Rethinking Fairness for Human-AI Collaboration

Haosen Ge
Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Hamsa Bastani
Department of Operations, Information and Decisions, Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Osbert Bastani
Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Most work on algorithmic fairness focuses on whether the algorithm makes fair decisions *in isolation*. Yet, these algorithms are rarely used in high-stakes settings without human oversight, since there are still considerable legal and regulatory challenges to full automation. Moreover, many believe that human-AI collaboration is superior to full automation because human experts may have auxiliary information that can help correct the mistakes of algorithms, producing better decisions than the human or algorithm alone. However, human-AI collaboration introduces new complexities – the overall outcomes now depend not only on the algorithmic recommendations, but also on the subset of individuals for whom the human decision-maker complies with the algorithmic recommendation. Recent studies have shown that selective compliance with algorithms can *amplify* discrimination relative to the prior human policy, even if the algorithmic policy is fair in the traditional sense. As a consequence, ensuring equitable outcomes requires fundamentally different algorithmic design principles that ensure robustness to the decision-maker's (a priori unknown) compliance pattern.

To resolve this state of affairs, we introduce the notion of *compliance-robust* algorithms – i.e., algorithmic decision policies that are guaranteed to (weakly) improve fairness in final outcomes, regardless of the human's (unknown) compliance pattern with algorithmic recommendations. In particular, given a human decision-maker and her policy (without access to AI assistance), we characterize the class of algorithmic recommendations that never result in collaborative final outcomes that are less fair than the pre-existing human policy, even if the decision-maker's compliance pattern is adversarial. Next, we prove that there exists considerable tension between traditional algorithmic fairness and compliance-robust fairness. Unless the true data-generating process is itself perfectly fair, it can be infeasible to design an algorithmic policy that simultaneously satisfies traditional algorithmic fairness, is compliance-robustly fair, and is more accurate than the human-only policy; this raises the question of whether traditional fairness is even a desirable constraint to enforce for human-AI collaboration. If the goal is to improve fairness and accuracy in human-AI collaborative outcomes, it may be preferable to design an algorithmic policy that is accurate and compliance-robustly fair, but not traditionally fair. Our last result shows that the tension between traditional fairness and compliance-robust fairness is prevalent. Specifically, we prove that for a broad class of fairness definitions, fair policies are not necessarily compliance-robustly fair, implying that compliance-robust fairness imposes fundamentally different constraints compared to traditional fairness.

2012 ACM Subject Classification Computing methodologies → Machine learning

Keywords and phrases fairness, human-AI collaboration, selective compliance

Digital Object Identifier 10.4230/LIPIcs.ITCS.2024.52

Category Extended Abstract

Related Version *Full Version*: https://arxiv.org/abs/2310.03647 [1]

Funding *Hamsa Bastani*: Analytics at Wharton

References

1 Haosen Ge, Hamsa Bastani, and Osbert Bastani. Rethinking fairness for human-ai collaboration. *arXiv preprint*, 2023. arXiv:2310.03647.

15th Innovations in Theoretical Computer Science Conference (ITCS 2024).
Editor: Venkatesan Guruswami; Article No. 52; pp. 52:1–52:1
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany