of Working with AI

Table 1: Regression: Skills and Working with AI

	Outcome: Accuracy (Treated Block)			
Skills (Control Block)	(1)	(2)	(3)	(4)
Treatment	6.93	6.27	6.16	6.27
	(0.56)	(0.47)	(0.47)	(0.47)
Accuracy		4.05	3.96	3.60
		(0.38)	(0.45)	(0.45)
Accuracy × Treatment		-1.57	-2.16	-1.83
		(0.50)	(0.57)	(0.57)
Calibration			0.19	0.11
			(0.44)	(0.43)
Calibration × Treatment			1.35	1.40
			(0.54)	(0.54)
IQ				1.17
				(0.43)
IQ × Treatment				-1.01
				(0.51)
Constant	65.46	65.96	65.99	65.88
	(0.48)	(0.39)	(0.40)	(0.39)
Observations	732	732	732	732

Note: Observations are at the subject level. Robust standard errors are in parentheses. Skills are standardized within session.

Table 1 presents estimates from our main regression specification, equation (3). We standardize all three skill measures (accuracy, calibration, IQ) to compare the relative heterogeneity of treatment effects across skills in the sample.

Column 1 shows the impact of AI assistance on accuracy in a simple bivariation regression.[9] AI improves participants' ability to accurately identify an image by 6.9 percentage points, an increase of more than 10 percent relative to the control mean of 65.5 percent. Column 2 adds participants' baseline accuracy with control images and an interaction between accuracy and treatment. Consistent with past work, we find that AI

[9] Table 3 replicates the main results but with demographic controls. We find very similar results. However, since demographic variables were missing for some participants, we present the main results without them.

improves accuracy more for participants with low baseline skill – a one standard deviation increase in accuracy decreases the impact of AI assistance by 1.6 percentage points.

Column 3 adds baseline calibration and the interaction between calibration and treatment. Unlike the interaction with accuracy, we find that better-calibrated participants benefit more from AI assistance. A one standard deviation increase in calibration increase the impact of AI assistance by 1.4 percentage points, or about 20 percent of the total treatment effect of 6.9 percentage points. Both interaction terms are statistically significant at the less than five percent level. Column 4 adds IQ and its interaction with treatment. Similar to the accuracy measure, we find that low IQ participants benefit more from AI help. Adding IQ in Column 4 does not meaningfully change the coefficients on calibration. Overall, we find that the impact AI assistance is increasing in the baseline accuracy of participants' beliefs (calibration) but decreasing in the other two skills (accuracy and IQ).

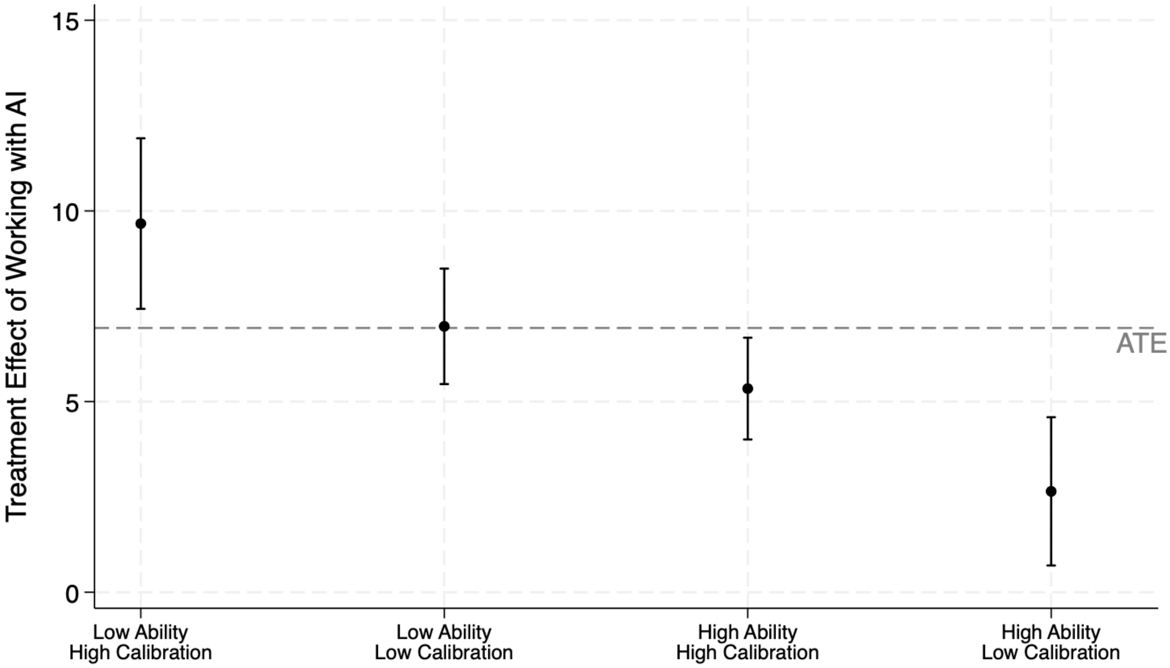

Figure 2: Heterogeneous treatment effects by skill-calibration groups.

To give a visual sense of the treatment effect variation, Figure 2 plots the estimated treatment effects for hypothetical participants with ability and calibration each one standard deviation above or below the mean, which we label respectively as *High* or *Low* Ability or Calibration. The coefficients and 95 percent confidence intervals in Figure 2 come from linear projections of the estimates from Column 3 of Table 1, with the average treatment effect from Column 1 added for clarity. Participants with low ability but accurate beliefs benefit the most – nearly 10 percentage points – from AI assistance. By contrast, participants with high accuracy but miscalibrated beliefs benefit the least.

Over- and under-confident participants may be harmed for different reasons. Under-confident participants may commit too many Type I errors (following the AI when it is incorrect), while over-confident individuals may commit too many Type II errors (failing to follow the AI when it is correct). In Table 4 of Appendix A.2, we repeat our main specification with net confidence instead of calibration and disaggregating by whether the AI prediction was correct. We find that the less confident are relatively more disadvantaged for images where the AI is ex post incorrect, whereas the more confident are more disadvantaged for images where the AI is ex post correct.

A remaining potential concern with our results is the presence of measurement error in our skills estimates. In a univariate setting, this would manifest as simple attenuation bias working against finding skill effects. However, in our multivariate context, Gillen, Snowberg and Yariv (2019) show that the attenuation bias in one covariate (in our case, individual ability A_i) could lead to a spurious effect of another covariate (calibration C_i) when the latter is correlated and thus picks up on the remnants of the true former effect.

We address this concern in two ways. First, if our measure of calibration were picking up on the effects of imperfectly measured ability, we would also expect to observe this spurious effect among control participants who never worked with AI; however, the coefficient on standardized calibration alone is economically and statistically insignificant across specifications in Table 1. Second, in Appendix A.3 we repeat the preceding analysis with an adaptation of the Obviously Related Instrumental Variables (ORIV) methodology of Gillen, Snowberg and Yariv (2019). Doing so, we find larger effects of calibration on the treatment effect, which while less precise (an expected outcome of this procedure), remain statistically significant at conventional levels.

A final possibility is that our results depend on our choice of performance measures. In Appendix A.4 we repeat the main analysis replacing our measure of ability (1) with AUC (Bamber, 1975; Hanley and McNeil, 1982), and our measure of calibration (2) with an alternative based on the functional form of Grether (1980). We replicate our main findings.

4.3 The Impact of Calibration on Performance Inequality

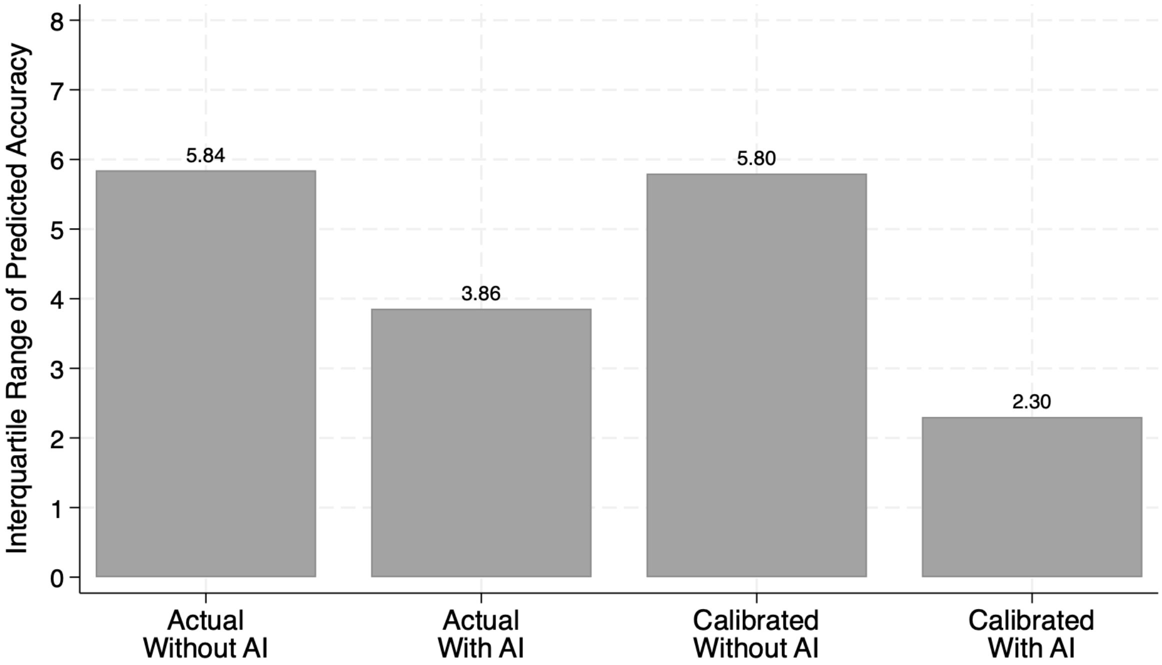

Figure 3: (Counterfactual) Productivity Gap

How does (mis)calibration affect the distribution of productivity when working with AI? Figure 3 presents the results of a counterfactual exercise where we estimate performance inequality across four scenarios. For ease of interpretation, we measure inequality using the the interquartile range (IQR) of predicted performance across participants. In the first two scenarios, we estimate inequality with and without AI, given participants' *actual* (mis)calibration in the experiment.[10]

[10] Specifically, we reestimate equation (3) using unscaled measures of each skill, compute the predicted

We estimate an IQR of 5.84 when participants work without AI, compared to an IQR of 3.86 with AI, representing a 34% reduction in performance inequality from AI assistance. This is consistent with prior studies which find that AI reduces inequality by helping low-performing workers more (Autor, 2024; Brynjolfsson, Li and Raymond, 2023; Noy and Zhang, 2023).

Scenarios 3 and 4 repeat the exercise, but now we simulate the impact of AI assistance in a counterfactual scenario where all participants are perfectly calibrated, meaning their accuracy and confidence take on exactly the same value. We estimate an IQR of 5.80 without AI. Thus, relative to the first scenario, improving calibration has no impact on accuracy when there is no AI assistance. However, with AI assistance, perfect calibration reduces the IQR to 2.30, which represents a 61% reduction in performance inequality relative to baseline (Scenario 1).

Our counterfactual implies that if everyone were perfectly calibrated, working with AI would reduce performance inequality by an additional 27 percentage points beyond the 34 percentage point reduction from AI assistance alone. This is because of the joint distribution of ability and calibration: the low-ability individuals who stand to gain the most from working with AI are also among the least calibrated. A natural set of questions that we leave to future work is if, how, and for whom individual calibration can be trained as a skill for the collaborative nature of work with AI.

5 Conclusion

In this paper we introduce an experimental framework to measure the joint impacts of accuracy and calibration on the benefits of working with AI. We find that AI assistance improves overall prediction accuracy on average, but that the extent and direction of the impact depends on whether participants are appropriately confident about their own ability. We also find that lower ability participants who are well-calibrated benefit the most from AI assistance, consistent with other studies finding that AI helps lower performers

values, and then calculate the IQR implied by performance at the 25th and 75th percentiles. For the third and fourth scenarios, we compute the same statistics but with the coefficients on calibration taking on the values implied by a participant whose accuracy and confidence are exactly equal to each other.

close the gap with their more expert peers (Brynjolfsson, Li and Raymond, 2023; Noy and Zhang, 2023; Autor, 2024).

Finally, we perform a counterfactual analysis that shows if miscalibration was eliminated, performance inequality would shrink nearly twice as much as it does already with the introduction of AI assistance. In addition to showing the importance of calibration in performance inequality, this result might motivate the use of inventions targeting calibration. Among the characteristics that we show are related to performance when working with AI (baseline skill, IQ, and calibration), calibration appears to be a relatively strong candidate for policy and training. For example, training interventions have been shown to improve calibration (Lichtenstein and Fischhoff, 1980; Benson and Önkal, 1992; Haddara and Rahnev, 2022), even with some short interventions lasting less than an hour (Gruetzemacher, Lee and Paradice, 2024). In domains where direct upskilling is costly or time consuming, our results suggest that the combination of calibration training and AI assistance may be particularly valuable.

References

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining human expertise with artificial intelligence: Experimental evidence from radiology." National Bureau of Economic Research.

Autor, David. 2024. "Applying AI to Rebuild Middle Class Jobs." National Bureau of Economic Research.

Bamber, Donald. 1975. "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph." *Journal of Mathematical Psychology*, 12(4): 387–415.

Benjamin, Daniel J. 2019. "Errors in probabilistic reasoning and judgment biases." *Handbook of Behavioral Economics: Applications and Foundations 1*, 2: 69–186.

Benson, P George, and Dilek Önkal. 1992. "The effects of feedback and training on the performance of probability forecasters." *International Journal of Forecasting*, 8(4): 559–573.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond. 2023. "Generative AI at work." National Bureau of Economic Research.

Caplin, Andrew, Dániel Csaba, John Leahy, and Oded Nov. 2020. "Rational inattention, competitive supply, and psychometrics." *The Quarterly Journal of Economics*, 135(3): 1681–1724.

Chan, David C, Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostick Skill: Evidence from Radiologists." *Quarterly Journal of Economics*, 137(2): 729–783.

Choi, Jonathan H., Adam Monahan, and Daniel Schwarcz. 2023. "Lawyering in the age of artificial intelligence." *Available at SSRN 4626276.*

Danz, David, Lise Vesterlund, and Alistair J Wilson. 2022. "Belief elicitation and behavioral incentive compatibility." *American Economic Review*, 112(9): 2851–2883.

Dargnies, Marie-Pierre, Rustamdjan Hakimov, and Dorothea Kübler. 2024. "Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence." *Management Science.*

DeGroot, Morris H, and Stephen E Fienberg. 1983. "The comparison and evaluation of forecasters." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2): 12–22.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey. 2015. "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of experimental psychology: General*, 144(1): 114.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with measurement error: Techniques with applications to the caltech cohort study." *Journal of Political Economy*, 127(4): 1826–1863.

Green, Ben, and Yiling Chen. 2019. "The principles and limits of algorithm-in-the-loop decision making." *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24.

Grether, David M. 1980. "Bayes rule as a descriptive model: The representativeness heuristic." *The Quarterly journal of economics*, 95(3): 537–557.

Gruber, Jonathan, Benjamin R Handel, Samuel H Kina, and Jonathan T Kolstad. 2020. "Managing intelligence: Skilled experts and AI in markets for complex products." National Bureau of Economic Research.

Gruetzemacher, Ross, Kang Bok Lee, and David Paradice. 2024. "Calibration training for improving probabilistic judgments using an interactive app." *Futures & Foresight Science*, e177.

Haddara, Nadia, and Dobromir Rahnev. 2022. "The impact of feedback on perceptual decision-making and metacognition: Reduction in bias but no change in sensitivity." *Psychological Science*, 33(2): 259–275.

Hanley, James A, and Barbara J McNeil. 1982. "The meaning and use of the area under a receiver operator characteristic (ROC) curve." *Radiology*, 143(1): 29–36.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li. 2018. "Discretion in hiring." *The Quarterly Journal of Economics*, 133(2): 765–800.

Hossain, Tanjim, and Ryo Okui. 2013. "The binarized scoring rule." *Review of Economic Studies*, 80(3): 984–1001.

Lichtenstein, Sarah, and Baruch Fischhoff. 1980. "Training for calibration." *Organizational behavior and human performance*, 26(2): 149–171.

Merali, Ahmed. 2024. "Scaling Laws for Economic Productivity: Experimental Evidence in LLM-Assisted Translation." *arXiv preprint arXiv:2409.02391.*

Moore, Don A, and Paul J Healy. 2008. "The trouble with overconfidence." *Psychological review*, 115(2): 502.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science*, 381(6654): 187–192.

Rothe, Rasmus, Radu Timofte, and Luc Van Gool. 2018. "Deep expectation of real and apparent age from a single image without facial landmarks." *International Journal of Computer Vision*, 126(2): 144–157.

Tejeda, Heliodoro, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2022. "AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies." *Computational Brain & Behavior*, 5(4): 491–508.