CHICAGO BOOTH

The University of Chicago Booth School of Business

Sophie Wang & Berkeley J. Dietvorst

Neither *Biased Algorithms* nor *Biased Humans* are Desirable, But **Combining Them** may be Permissible

Key Findings

- People prefer biased humans to consult unbiased AI;
- People prefer unbiased humans to avoid consulting biased AI;
- People typically prefer biased humans to consult a biased AI
  ➡ don't always object to the use of biased AI

Background

- **Multiple Notions of Fairness**
  - *Group Fairness*: Aims for equal outcomes across distinct groups.
  - *Conditional Statistical Parity*: Seeks equal outcomes across groups when conditioned on specific background factors.
- **Conflicts & Challenges**
  It's challenging to achieve all fairness notions simultaneously due to their conflicting nature.[1]
- **Bias: An Inevitable Outcome**
  Given the multiple notions of fairness, virtually any decision-maker, whether human or algorithm, may be biased or unfair under some fairness criteria.

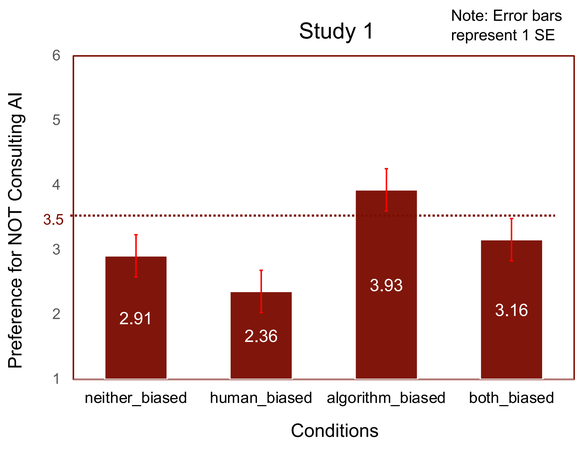
Methods

- Overview: Three pre-registered studies (N = 2,411) examining people's preferences for <u>potentially biased human decision-makers consulting possibly biased AI algorithms</u>.
- Participants: Studies 1 & 2: MTurk
  Study 3: CloudResearch
- Survey structure: In all studies, we used a 2x2 factorial design (Human/AI biased: Yes/No). Participants were randomly assigned to one of four conditions.

Studies 1 & 2

- **Study 1** (N = 806): examined people's preferences for doctors to consult (or not) AI in diagnostic decisions.
  Study 2 (N = 804): replicated Study 1, examining people's preferences for judges to consult (or not) AI in sentencing decisions.
- **Dependent Variable**: Preference (1-6 Scale)
  1 = Strong preference for AI consultation by doctor/judge
  6 = Strong preference against AI consultation by doctor/judge

Analysis 1 – average preference



Analysis 2 - OLS regression

Results

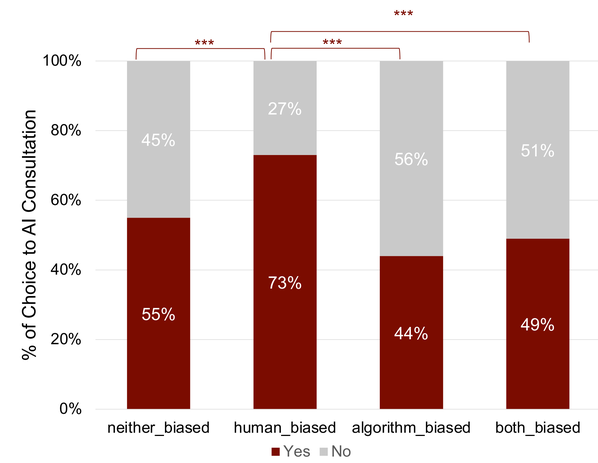
1. **Less** preference for AI assistance when *AI* framed as biased
   Algorithm_biased
   Study 1: t = 8.90, p < .001, d = .46
   Study 2: t = 5.84, p <. 001, d = .33

2. **More** preference for AI assistance when the *human* framed as biased
   Human_biased
   Study 1: t = -6.43, p < .001, d = -.33
   Study 2: t = -13.10, p <. 001, d = -.74

3. Independence of Effects
   Interaction
   Study 1: t = -1.01, p = .31
   Study 2: t = -.55, p = p = .582.

Study 3

- **Study 3** (N = 801): Incentivized Prize Allocation Study
- **Dependent Variable:** Binary choice for AI consultation in essay grading: "Yes" or "No"
- **Analyses:** Average preference and logistic regression
- **Results:**



Discussion & Directions

- **Public Perception:** Biased AI isn't always viewed negatively. Its pairing with potentially biased humans can be seen as beneficial.
- **Ongoing Research:** Replicating Study 1 using refined DV scales from organizational fairness literature.[2]
- **Future Studies:** Investigating factors that shape perceptions of fairness.

References

1. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining (pp. 797–806).

2. Colquitt, J. A. (2001). On the dimensionality of organizational justice: a construct validation of a measure. Journal of applied psychology, 86(3), 386.

Contact Sophie Wang at sophiewang@chicagobooth.edu