Recalibrated Wisdom of Crowds in Detection of AI-Generated Images

Phillip Hegeman, Jennifer S. Trueblood

DEPARTMENT OF PSYCHOLOGICAL AND BRAIN SCIENCES

Computational Decision Making Lab, Indiana University Bloomington

Contact: phegeman@iu.edu

Introduction

- AI-generated images: easier than ever to create

WITH GREAT POWER COMES GREAT RESPONSIBILITY! [1]

- Potential for harmful uses, e.g., spread of misinformation
- Harm is largely conditional on inability to detect fake images

How difficult is it really?
You decide: **real human** or **AI-generated image** (StyleGAN2)? [3]

A.

B.

C.

Individual judgement of face authenticity:

- **near chance overall:** average 2AFC accuracy 48.2% (n=315) [2]
- **worse than chance for White AI faces:** accuracy 31.5% [3]

Question:
Can the Wisdom of Crowds succeed in such a difficult perceptual judgment task?

Methods - Data

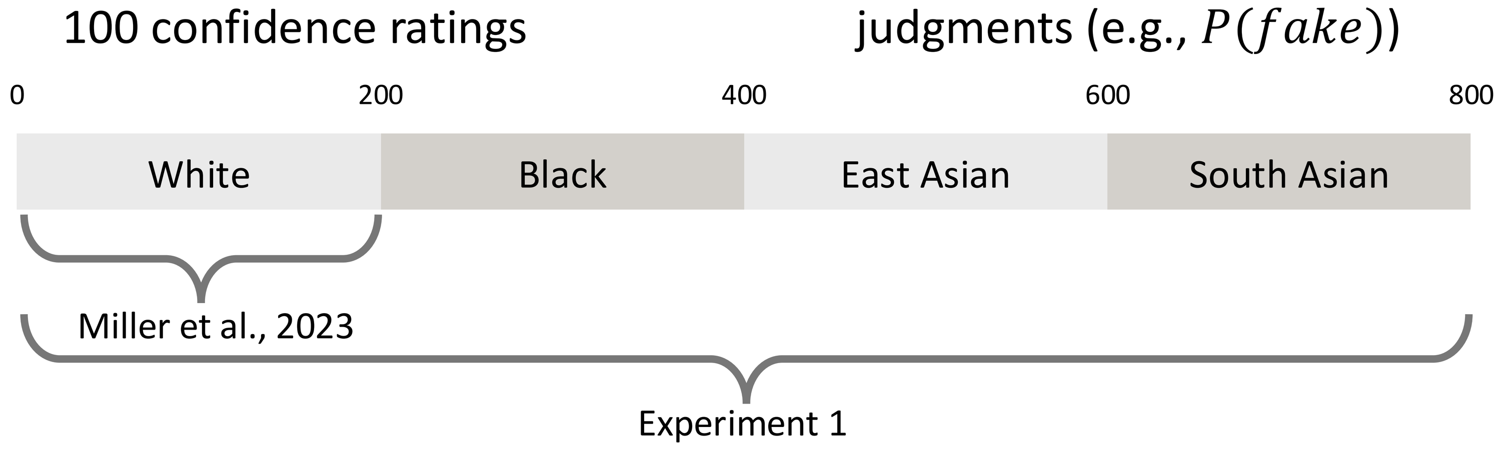
To investigate, we found open access data with judgments of face authenticity, and collected our own data with some key differences

Miller et al., 2023

- $n = 121$ participants, Mturk
- binary judgments followed by 0-100 confidence ratings

Our Experiment

- $n = 147$ participants, Mturk
- continuous 0-100 probability judgments (e.g., $P(fake)$)



Stimuli: face images from Nightingale and Farid, 2022 [2]
400 AI-generated (StyleGAN2), 400 real faces from AI training set

Methods - Wisdom of Crowds

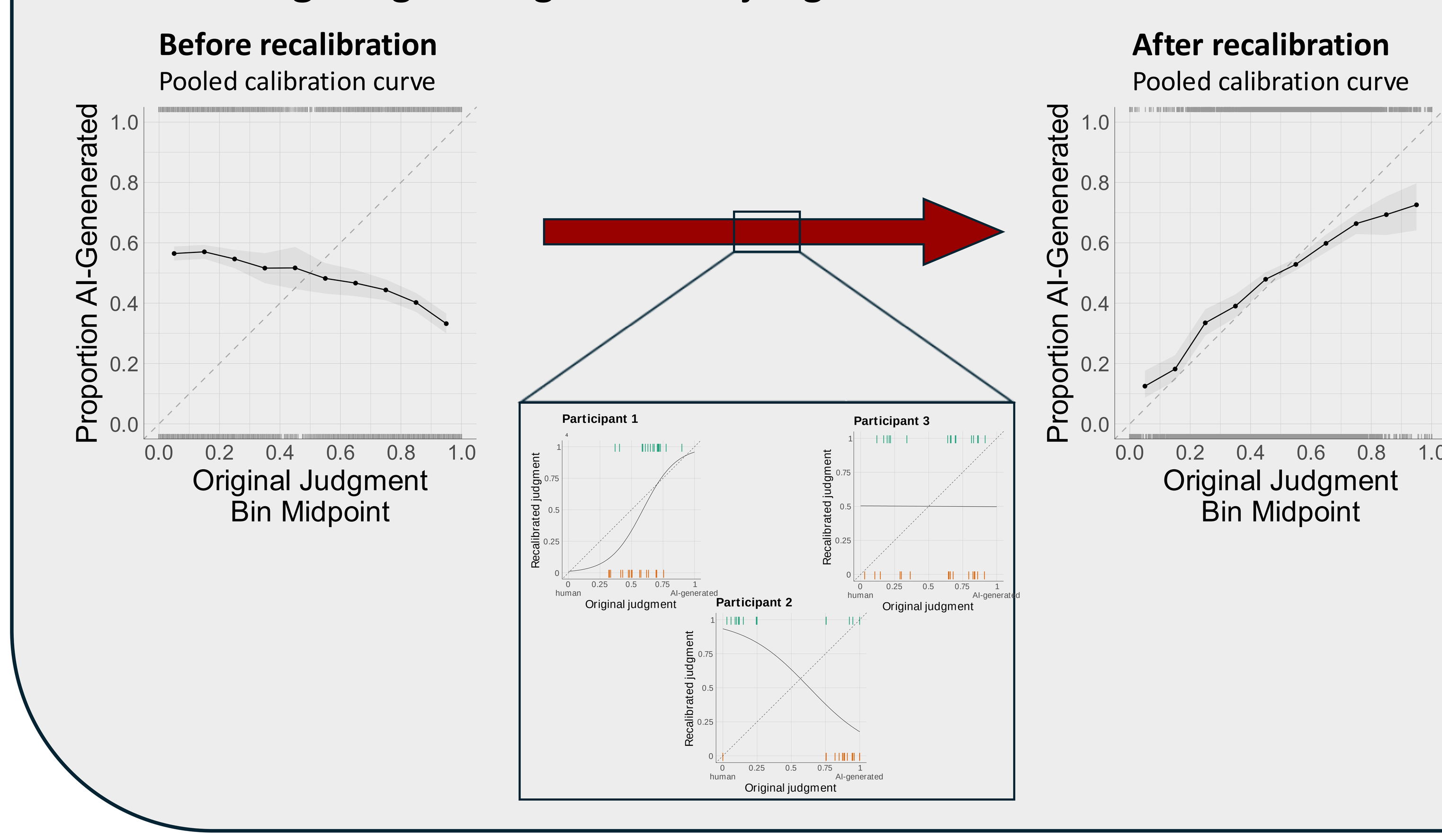
Intuition: diverse collections of (independent) judges often outperform individuals and even experts through elimination of random errors and competing biases

Possible implementations:

- **Simple crowd:** mean of judgments
  - Key idea: every member contributes equally
- **Performance-based crowd:** weighted mean of judgments
  - Key idea: pay more attention to the most able individuals
  - proportional **accuracy weighting** or **chose top-n** most accurate
- **Recalibrated crowd:** mean of recalibrated judgments

Recalibration

- Building on Turner et al., 2014 [4]
  - Their main idea: "…we might improve forecast aggregation by correcting for forecasters' systematic biases"
- **Platt scaling** – logistic regression of judgment onto truth



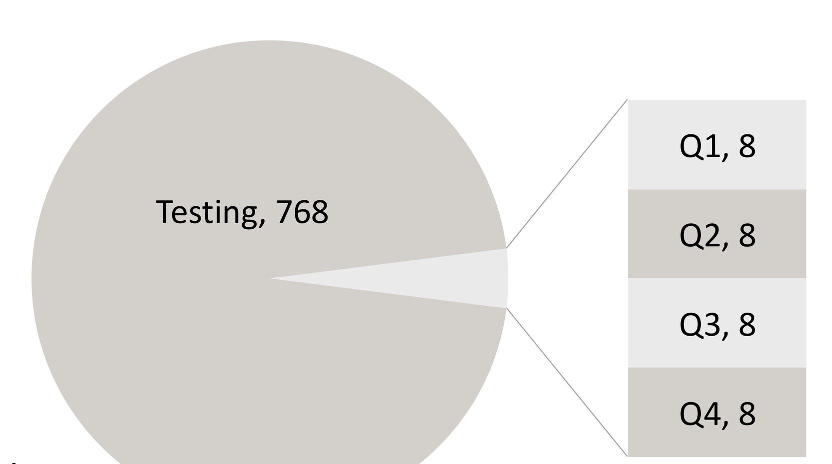
Common design

- **Holdout** for accuracy calculation or recalibration
  - 32 images
  - balanced real/AI
  - balanced difficulty
- **Testing** for evaluation

Miller et al, 2023: resampled 2,000 holdout sets meeting above criteria

Our Experiment: single holdout set judged by all participants

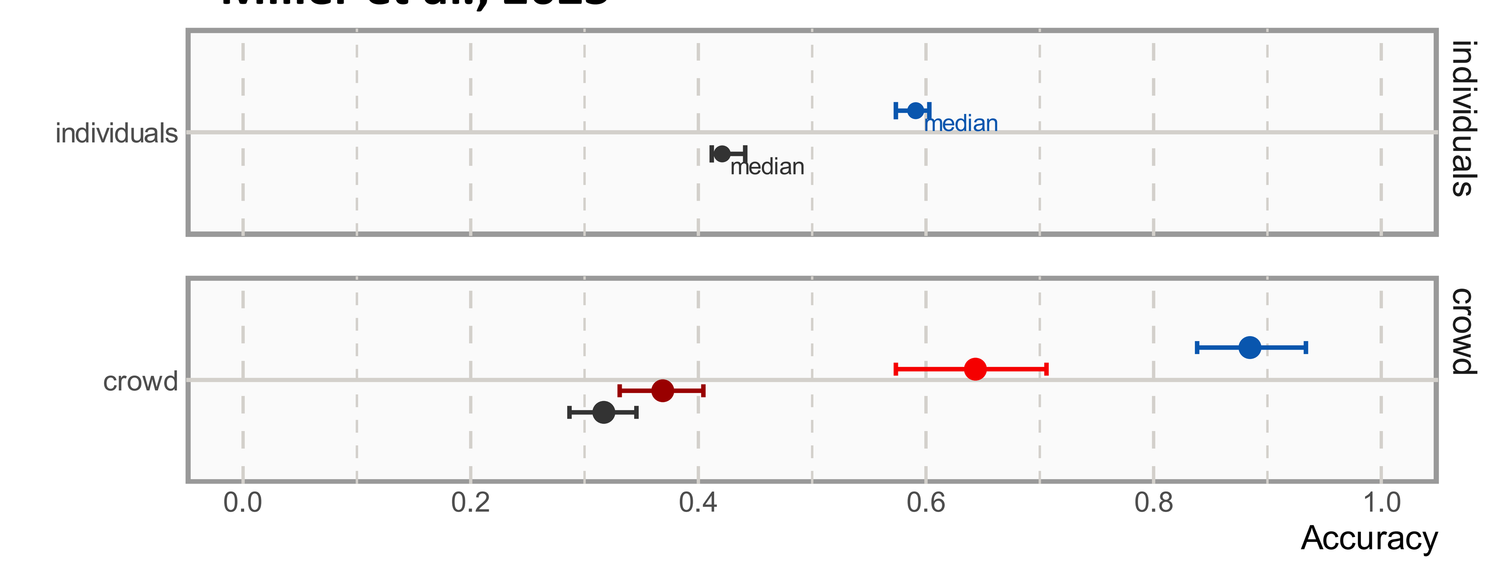
Experiment 1 Stimuli Breakdown



Holdout, 32
Quartiles based on difficulty (correct judgment rate[2])

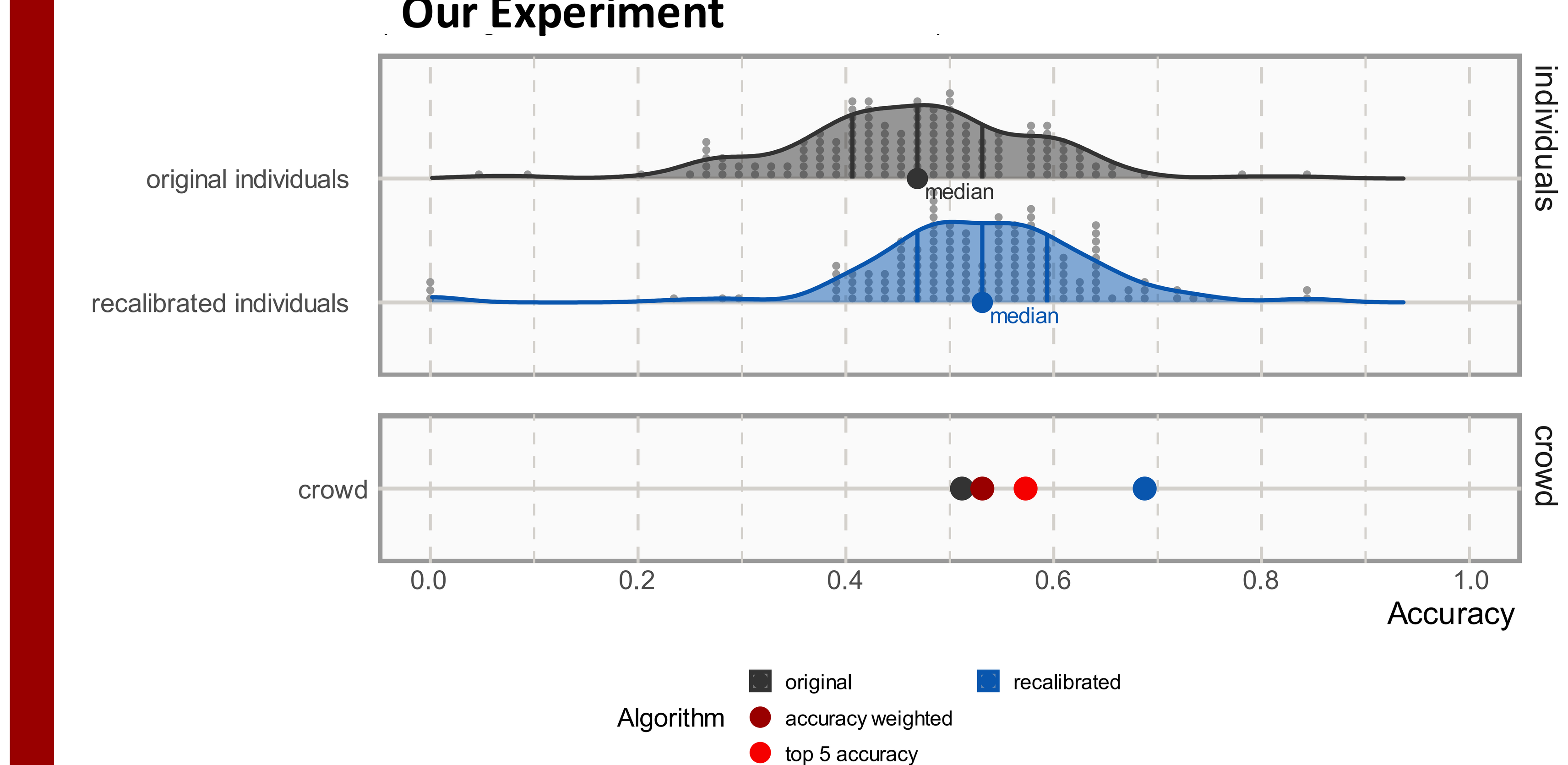
Results

Miller et al., 2023



Mean accuracies and 95% bootstrapped CI calculated via resampling holdout sets of images.

Our Experiment



Conclusions

- In a difficult perceptual judgment task
  - Simple wisdom of crowds **fails**
  - Common performance-based weighting methods have **some benefit**
  - **Recalibration succeeds** by correcting individuals' systematic errors
- Even a very simple recalibration model, fit to little data, works well
- Performance can depend on the difficulty homogeneity of stimuli

References

1. Lee, S. (1962). Amazing Fantasy #15 – Uncle Ben's advice to Peter Parker
2. Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119.
3. Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. (2023). AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science*, *34*(12), 1390–1403.
4. Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*(3), 261–289.