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

Andrew Caplin[1], David Deming[2], Shangwen Li[1], Daniel Martin[3], Philip Marx[4], Ben Weidmann[2], and Kadachi Ye[1]

Supported by Sloan Foundation Grant "Cognitive Economics at Work"

[1]New York University
[2]Harvard University
[3]University of California, Santa Barbara
[4]Louisiana State University

ESA Columbus
Large Language Models and AI Session

Motivation

- In the future, many workers will use AI tools
- AI can perform a variety of tasks at, or near, the level of human experts, but rarely outperforms the best humans
- Workers can complement AI both with contextual knowledge and the ability to deal with atypical examples
- We propose that these complementarities will be greatest when workers have an accurate appraisal of their own abilities

Calibration

- In line with this, we demonstrate that workers with calibrated beliefs make better use of AI assistance
  - Calibration means that beliefs are aligned with objective likelihoods
  - When 90% sure, a calibrated agent is correct about 90% of the time
- Why would calibration (before receiving AI advice) be important?
  - Overconfident agents might override confident AI predictions
  - Under-confident agents might replace their own better predictions

Question and Challenge

- Question: Do those with better calibrated beliefs make more effective use of AI advice?
- It is challenging to answer this question because of requirements for accurately measuring ability and calibration
- Need probabilistic assessments, clear ground truth, a large sample of participants and number of reports per participant, "right" level of task difficulty, distribution of probability reports to be as even as possible, accuracy and confidence levels to be as balanced as possible across those with and without AI help, etc.

Results

- We design and implement an experiment and empirical strategy to overcome these (and other) challenges
- Our results are clear-cut:
  1. We find individual differences in how well calibrated participants are
  2. Holding ability and IQ fixed, those who are well calibrated have higher performance with AI advice than those who are not
  3. The group that benefits most from working with AI are those of lower ability whose beliefs are well calibrated

Potential Impact

- How will AI affect worker productivity and labor market inequality?
- Working with AI has been shown to reduce the productivity gap between high and low ability workers
- We show that if miscalibration was eliminated, the productivity gap would shrink nearly twice as much as it does already

Contribution

- Mounting evidence suggests that AI adoption has heterogeneous impacts across individuals

Contribution

- Mounting evidence suggests that AI adoption has heterogeneous impacts across individuals
- Our contribution is to examine whether this heterogeneity is a function of measurable, individual differences in skills

Contribution

- Mounting evidence suggests that AI adoption has heterogeneous impacts across individuals
- Our contribution is to examine whether this heterogeneity is a function of measurable, individual differences in skills
- Using controlled experiment, examine extent that differences in AI value-add are driven by ability and belief calibration

Contribution

- Mounting evidence suggests that AI adoption has heterogeneous impacts across individuals
- Our contribution is to examine whether this heterogeneity is a function of measurable, individual differences in skills
- Using controlled experiment, examine extent that differences in AI value-add are driven by ability and belief calibration
(the 'ABCs of who benefits')

Contribution

- Mounting evidence suggests that AI adoption has heterogeneous impacts across individuals
- Our contribution is to examine whether this heterogeneity is a function of measurable, individual differences in skills
- Using controlled experiment, examine extent that differences in AI value-add are driven by ability and belief calibration
  (the 'ABCs of who benefits')
- This requires bringing together several literatures:
  - Experimental economics: human-AI interactions
  - Labor economics: impacts of working with AI
  - Cognitive economics: belief calibration

Experiment: Task

- In each of 160 rounds, participants were presented with an image of a human face
- They were asked to report the probability that the individual was over 21 years old at the time the image was taken
- They were informed that all images were taken between 2010 and 2014 and half of the images depicted individuals who were over 21

Experiment: Task

What is the probability that the person in this image was **over** 21 years old?

Time left to complete this page: 52

Click anywhere on the bar to move around your choice.

Under 21 **Over** 21

Submit

Round 1 out of 160

Experiment: Task

What is the probability that the person in this image was **over** 21 years old?

Time left to complete this page: 44

Click anywhere on the bar to move around your choice.

Under 21 — **Over** 21

Probability over 21 in image is

74%

Submit

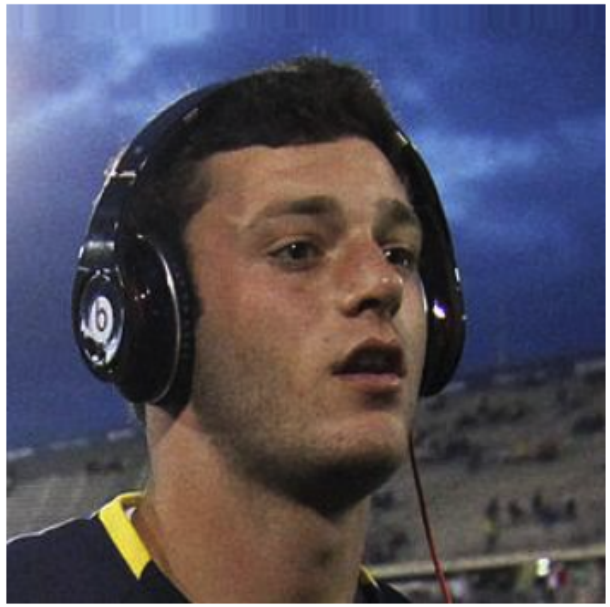
Round 1 out of 160

Experiment: AI Assistant

- Participants in the treatment group were also shown an "AI Assistant guess" in half of their rounds
- This was an AI prediction of whether the individual was over 21 (a confidence score between 0% and 100%)
- Participants were told that the AI Assistant was more accurate than the average human, but worse than the most skilled human

Experiment: Task w/ AI Assistant

What is the probability that the person in this image was **over** 21 years old?



Time left to complete this page: 51

AI Assistant's guess is

62%

Click anywhere on the bar to move around your choice.

Under 21 **Over** 21

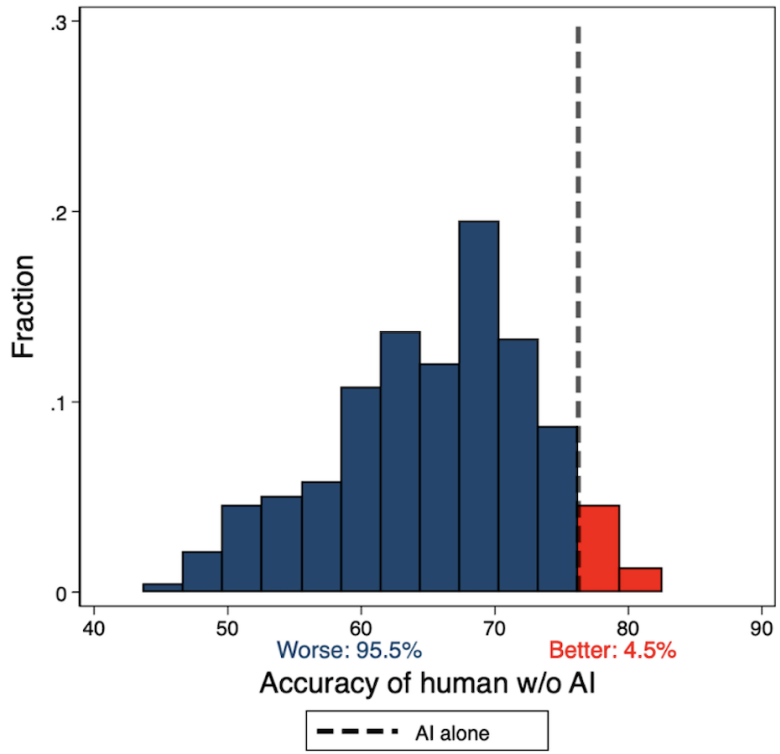
Submit

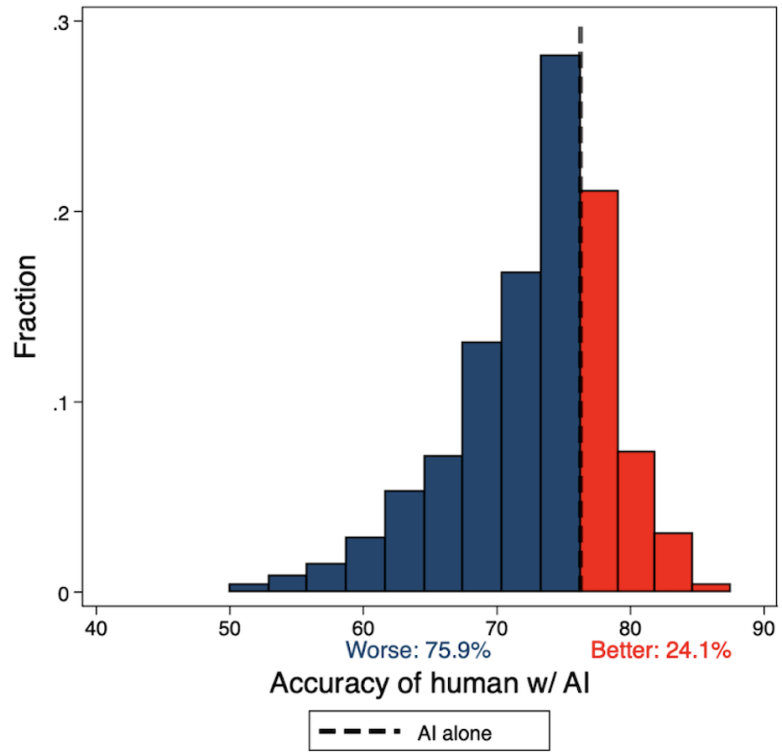
Round 21 out of 160

Experiment: Other Details

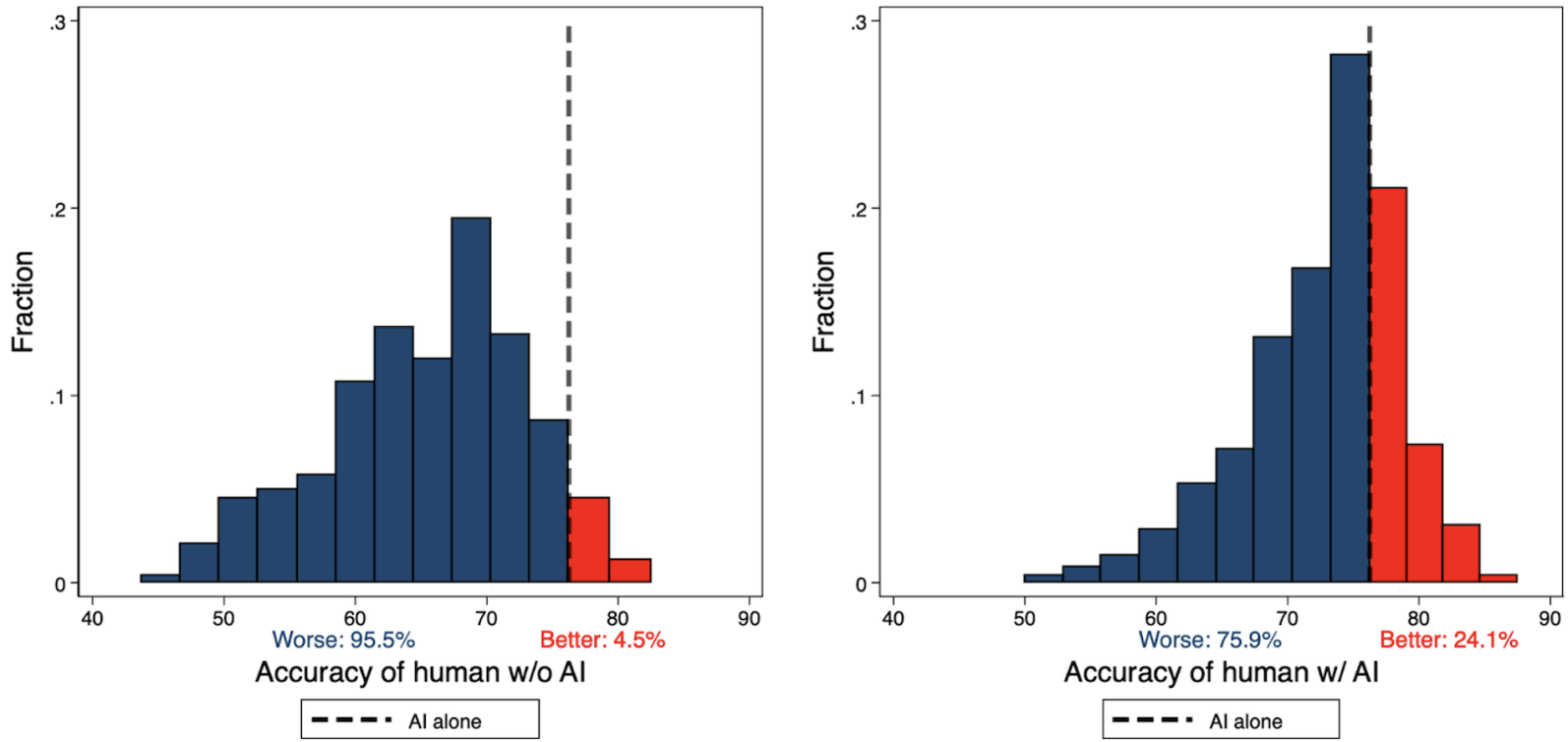
- Ran on Prolific: 497 participants in treatment, 242 in control
- 4 practice rounds and no feedback in practice or incentivized rounds
- 160 incentivized rounds followed by incentivized Raven's Progressive Matrices (a measure of IQ)
  - 80 *untreated* rounds + 80 *treated* rounds (w/ AI if treatment group)
- $5 participation fee for finishing the experiment
- $5 bonus payment using binarized scoring rule (told likelihood of receiving bonus was maximized by truthful reporting their belief)
- $1 bonus payment if random matrix was correct

Performance is average correctness (correct = report closer to truth)



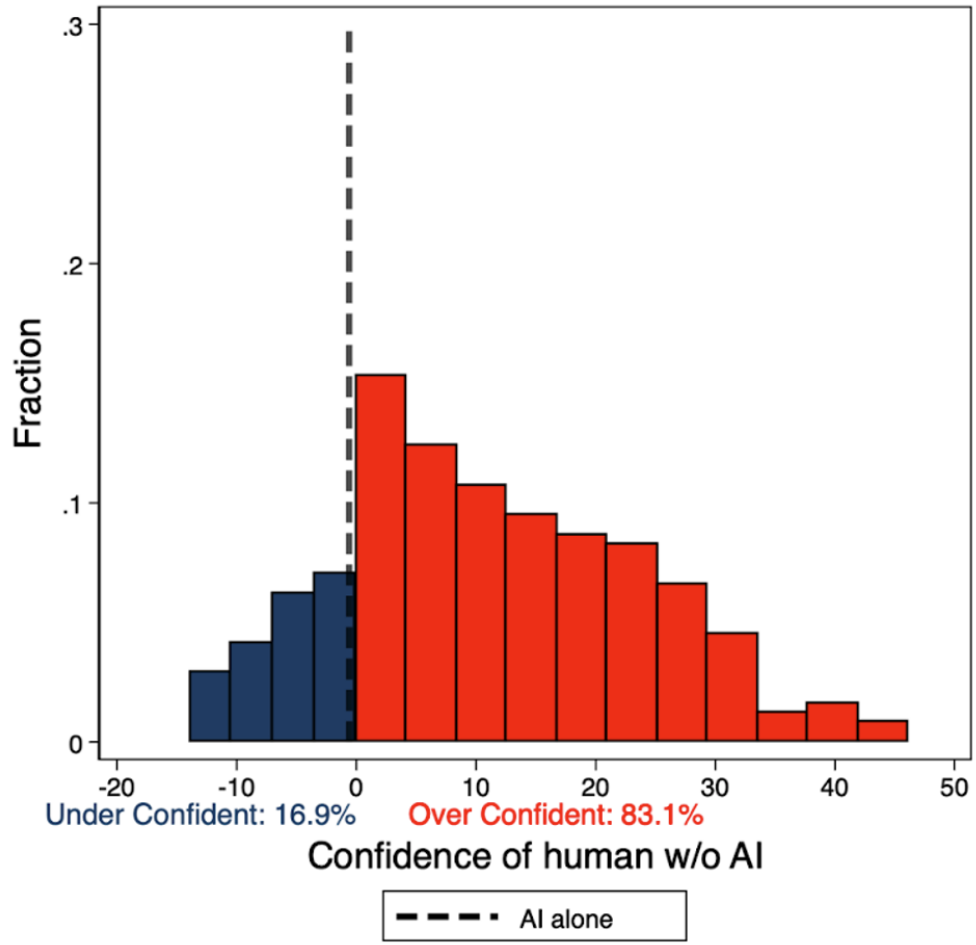


Performance is average correctness (correct = report closer to truth)

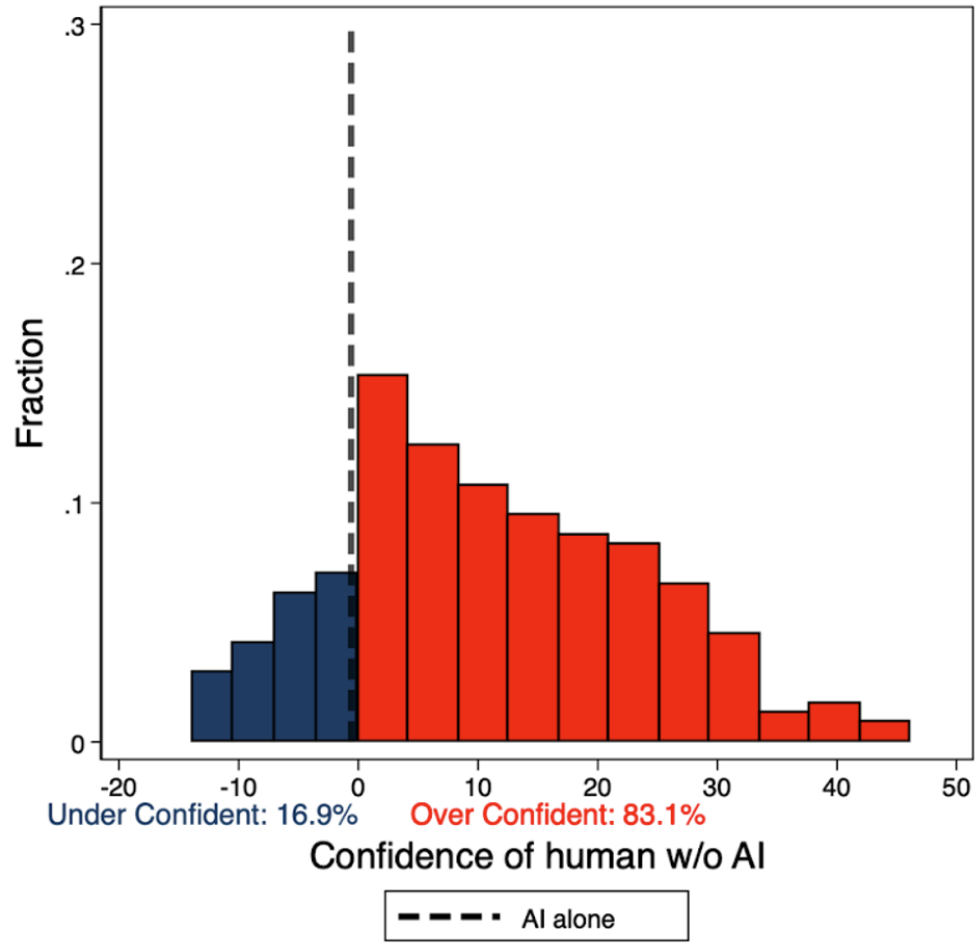


Performance without AI is our measure for ability & performance with AI is our outcome of interest

Confidence is average of report confidence minus its correctness



Confidence is average of report confidence minus its correctness



Result #1: Individual differences in calibration (= |confidence|)

Is calibration related with performance without AI?

	Accuracy in treated images
Treatment	6.27
	(0.47)
Accuracy in untreated images	3.60
	(0.45)
Accuracy in untreated images × Treatment	-1.83
	(0.57)
Calibration in untreated images	0.11
	(0.43)
Calibration in untreated images × Treatment	1.40
	(0.54)
IQ	1.17
	(0.43)
IQ × Treatment	-1.01
	(0.51)
Constant	65.88
	(0.39)
Observations	732

Is calibration related with performance with AI?

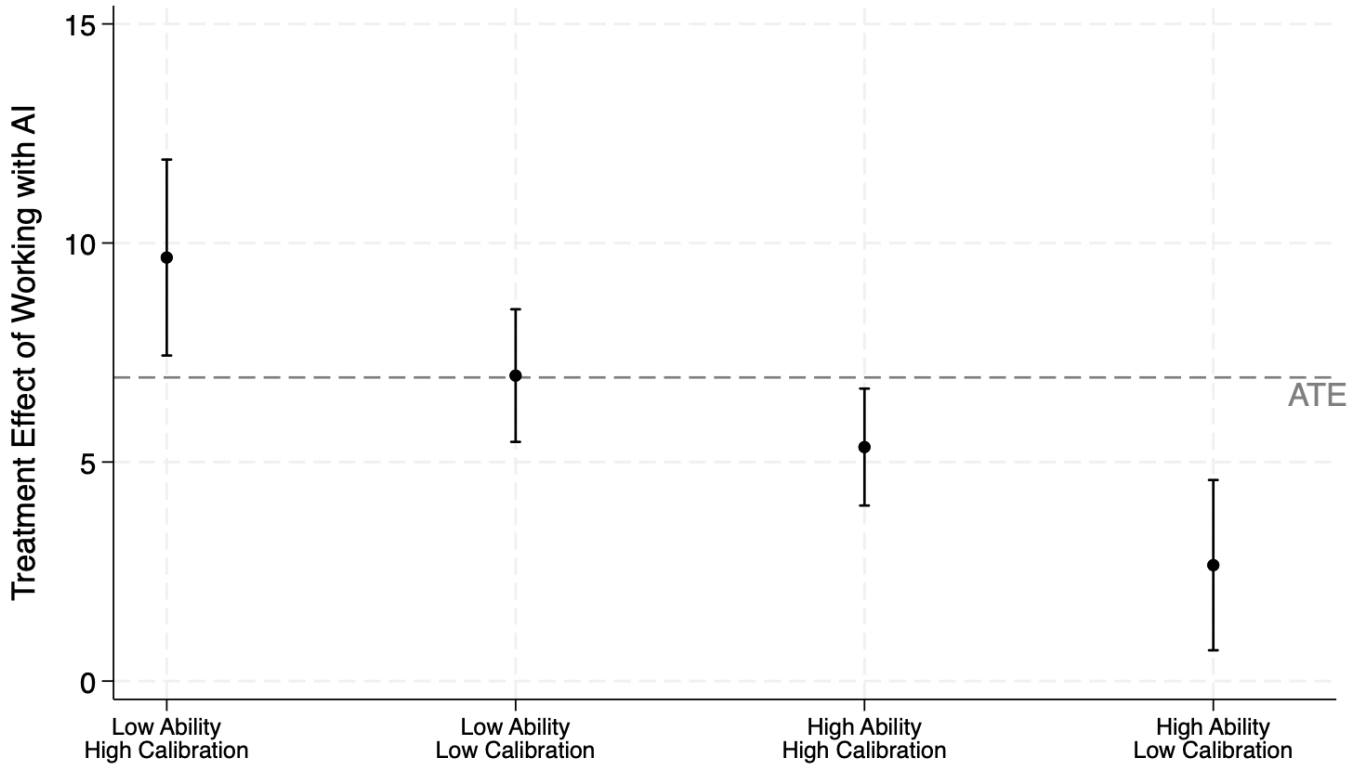
	Accuracy in treated images
Treatment	6.27
	(0.47)
Accuracy in untreated images	3.60
	(0.45)
Accuracy in untreated images × Treatment	-1.83
	(0.57)
Calibration in untreated images	0.11
	(0.43)
Calibration in untreated images × Treatment	1.40
	(0.54)
IQ	1.17
	(0.43)
IQ × Treatment	-1.01
	(0.51)
Constant	65.88
	(0.39)
Observations	732

Is calibration related with performance with AI?

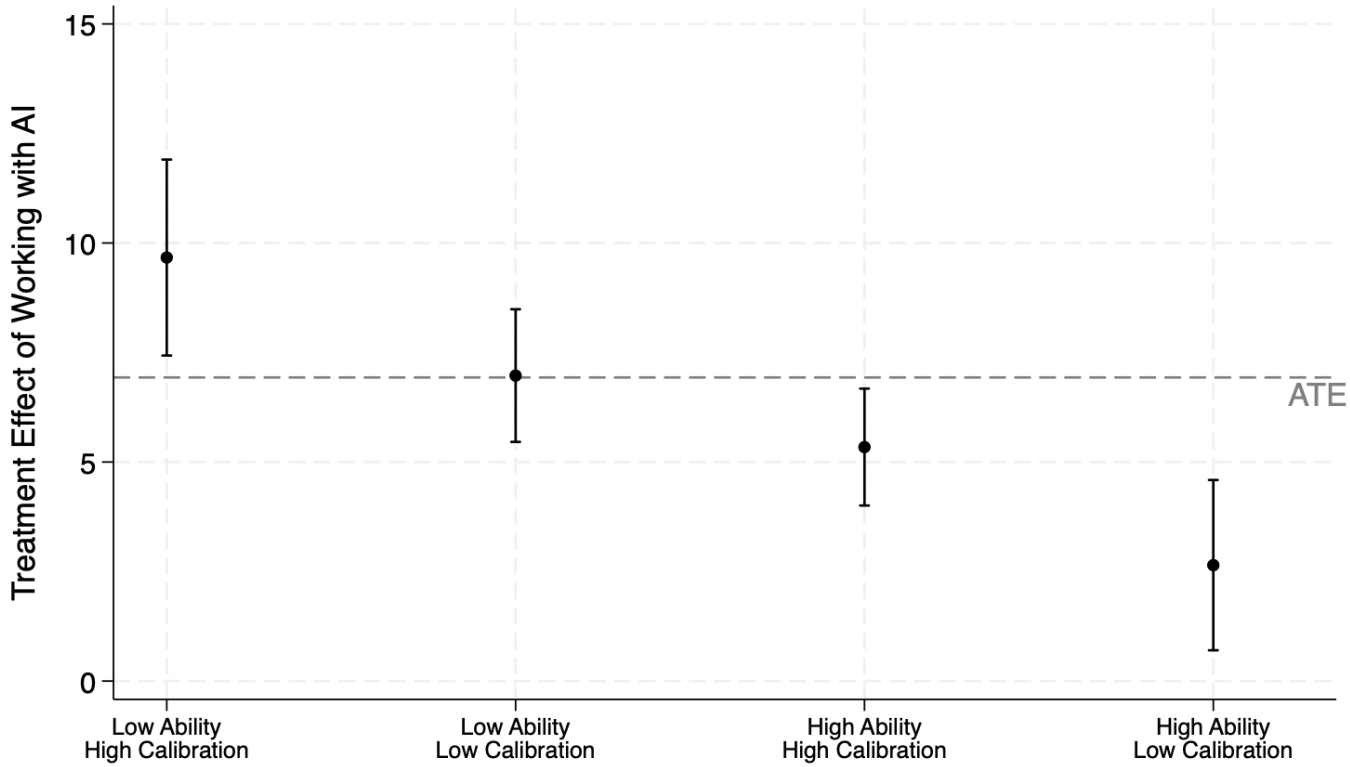
	Accuracy in treated images
Treatment	6.27
	(0.47)
Accuracy in untreated images	3.60
	(0.45)
Accuracy in untreated images × Treatment	-1.83
	(0.57)
Calibration in untreated images	0.11
	(0.43)
Calibration in untreated images × Treatment	1.40
	(0.54)
IQ	1.17
	(0.43)
IQ × Treatment	-1.01
	(0.51)
Constant	65.88
	(0.39)
Observations	732

Result #2: Holding ability and IQ fixed, those who are well calibrated have higher performance with AI advice than those who are not

Who is affected?

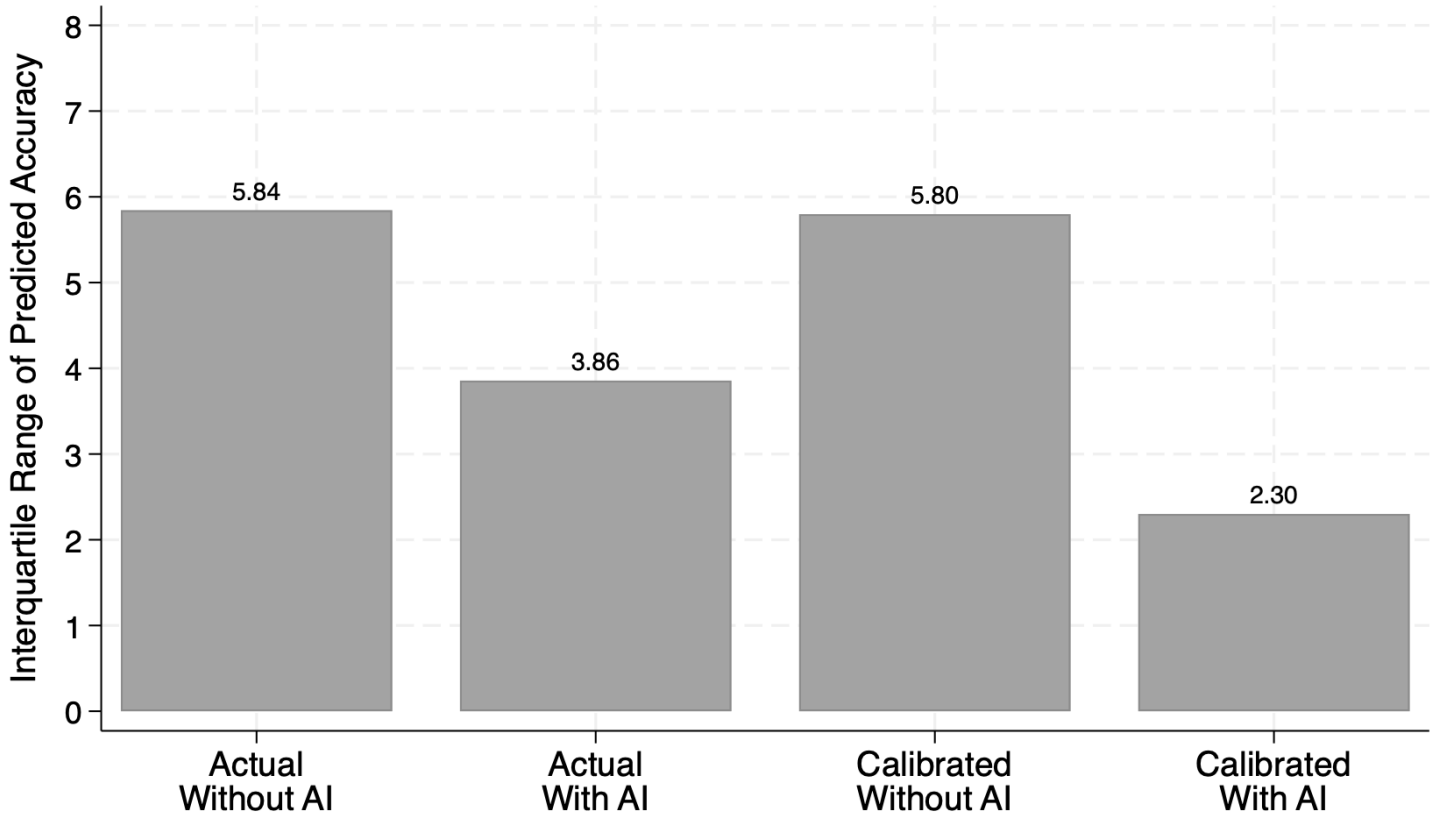


Who is affected?



Result #3: Those of lower ability whose beliefs are well calibrated benefit the most from AI advice

Counterfactual: If miscalibration eliminated, productivity gap would shrink nearly twice as much as it does already



IQR = interquartile range (25th to 75th percentile)

Summary

- Our results are clear-cut:
  1. We find individual differences in how well calibrated participants are that are positively correlated with ability
  2. Holding ability fixed, those who are well calibrated have higher performance with AI advice than those who are not
  3. The group that benefits most from working with AI are those of lower ability whose beliefs are well calibrated
- Counterfactual: If miscalibration eliminated, productivity gap would shrink nearly twice as much as it does already

Thank you!

Confidence and Agreement

One percentage point of confidence corresponds to a 0.21 percentage point reduction in probability of agreement with AI

	Agree
Accuracy w/o AI	0.107
	(0.0979)
Net confidence w/o AI	-0.210
	(0.0538)
Confidence of AI	0.436
	(0.0177)
Correct AI	15.21
	(0.610)
Constant	29.07
	(6.927)
Observations	39106

Summary Stats

	Control	Treatment	P-value
Total Time (Minutes)	34.14	34.99	0.44
	(13.42)	(14.13)	
Bonus ($)	4.27	4.29	0.90
	(2.16)	(2.08)	
IQ (0-14)	5.03	4.86	0.42
	(2.80)	(2.79)	
Age	40.74	40.45	0.79
	(14.39)	(13.56)	
Female	0.56	0.58	0.69
	(0.50)	(0.49)	
Nonwhite	0.39	0.35	0.35
	(0.49)	(0.48)	
Sample Size	239	493	