On the Fairness of Machine-Assisted Human Decisions

Talia Gillis
Columbia Law

Bryce McLaughlin
Stanford GSB

Jann Spiess
Stanford GSB

arxiv.org/abs/2110.15310

Lab for Inclusive FinTech LIFT–IFC conference, December 2023

Algorithms in high-stakes decisions

The New York Times

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.

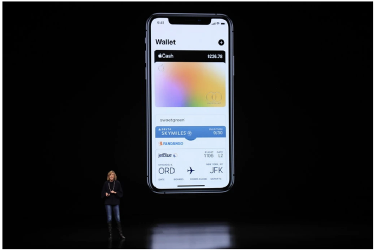
By Neil Vigdor

Nov. 10, 2019

...

"My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time," Mr. Hansson wrote Thursday on Twitter. "Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does."

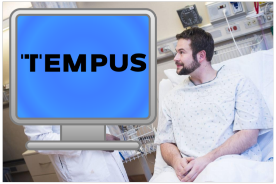
...



Jennifer Bailey, vice president of Apple Pay. Regulators are investigating Apple Card's algorithm, which is used to determine applicants' creditworthiness. Jim Wilson/The New York Times

Algorithms in high-stakes decisions

Algorithms diagnose diseases



AI-DRIVEN TO
TEST

The Tempus Tumor Origin (TO) test uses tumor RNA expression results to predict the patient's most likely cancer type(s) from 68 possible cancer types. The Tempus TO test was developed using a large internal database of clinical and annotated molecular tumor data.

Algorithms set bail

PTRA risk categories	Number of defendants	Any adverse event
PTRA One	28,033	20:1
PTRA Two	24,017	9:1
PTRA Three	20,992	4:1
PTRA Four	9,836	2:1
PTRA Five	2,491	2:1

Algorithms hire employees

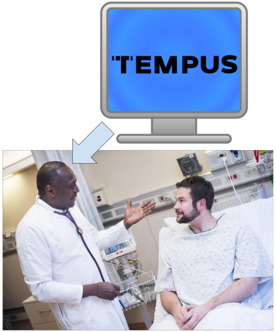
Uniquely predicts candidate interest and conversion

Ensures fair and unbiased passive talent qualification

Literature on defining fairness, diagnosing bias, fairness/accuracy trade-offs. But...

Algorithms *assisting* in high-stakes decisions

Doctors diagnose diseases



Judges set bail

Managers hire employees

Not just a design decision, often a legal requirement

Main idea

Commonly analyze algorithms through lens of direct implementation (*automation*)

$$\text{training data} \xrightarrow{\text{machine}} \text{decision}$$

But often algorithm provides *assistance* to decision-maker who retains authority

$$\text{training data} \xrightarrow{\text{machine}} \text{prediction} \xrightarrow{\text{human}} \text{decision}$$

This project: How does design of algorithm affect decision accuracy and fairness when algorithm *assists* a possibly biased decision-maker rather than *automates*?

1. Thought experiment
2. Online lab experiment

Automation

Accuracy vs *fairness* trade-off when including/excluding sensitive covariates

Assistance

Exclusion may hurt rather than help, for *accuracy* and *fairness*

Related work ▸ Empirical evidence

- **Algorithmic fairness:** Kleinberg et al. (2016); Chouldechova (2016); Lakkaraju et al. (2017); Jiang and Nachum (2020); Liang et al. (2021)
- **Regulation:** Bent (2019); Huq (2020); Gillis (2022); Enarsson et al. (2022); Kim (2022)
- **Statistical communication, delegation:** Kamenica and Gentzkow (2011); Spiess (2018); Athey et al. (2020); Andrews and Shapiro (2020); Ibrahim et al. (2021)
- **Sources of bias:** Bordalo et al. (2019); Bohren et al. (2019a,b); Coffman et al. (2021)
- **Empirical analysis of human–machine interaction:** Dietvorst et al. (2018); Green and Chen (2019); Stevenson and Doleac (2019); Imai et al. (2020); De-Arteaga et al. (2020); Lai et al. (2021); Ludwig and Mullainathan (2021); Bastani et al. (2021); Fogliato et al. (2022); Snyder et al. (2022); Donahue et al. (2022)

Algorithmic risk assessments in felony sentencing

1. changes sentencing
2. does not lower prison populations, risk to public safety
3. does not seem to improve racial disparities in sentencing

1. Setup and model
2. Implications for fairness–accuracy trade-offs
3. Lab experiment
4. Summary and conclusion

Human–machine interaction as a game between principal and agent

$(Y, X, G, H) \sim \mathrm{P}$ $X \in \mathcal{X}$ discrete $G \in \mathcal{G}=\{M, F\}$

- **Principal:** designs algorithm
  - Algorithm maps training data to predictions $\hat{f}(x, g)$; here, consider:
    1. Group-blind average given $X{=}x$ only: $\hat{f}_{-}(x) = \widehat{\mathrm{E}}[Y|X{=}x] = \frac{\sum_{X_i=x} Y_i}{\sum_{X_i=x} 1}$
    2. Group-aware average given $X{=}x$, $G{=}g$: $\hat{f}_{+}(x, g) = \widehat{\mathrm{E}}[Y|X{=}x, G{=}g]$
- **Agent:** takes decision
  - Observes an instance (X, G, H) and algorithmic prediction $\hat{Y} = \hat{f}(X, G)$
  - Takes a decision $\hat{D} = h(X, G, H, \hat{Y})$
  - Has a subjective model/belief of the world (prior P^* over distribution P)

Preferences and decisions

$$\text{data } (Y_i, X_i, G_i) \xrightarrow{\text{machine}} \text{prediction } \hat{Y} = \hat{f}(X, G) \xrightarrow{\text{human}} \text{decision } \hat{D} = h(X, G, H, \hat{Y})$$

- **Principal:** trades off accuracy and fairness

  Accuracy (risk)
  $$r(h) = \mathrm{E}[(Y - \hat{D})^2]$$

  Fairness (conditional disparity)
  $$\Delta(h, X^*) = \mathrm{E}[\hat{D}|G{=}M, X^*] - \mathrm{E}[\hat{D}|G{=}F, X^*]$$

- **Agent:** maximizes accuracy
  - Minimize risk $r(h) = \mathrm{E}[(Y - \hat{D})^2]$, averaged over P^*
  - Optimal decision: $\hat{D} = \mathrm{E}^*[Y|X, G, H, \hat{Y}]$

1. Setup and model

2. Implications for fairness–accuracy trade-offs

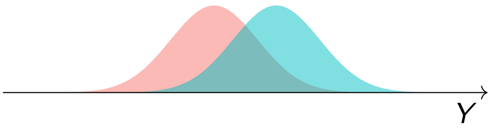
3. Lab experiment

4. Summary and conclusion

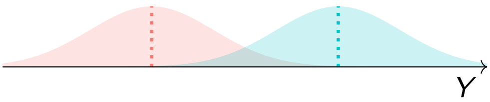
Illustration in a (very) simple example

data (Y_i, X_i, G_i) $\xrightarrow{\text{machine}}$ prediction $\hat{Y} = \hat{f}(X, G)$ $\xrightarrow{\text{human}}$ decision $\hat{D} = \mathrm{E}^*[Y|X, G, H, \hat{Y}]$

Actual data



Prior belief

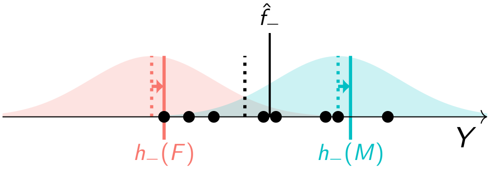


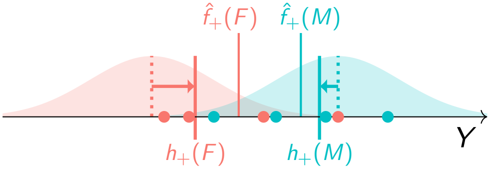
Data $Y|G{=}g \sim \mathcal{N}\left(\mu(g), \sigma^2\right)$
with $\mu(g)$ unknown

Prior $\mu(g) \sim \mathcal{N}\left(\pi(g), \tau^2\right)$
independent across g

$\mathrm{P}(G{=}F) = 1/2 = \mathrm{P}(G{=}M) \quad n(M) = n/2 = n(F)$

$$\text{data } (Y_i, G_i) \xrightarrow{\text{machine}} \text{prediction } \hat{Y} = \hat{f}(G) \xrightarrow{\text{human}} \text{decision } \hat{D} = \mathrm{E}^*[Y|\hat{Y}, G]$$





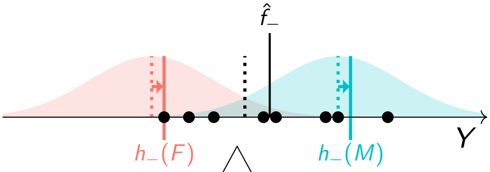
$$\Delta(\hat{f}_-) = 0$$

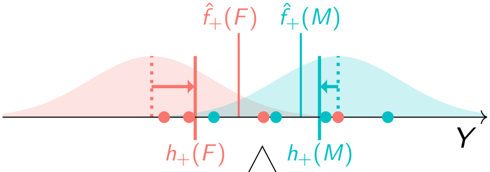
$$\Delta(h_-) = \underbrace{\Delta^*}_{\text{prior disparity}} = \mathrm{E}^*[Y|G{=}M] - \mathrm{E}^*[Y|G{=}F]$$

$$\Delta(\hat{f}_+) = \overbrace{\Delta_Y}^{\text{true disparity}} = \mathrm{E}[Y|G{=}M] - \mathrm{E}[Y|G{=}F]$$

$$\Delta(h_+) = \underbrace{\frac{2\sigma^2}{n\,\tau^2+2\sigma^2}}_{\text{weight } w \,\to\, 0} \Delta^* + \underbrace{\frac{n\,\tau^2}{n\,\tau^2+2\sigma^2}}_{\text{weight } 1-w \,\to\, 1} \Delta_Y$$

data (Y_i, G_i) $\xrightarrow{\text{machine}}$ prediction $\hat{Y} = \hat{f}(G)$ $\xrightarrow{\text{human}}$ decision $\hat{D} = \mathrm{E}^*[Y|\hat{Y}, G]$





This prediction does not include info about gender differences, so I'll use my belief

This prediction includes gender-specific info; I'll update my belief about differences

Take-aways from the Normal illustration

General results

$$\text{data } (Y_i, G_i) \xrightarrow{\text{machine}} \text{prediction } \hat{Y} = \hat{f}(G) \xrightarrow{\text{human}} \text{decision } \hat{D} = \mathrm{E}^*[Y|\hat{Y}, G]$$

- Biased decision-maker: $\underbrace{\Delta^*}_{\text{prior disparity}} > \underbrace{\Delta_Y}_{\text{true disparity}}$
- True differences in large sample: $\Delta_Y \neq 0$, $n \to \infty$

Group-blind predictions $\hat{Y} = \hat{f}_-$

Automation: $\hat{Y}$ has less disparity and is less accurate

Assistance: $\hat{D}$ has more disparity and is less accurate

Group-aware predictions $\hat{Y} = \hat{f}_+(G)$

Automation: $\hat{Y}$ has more disparity and is more accurate

Assistance: $\hat{D}$ has less disparity and is more accurate

1. Setup and model

2. Implications for fairness–accuracy trade-offs

3. Lab experiment

4. Summary and conclusion

- **Our experiment:** 1250 online study subjects predict performance of female, male test-takers on math test
- **Baseline data:** Math scores (six questions) of test takers from Cecilia Ridgeway and Tamar Kricheli-Katz, “Behavioral responses to the changing world of gender”

Shelley cuts 3 cucumbers in 5 minutes.
David cuts 4 tomatoes in 7 minutes.
Shelley cut cucumbers and Davis cut tomatoes for 35 minutes.
How many cucumbers and tomatoes (in total) did Shelley and David cut?

- ○ 60
- ○ 52
- ○ 45
- ○ 41

	Obs	**Mean**	**SD**
Male	207	2.20	1.46
Female	189	2.58	1.56

Average conditional on age

Subject 322 : His age is between 50 and 55 years old and he **has** a 4-year college degree.

Assistant: The average (mean) score we have seen from men and women over 45 years old (125 observations) is 44.7%.

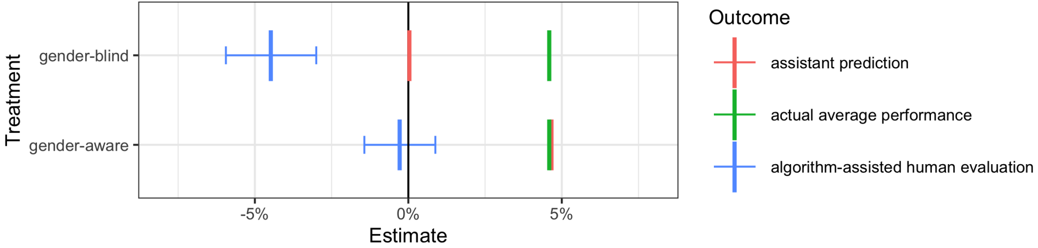
What do you think he (Subject 322) scored on the math test?

Average conditional on age and gender

Subject 322 : His age is between 50 and 55 years old and he **has** a 4-year college degree.

Assistant: The average (mean) score we have seen from men over 45 years old (62 observations) is 40.3%.

What do you think he (Subject 322) scored on the math test?



Weighted for test-taker population distribution by age bracket, education, gender; standard error estimates clustered at subject level

- No evidence for strong **explicit** bias
- **Implicit** bias by failing to adjust for differences in ability–education relationship
- Exclusion has unintended consequences

1. Setup and model

2. Implications for fairness–accuracy trade-offs

3. Lab experiment

4. Summary and conclusion

Summary and conclusion ‣ Generalizations

- We model the relationship between design of algorithm and resulting fairness properties of machine-assisted human decisions with biased beliefs
- We illustrate that common trade-offs between fairness and accuracy may revert
- We provide evidence for reversal in lab study
- Narrowly, adds another reason to be skeptical about input restrictions
- More broadly, a need for modeling context, beliefs, preferences, and frictions when analyzing human-machine decisions

Thank you!

jspiess@stanford.edu

Backup Slides

- When algorithms assist human decision-makers with decision authority, they often affect human decisions, but do not necessarily improve decisions
- Stevenson and Doleac (2019): Algorithmic risk assessments in felony sentencing
  1. changes sentencing
  2. does not lower prison populations, risk to public safety
  3. does not seem to improve racial disparities in sentencing
- Ludwig and Mullainathan (2021): Pre-trial release decisions in NYC

Table 4

Results from Algorithm for Pre-trial Release Decisions in New York City

	Release recommendations under old tool	*Release recommendations under new tool*	*Judge release decisions under new tool (2019–20 data)*
Black defendants	31.7%	83.9%	69.4%
White defendants	41.1%	83.5%	72.0%
Black-White gap	9.4 percentage points	0.4 percentage points	2.6 percentage points

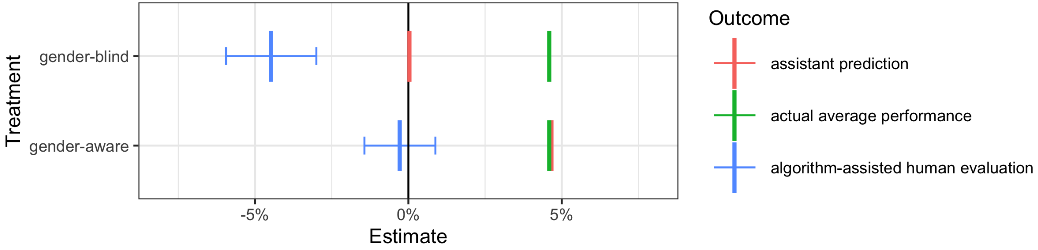
Source: Peterson (2020). The new algorithmic tool was built by the University of Chicago Crime Lab in partnership with Luminosity and the NYC Criminal Justice Agency.

Definition (δ-disparate beliefs). The decision-maker's belief P^* about means at $X{=}x$ assumes disparity of at least $\delta > 0$ between groups $G = M$ and $G = F$ with all else known, $\mathsf{E}^*[Y|X{=}x, G{=}M, \bar{\mu}(x)] - \mathsf{E}^*[Y|X{=}x, G{=}F, \bar{\mu}(x)] \geq \delta$ for $\bar{\mu}(x) = \frac{n(x,M)\,\mathsf{E}[Y|X=x,G=M] + n(x,F)\,\mathsf{E}[Y|X=x,G=F]}{n(x,M)+n(x,F)}$.

Theorem (Trade-off reversal). Assume that the decision-maker has δ-disparate beliefs, that the regularity conditions hold, and that $0 < \Delta_\mu(x) < \delta$. Then π-almost surely for every $\eta > 0$ and $\zeta \in \left(0, \frac{1}{2}\right]$ there exists some M such that with probability (over draws of the training data) at least $1-\eta$ we have that $\Delta_{\hat{D}_+}(x) < \Delta_{\hat{D}_-}(x)$ and $\mathsf{E}[\ell(Y, \hat{D}_+)|X{=}x] < \mathsf{E}[\ell(Y, \hat{D}_-)|X{=}x]$ while $\Delta_{\hat{Y}_+}(x) > \Delta_{\hat{Y}_-}(x)$ and $\mathsf{E}[\ell(Y, \hat{Y}_+)|X{=}x] < \mathsf{E}[\ell(Y, \hat{Y}_-)|X{=}x]$ whenever $\zeta \leq \frac{n(x,F)}{n(x,F)+n(x,M)}, \frac{n(x,M)}{n(x,F)+n(x,M)} \leq 1 - \zeta$ and $n(x, F) + n(x, M) \geq M$.

Main result

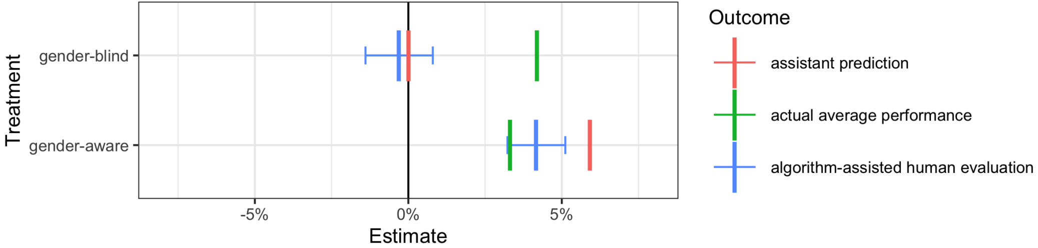
By group · Main results · Summary



Weighted for test-taker population distribution by age bracket, education, gender; standard error estimates clustered at subject level

On average, women evaluated lower than men, but what is the mechanism?

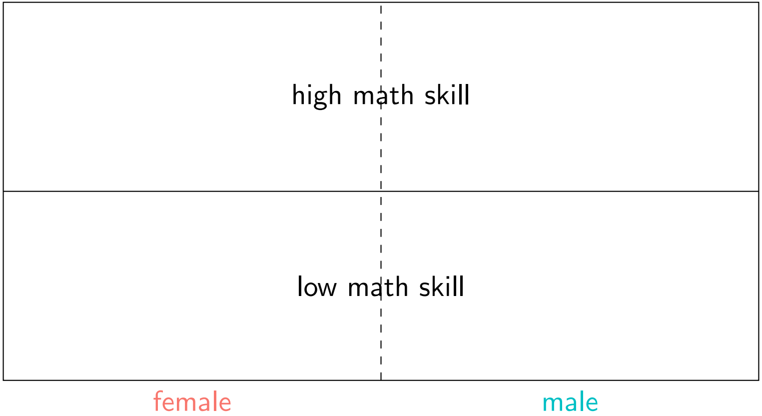
$$\Delta(h, X) = \mathrm{E}[\hat{D}|G{=}M, X] - \mathrm{E}[\hat{D}|G{=}F, X]$$

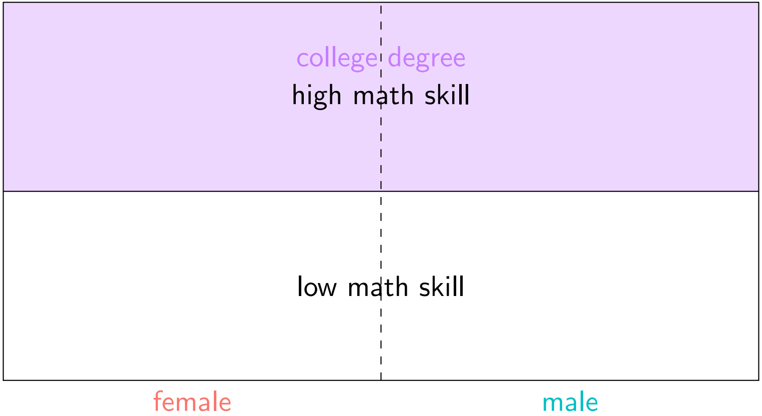


Gender-balanced test-takers; standard error estimates clustered at subject level

Stylized mechanism

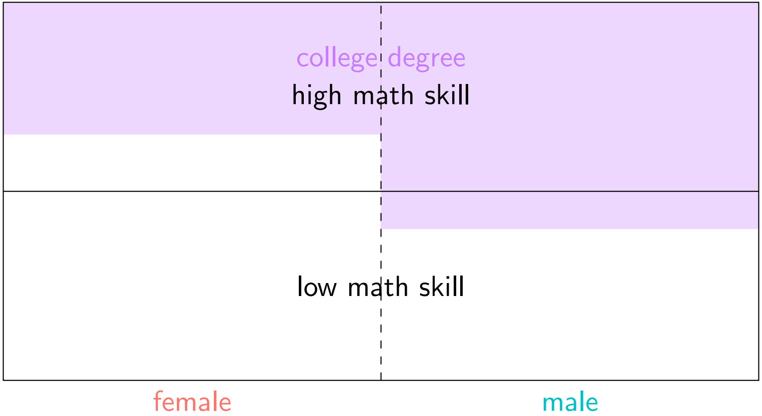
Main results Summary




college degree
high math skill
low math skill
female
male

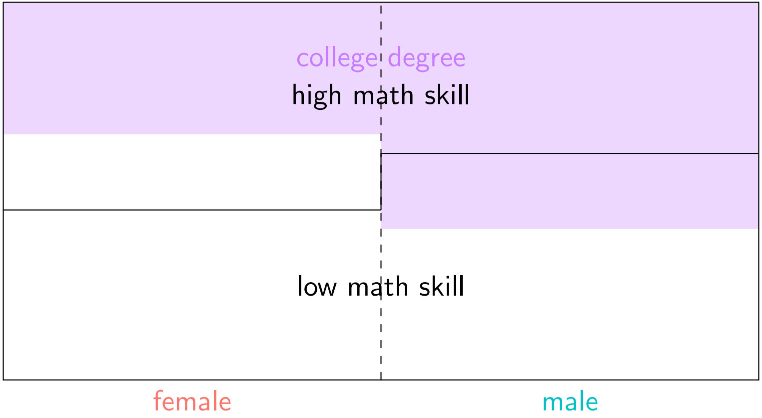
Stylized mechanism

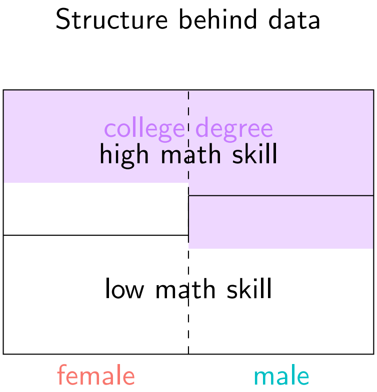
▸ Empirical evidence ◂ Main results

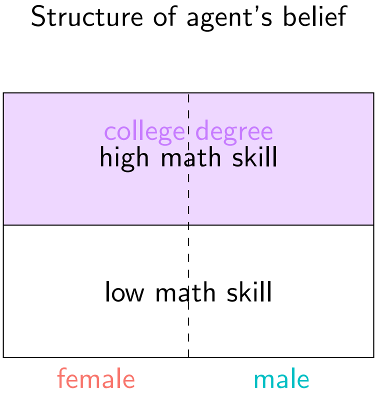


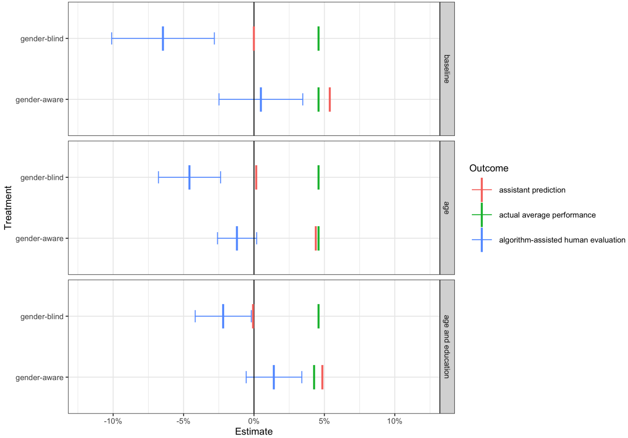
Stylized mechanism

▸ Empirical evidence ◂ Main results




Structure behind data
college degree
high math skill
low math skill
female
male

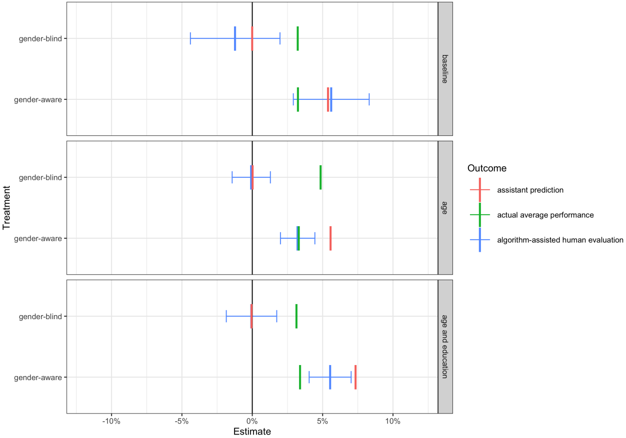

Structure of agent's belief
college degree
high math skill
low math skill
female
male



Weighted for test-taker population distribution by age bracket, education, gender; standard error estimates clustered at subject level

Gender-balanced results

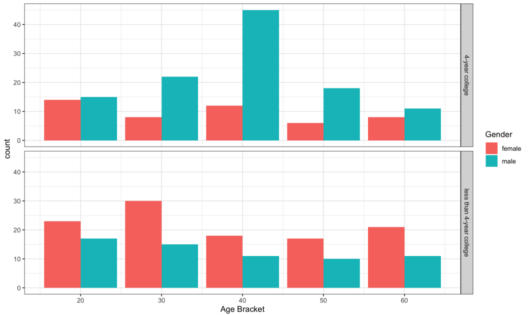
◂ Balanced results ◂ Main results



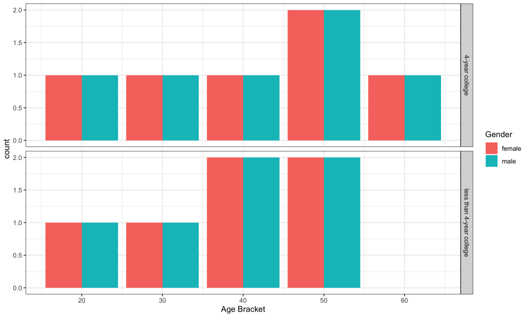
Gender-balanced test-takers; standard error estimates clustered at subject level

Covariate distribution

◂ Balanced results ◂ Main results

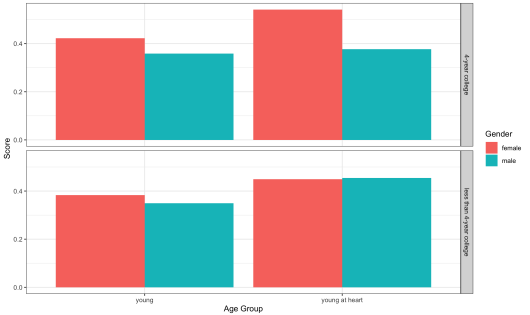


Empirical distribution of test-takers

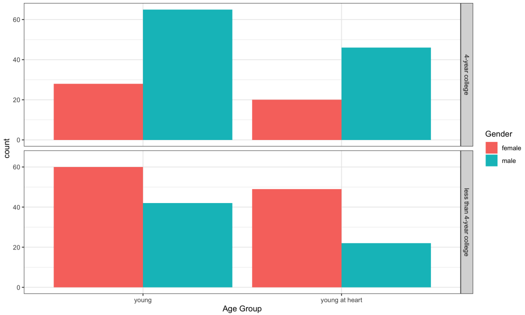


Balanced distribution of held-out test-takers

Performance and distribution of test-takers



Performance



Distribution

- **Correlated features:** G may be prohibited; consider inclusion of Z correlated with G

$$\text{data } (Y_i, X_i, Z_i, G_i) \quad \xrightarrow{\text{machine}} \quad \text{prediction } \hat{Y} = \hat{f}(X, G) \quad \xrightarrow{\text{human}} \quad \text{decision } \hat{D} = \mathrm{E}^*[Y|\hat{Y}, X, Z, G, H]$$

- **Beyond input restrictions:**
  - In many cases, input restrictions ineffective, suboptimal, or even counterproductive in first place (Kleinberg et al., 2018; Gillis and Spiess, 2018)
  - Kleinberg et al. (2018): Use protected characteristic, adjust across groups, e.g.

$$\hat{f}(x, g) = \widehat{\mathrm{E}}[Y|X{=}x, G{=}g] + \hat{\alpha}(g)$$

  - However, data still uninformative about group differences $\rightarrow$ prior disparity prevails

- **Optimal design:** Principal–agent problem where principal chooses $\hat{f}$ to minimize

$$\mathrm{E}[\ell(Y, \hat{D})] + \lambda(\mathrm{E}[\hat{D}|G{=}M] - \mathrm{E}[\hat{D}|G{=}F])^2 \qquad \hat{D} = \mathrm{E}^*[Y|X, G, H, \hat{f}(X, G)]$$

Andrews, I. and Shapiro, J. M. (2020). A Model of Scientific Communication. page 41.

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020). The Allocation of Decision Authority to Human and Artificial Intelligence. *AEA Papers and Proceedings*, 110:80–84.

Bastani, H., Bastani, O., and Sinchaisri, W. P. (2021). Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454*.

Bent, J. R. (2019). Is algorithmic affirmative action legal. *Geo. LJ*, 108:803.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2019a). Inaccurate Statistical Discrimination. *SSRN Electronic Journal*.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019b). The Dynamics of Discrimination: Theory and Evidence. *American Economic Review*, 109(10):3395–3436.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about Gender. *American Economic Review*, 109(3):739–773.

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv:1610.07524 [cs, stat]*. arXiv: 1610.07524.

Coffman, K. B., Exley, C. L., and Niederle, M. (2021). The Role of Beliefs in Driving Gender Discrimination. *Management Science*, 67(6):3551–3569.

De-Arteaga, M., Fogliato, R., and Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Honolulu HI USA. ACM.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, 64(3):1155–1170.

Donahue, K., Chouldechova, A., and Kenthapadi, K. (2022). Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. *arXiv preprint arXiv:2202.08821*.

Enarsson, T., Enqvist, L., and Naarttijärvi, M. (2022). Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, 31(1):123–153.

Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., Parkinson, M., Horvitz, E., Inkpen, K., and Nushi, B. (2022). Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1362–1374.

References II

Gillis, T. B. (2022). The input fallacy. *Minnesota Law Review*, 2022.

Gillis, T. B. and Spiess, J. L. (2018). Big Data and Discrimination. *The University of Chicago Law Review*, page 29.

Green, B. and Chen, Y. (2019). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99, Atlanta GA USA. ACM.

Huq, A. Z. (2020). A right to a human decision. *Va. L. Rev.*, 106:611.

Ibrahim, R., Kim, S.-H., and Tong, J. (2021). Eliciting Human Judgment for Prediction Algorithms. *Management Science*, 67(4):2314–2325. Publisher: INFORMS.

Imai, K., Jiang, Z., Greiner, J., Halen, R., and Shin, S. (2020). Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment.

Jiang, H. and Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR. ISSN: 2640-3498.

Kamenica, E. and Gentzkow, M. (2011). Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615.

Kim, P. (2022). Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *California Law Review, Forthcoming*.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018). Algorithmic fairness. In *AEA papers and proceedings*, volume 108, pages 22–27.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. (2021). Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv:2112.11471*.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, Halifax NS Canada. ACM.

Liang, A., Lu, J., and Mu, X. (2021). Algorithmic design: Fairness versus accuracy.

Ludwig, J. and Mullainathan, S. (2021). Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System. *The journal of economic perspectives: a journal of the American Economic Association*, 35(4):71–96.

Snyder, C., Keppler, S., and Leider, S. (2022). Algorithm Reliance Under Pressure: The Effect of Customer Load on Service Workers. *SSRN 4066823*.

Spiess, J. (2018). Optimal Estimation when Researcher and Social Preferences are Misaligned.

Stevenson, M. and Doleac, J. L. (2019). Algorithmic Risk Assessment in the Hands of Humans. *IZA Discussion Paper No. 12853*.