Public Attitudes on Performance for Algorithmic and Human Decision-Makers (Extended Abstract)

Kirk Bansak[1], **Elisabeth Paulson**[2]
[1]University of California, Berkeley
[2]Harvard Business School
kbansak@berkeley.edu, epaulson@hbs.edu

Concerns about the efficacy, bias, and fairness of artificial intelligence (AI) and algorithmic decision-making have intensified as algorithmic decision-making has increasingly diffused into public policy and other high-stakes decision-making contexts, such as criminal justice, healthcare, immigration, and other policy systems. There has been a limited amount of research focused on how the public (as *observers and voters*) views the use of AI and algorithmic decision-making tools in society. A larger body of research has focused on the views and behavior of human *users* of AI and algorithmic decision-making tools, and has found there to be aversion to (or appreciation for) algorithmic decision-making among users that is largely context-dependent. However, there are many open questions about attitudes toward different performance metrics and how they interact with decision-maker (DM) type. For example, it is well-known that simultaneously maximizing the fairness and accuracy of predictive algorithms is impossible, leading to unavoidable tradeoffs. However, what is not well-known is how the public prefers to trade off these objectives and whether such preferences differ when considering algorithmic versus human DMs. Also unclear is the degree to which public attitudes toward algorithmic DMs in society can be shaped through transparency about performance metrics.

This study examines the extent to which a DM's performance metrics (efficiency and fairness metrics) impact the public's evaluation and preference over human and algorithmic DMs in high-stakes contexts. Our goal is to answer the following questions: What are the public's expectations and preferences regarding DM type (human or algorithm)? To what degree do those preferences hinge upon the efficiency and fairness metrics of the DM? How does the public weigh tradeoffs between efficiency and fairness, and do these tradeoffs differ when the DM is an algorithm vs. a human?

To answer these questions, we designed, pre-registered, and implemented a high dimensional conjoint experiment on a large opt-in sample ($n \approx 9,000$) constructed to be representative of the U.S. population. In our conjoint experiment, respondents chose between pairs of DM profiles in two scenarios: pre-trial release decisions and bank loan decisions. DM profiles varied on the DM's type (human vs. algorithm) and on three metrics—defendant crime rate/loan default rate, false positive rate (FPR) among white defendants/applicants, and FPR among minority defendants/applicants—as well as an implicit (un)fairness metric defined by the absolute difference between the two FPRs.

Our results produced three sets of findings. First, they revealed an average preference for human DMs over algorithmic DMs, controlling for performance. However, this aggregate finding is driven by a sizeable proportion of respondents who believe human DMs are more likely to perform better in the real world. At the same time, there is a roughly equally sized proportion of respondents with the opposite belief that algorithmic DMs are superior in the real world and a corresponding (though slightly weaker) preference for algorithmic DMs. In other words, there is substantial heterogeneity in people's preferences for DM type (controlling for performance) that appears to be linked to heterogeneity in their *beliefs* about DM type.

Second, our results also revealed the importance of DM performance on preferences for specific DMs, and provided evidence on the degree to which people emphasize specific performance metrics. The relative sizes of the effects of the four metrics considered—crime/default rate, minority FPR, white FPR, and (un)fairness—were similar across both the pre-trial release and bank loan scenarios. The results also showed that while fairness considerations did factor into respondents' choices on average, they were not given a high priority.

Third, when both DM type and performance were considered together, our results revealed that people's preferences with respect to DM performance are *not* contingent upon DM type. That is, regardless of the DM type, the relative impacts of the performance metrics on DM choice remained stable, and any differences in the effects were limited and mostly small in magnitude. In other words, we do not find evidence that algorithmic DMs are held to a fundamentally different set of standards on fairness or efficiency.

Taken together, the results collectively suggest that people have very *different beliefs* about what type of DM (human or algorithm) will deliver better performance and should be preferred, but they have much more *similar desires* in terms of what they want that performance to be regardless of DM type.

The full version of our paper can be accessed at: https://osf.io/preprints/osf/pghmx