Algorithm-in-the-Loop Decision Making

Ben Green,[1] **Yiling Chen**[1]
[1]Harvard University
bgreen@g.harvard.edu, yiling@seas.harvard.edu

Abstract

We introduce a new framework for conceiving of and studying algorithms that are deployed to aid human decision making: "algorithm-in-the-loop" systems. The algorithm-in-the-loop framework centers human decision making, providing a more precise lens for studying the social impacts of algorithmic decision making aids. We report on two experiments that evaluate algorithm-in-the-loop decision making and find significant limits to these systems.

Introduction

Machine learning models are increasingly being incorporated into important decision making processes (such as criminal sentencing) under the assumption that they will improve decision making. These models are typically evaluated according to statistical metrics related to considerations such as accuracy and fairness.

Yet these evaluations fail to fully capture the impacts of algorithmic decision making aids. In practice, these tools do not make definitive judgments, but instead are typically used to inform human decision makers. It is therefore essential that considerations of algorithmic decision making aids be informed by rigorous studies of how people actually interpret and use them.

With this in mind, we introduce a new framework of "algorithm-in-the-loop" systems: processes that employ algorithmic aids to enhance human decision making. In contrast to the human-in-the-loop paradigm, which privileges algorithms as the central focus and uses people to improve algorithmic performance, the algorithm-in-the-loop perspective privileges people as the central focus and uses algorithms to improve human decision making.

The algorithm-in-the-loop framework can inform the design and evaluation of algorithmic decision making aids. In terms of design, it emphasizes developing systems for integration into sociotechnical contexts rather than for isolated decision making. In terms of evaluation, it emphasizes the human's decisions—rather than the algorithm's decisions—as the primary outcome of interest.

We report here on two sets of experiments studying how people make predictions when presented with the aid of a machine learning model. We ran experiments on Amazon Mechanical Turk, asking participants to make predictions in two settings: pretrial release and financial lending. Participants were presented with narrative profiles about people—either criminal defendants or loan applicants—and were asked to predict how likely those people were to take a certain action in the future (for defendants, failing to appear in court for trial or being arrested before trial; for applicants, defaulting on the loan). Some of the experimental participants were shown, in addition, the prediction of a machine learning-based risk assessment for that individual. These risk assessments were trained on historical data from the U.S. Department of Justice and a financial lending company. Participants were financially incentivized to report their true predictions. After making these predictions, participants were asked to report their beliefs about their own performance and the risk assessment's performance.

Disparate Interactions

Our first experiments studied algorithm-in-the-loop decision making in the context of pretrial release (Green and Chen 2019a). We particularly focused on how people respond to the risk assessment's predictions based on defendant race. This evaluation shifts the focus of algorithmic fairness from the model itself to the decisions that people make with the model. We found two types of evidence for what we call "disparate interactions": racially disparate impacts that emerge through people's biased interactions with the risk assessment.

We looked first at the influence of risk scores on people's behavior, based on the race of defendants. When the risk assessment suggested that people reduce their predictions of risk, the risk assessment exerted a similar influence on participants regardless of the defendant's race. Yet when the risk assessment suggested that people predict a higher level of risk than they otherwise would have, it exerted a 25.9% stronger average influence on predictions about black defendants than on predictions about white defendants. In other words, our experiment participants were 25.9% more strongly influenced by the risk assessment to

increase their risk predictions when evaluating black defendants than white ones, leading to a 20.3% larger average increase for black than white defendants due to the risk assessment.

We then looked at how people deviated from the risk assessment when making predictions about black and white defendants. When evaluating white defendants, participants made predictions that were marginally below the risk assessment's predictions. Yet when evaluating black defendants, participants predicted higher levels of risk than the risk assessment did. Participants were 36.4% more likely to deviate positively from the risk assessment and 21.5% less likely to deviate negatively from the risk assessment when evaluating black defendants.

Principles and Limits

Our second experiments considered both the normative principles regarding how people *should* collaborate with algorithms and the empirical evidence regarding how people *do* collaborate with algorithms (Green and Chen 2019b). First, we articulated three principles that are essential to ethical and responsible decision making with algorithms: accuracy, reliability, and fairness. We then ran experiments to test whether people follow these principles when making decisions. We evaluated decision making across both pretrial release and financial lending as well as across six conditions for presenting the risk assessment.

Although presenting the risk assessment did increase people's accuracy in almost every case, our study participants made decisions that were both unreliable and racially biased. Across both settings and all six treatments, participants were unable to consistently evaluate the quality of their own or the risk assessment's predictions. In turn, in all but one case, participants did not differentiate their reliance on the risk assessment based on how it actually performed. This means that people are not properly adjusting their decision making strategy to account for the particular details of each case. We also found the presence of disparate interactions across all settings and treatments.

Conclusion

These results call into question foundational assumptions about the efficacy and reliability of algorithmic decision making aids. Before such algorithms are integrated into high-stakes decisions, we must be confident that the decision making processes that result will be ethical and responsible. It is therefore necessary both to further develop criteria that should govern algorithm-in-the-loop decision making and to develop a deeper science of human-algorithm interactions for decision making.

The framework of algorithm-in-the-loop decision making brings these questions to the fore, expanding beyond evaluating algorithms in the abstract to investigating the full sociotechnical contexts in which people and algorithms interact. This approach provides a necessary corrective to assessments of algorithmic decision making aids that are overly sanguine because they fail to consider the sociotechnical context. It can also inform the development of algorithmic systems that more rigorously promote responsible and ethical decision making.

A key concern raised by these experiments is that the adoption of machine learning decision making aids will create an empty cycle of oversight. Algorithms are implemented to improve human decision making, then those same humans are asked to oversee the algorithm's decisions. Yet our experiments suggest that people are unable to provide the types of oversight that are required. Greater attention is therefore needed to more rigorously synthesizing human and algorithmic decision making.

A key aspect of future work will be to study algorithm-in-the-loop decision making in real-world rather than experimental contexts. Mechanical Turk experiments are no substitute for *in situ* evaluations. However, experiments such as these provide an effective tool for diagnosing the types of human-algorithm interactions that could arise in practice. Issues identified in experiments can inform the design and evaluation of real-world systems in order to prevent breakdowns when the stakes are high.

Acknowledgments

We thank Juntao Wang for help with setting up and running experiments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745303. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Green, B., and Chen, Y. 2019a. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 90–99. New York, NY, USA: ACM.

Green, B., and Chen, Y. 2019b. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):50:1–50:24.