Crowdworkers Are Not Judges: Rethinking Crowdsourced Vignette Studies as a Risk Assessment Evaluation Technique

Emma Lurie
University of California, Berkeley
emma_lurie@berkeley.edu

Deirdre K. Mulligan
University of California, Berkeley
dmulligan@berkeley.edu

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract

Algorithmic risk assessments are widely deployed as judicial decision-support tools in the U.S. criminal justice system. A review of recent CS/HCI/CSCW research around algorithmic risk assessments reveals a potentially troubling trend: the use of crowdworkers as a stand-in for judges when analyzing the impact of algorithmic risk assessments. We raise three concerns about this approach to understanding algorithms in practice, and call for a reevaluation of whether human-centered AI research should rely on experimental crowdworker studies as a means to assess the impact of algorithmic risk assessments in the criminal justice system.

Introduction

Understanding how algorithmic predictions affect human decision making is an important and pressing area of research. Since the public controversy around COMPAS in 2016 [2], algorithmic risk prediction in criminal justice has become a common case study for computer scientists exploring human-AI decision making. A growing number of CS/CSCW/HCI papers [13, 6, 7, 8] attempt to gain insight into how risk assessment algorithms change various outcomes in the criminal justice system (e.g. pre-trial release, recidivism rates). These studies often conduct experiments that present crowdworkers, usually Amazon Mechanical Turk workers, with 1) a vignette (a few sentences of back-

ground about a criminal case), and 2) a risk assessment score. With these two pieces of information, crowdworkers are then asked to make a judgment about the best course of action in the case.These crowdworker vignette studies require the assumption (to various extents) that crowdworkers in the experiment can provide insight into the decisions of judges in practice.

We argue that relying on crowdworkers to assess the impact of algorithmic risk assessments in the criminal justice system warrants reevaluation by the responsible AI community because of 1) ecological validity limitations; 2) framing difficulties; and, 3) ethical concerns.

Limited Ecological Validity

The proposed goal of these crowdworker studies is to understand human-algorithm interactions within a specific context. While the studies often take care to represent the technical aspects of the algorithmic system with some fidelity to those deployed in practice, they abstract away the education, training, professional identity, and organizational context of the individuals using the algorithmic system, replacing judges with undifferentiated crowdworkers. This creates an ecological validity problem: what crowdworkers do in these experiments is almost certainly not what judges do in practice.[1]

Previous research casts doubt on the notion that vignette studies can predict judges' actions on the bench. When judges are asked to evaluate short case histories (vignettes), they are swayed by factors such as a prior criminal record; however, when the researchers analyzed the courtroom behavior of the same judges, they find that the judges almost exclusively followed the bail recommendation of the district attorney [4]. This raises questions concerning how much we can learn about judicial decision making or criminal justice system outcomes from vignette experiments.

Additional research finds that professionals are less likely to heed advice generally, and less deferential to machine advice specifically [9]. While it has been well established that judges are susceptible to cognitive biases, the literature demonstrates that judges decision making patterns differ from "ordinary humans" in distinct ways that are connected to professional training and normative commitments in criminal justice [10].

Detailed ethnographic research on actual judges' interactions with risk assessment tools reveals that their appropriation of such systems is informed by routines, norms, obligations of professional identity, and their position relative to others within the organizational hierarchy [3]. Christin's work finds judges resisting risk-assessment tools for reasons including concerns that they do not capture professional judgement as well as wariness about the opaque and commercial nature of the systems. Other work by Stevenson and Doleac [12] finds that while judges sometimes use risk scores, judges diverge from the risk assessment algorithms for normative reasons not modeled in the risk scores. For example, judges are consistently lenient when sentencing defendants 23 years and younger despite risk assessments viewing age as a strong predictor of future rearrest. These concerns reflect professional training and context specific normative commitments that are unlikely to be present in the crowdworker community and therefore limit the insights crowdworker vignette studies provide into actual practice.

Given that non-experts on Mechanical Turk are being asked to make a high-stakes decision without understanding the

[1] While some of the existing crowdworker vignette studies acknowledge this limitation, such acknowledgements do not fully address the underlying validity issues.

reasoning requirements and normative commitments that constrain and guide judges in practice, it is unclear whether the results provide useful new information about the effects of risk assessments on judicial decision making or criminal justice system outcomes.

The Framing Trap

We believe the crowdworker vignette studies are an example of the "framing trap." The framing trap is described by Selbst et al. [11] as a way that technical interventions "[fail] to model the entire system over which a social criterion, such as fairness, will be enforced." In vignette crowdworker experiments, both the casting of the risk assessment score as a primary criteria for judicial decision making as well as the minimization of the importance of professional expertise result in a failure to model the sociotechnical frame of judicial decision making in the criminal justice system.

As we consider how to better measure the effects of algorithmic risk assessments, it is worth considering the framing of the research problem adopted by crowdworker vignette studies. The crowdworker vignette studies focus on assessing whether judges become more or less accurate at predicting future outcomes, but judges may be considering other factors. As in the age example referenced above, there may be times where judicial divergence from risk scores is a desirable societal outcome. Judges may be prioritizing justice over some technical measure of accuracy.

There are other fundamental critiques about risk assessments including evidence risk assessments data is 1) biased by the over-policing and unequal outcomes at every level in the criminal justice system for people of color and 2) conflates individual's danger to society with failure to appear in court [1]. Perhaps assessing humans' ability to predict risk with the assistance of an algorithmic tool, misses broader, more pressing issues with risk assessments in practice [5].

Ethical Concerns

Judges' decisions have serious consequences for people's lives, and disproportionately the lives of people of color. Research that swaps out judges for crowdworkers makes research faster and cheaper, but risks fueling faulty assumptions about the impact of risk assessment tools. We recognize the difficulty and cost of studying judges interactions with risk assessment tools in the lab and in the wild. Yet, given the research showing divergence between lay people and experts, and divergence between judges in the lab and in the wild, the notion that crowdworker studies can tell us something meaningful about the effects of risk assessment tools on the criminal justice system seems problematic. Similar to clinical trials of drugs recruiting exclusively male subjects because testing on women was more complicated due to pregnancy, the expedient use of crowdworkers may produce quicker results but creates risks if policymakers rely on the research to inform public policy or sociotechnical system design.

In the medical domain, researchers often must both compensate doctors to observe them in the lab or interview them, and gain their trust. Building the trust necessary to gain access requires researchers to establish that their research will produce meaningful information that is relevant to the professional community. While a barrier to research, this serves as a check on validity and beneficence.

As a beneficial AI research community, when we use crowdworkers to explore judicial decision making, we are concerned that we are broadcasting to the wider community that the professionals making the decisions, and the ac-

cused, whose futures are at stake, may not be worth the expense of more research that more rigorously captures the complexities of practice.

Conclusion

It is important to study how algorithmic risk assessments are shaping judicial decision making, but as these tools profoundly impact people's lives, the methods and experimental designs we employ should mirror the seriousness of the use case. We believe as a responsible and beneficial AI community that we should reevaluate whether crowdworker vignette experiments are a sound way to advance our understanding of how algorithmic risk assessments are impacting the criminal justice system.

REFERENCES

[1] 2019. Technical flaws of pretrial risk assessments raise grave concerns. `https://dam-prod.media.mit.edu/x/2019/07/16/TechnicalFlawsOfPretrial_ML%20site.pdf`. (2019).

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016).

[3] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017).

[4] Ebbe B Ebbesen and Vladimir J Konecni. 1975. Decision making and information integration in the courts: The setting of bail. *Journal of Personality and Social Psychology* 32, 5 (1975), 805.

[5] Ben Green. 2020. The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. In *Proceedings of FAT**.

[6] Ben Green and Yiling Chen. 2019a. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of FAT**. 90–99.

[7] Ben Green and Yiling Chen. 2019b. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[8] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[9] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[10] Jeffrey J Rachlinski and Andrew J Wistrich. 2017. Judging the judiciary by the numbers: Empirical research on judges. *Annual Review of Law and Social Science* 13 (2017), 203–229.

[11] Andrew D Selbst, danah boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of FAT**. 59–68.

[12] Megan T Stevenson and Jennifer L Doleac. 2019. Algorithmic Risk Assessment in the Hands of Humans. *Available at SSRN* (2019).

[13] Michelle Vaccaro and Jim Waldo. 2019. The effects of mixing machine learning and human judgment. *Commun. ACM* 62, 11 (2019), 104–110.