UT LAW CLE | THE UNIVERSITY OF TEXAS SCHOOL OF LAW

PRESENTED AT

34th Annual Technology Law Conference

May 26-28, 2021
Live Webcast

A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores

Maria De-Arteaga
Riccardo Fogliato
Alexandra Chouldechova

A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores

Maria De-Arteaga*
Heinz College
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA
mdeartea@andrew.cmu.edu

Riccardo Fogliato*
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA, USA
rfogliat@andrew.cmu.edu

Alexandra Chouldechova
Heinz College
Carnegie Mellon University
Pittsburgh, PA, USA
achould@cmu.edu

ABSTRACT

The increased use of algorithmic predictions in sensitive domains has been accompanied by both enthusiasm and concern. To understand the opportunities and risks of these technologies, it is key to study how experts alter their decisions when using such tools. In this paper, we study the adoption of an algorithmic tool used to assist child maltreatment hotline screening decisions. We focus on the question: Are humans capable of identifying cases in which the machine is wrong, and of overriding those recommendations? We first show that humans do alter their behavior when the tool is deployed. Then, we show that humans are less likely to adhere to the machine's recommendation when the score displayed is an incorrect estimate of risk, even when overriding the recommendation requires supervisory approval. These results highlight the risks of full automation and the importance of designing decision pipelines that provide humans with autonomy.

Author Keywords

Human-in-the-loop; Decision support; Algorithm aversion; Automation bias; Algorithm assisted decision making; Child welfare

CCS Concepts

•Human-centered computing → Human computer interaction (HCI); User studies; **•Information systems → Decision support systems; •Applied computing → Computing in government;**

INTRODUCTION

Risk assessment tools are increasingly being incorporated into expert decision-making pipelines across domains such as criminal justice, education, health, and public services [8, 26, 23, 7, 45]. These tools, which range from simple regression models to more complex machine learning models, distill available information on a given case into a risk score reflecting the likelihood of one or more adverse outcomes. Bolstered by decades of research showing that statistical models outperform human experts on prediction tasks, there is widespread optimism that these tools will increase the quality of decisions [32, 11, 20, 1, 25]. This optimism is tempered by evidence that, while humans provided with machine predictions may achieve improved performance, not only do they continue to underperform the machine predictions, but they may also uptake the information in ways that leads to increased disparities in decision outcomes across racial [19] and socioeconomic groups [42]. Such findings raise critical questions about the role of humans in the loop and human-machine complementarity in key societal domains. In this work we focus on one important question in this area: Are humans capable of identifying certain cases where the machine's recommendation is wrong, and appropriately overriding the recommendation in such cases?

We analyze a real world child welfare decision making context where call workers are tasked with deciding whether a call concerning potential child neglect or maltreatment should be screened in for investigation. While in many instances the information communicated in the call may be enough for the call worker to make a determination, in other instances the information may be vague and inconclusive. In an effort to better focus resources on investigating cases where the children are at greatest risk, Allegheny County has deployed a risk assessment tool called the Allegheny Family Screening Tool (AFST) to assist call workers in their screening decisions. The tool uses multi-system administrative data to assess the likelihood that children on the case will experience adverse child welfare events in the near future. More information about the tool and its development can be found on the County's AFST website [38].

Some time after the tool was deployed it was discovered that a technical glitch had resulted in a subset of model inputs being incorrectly calculated in realtime. This in turn resulted in misestimated risk scores being shown in some cases. While, as we elaborate on below, the misestimation was often mild, and the shown score generally provided reasonable risk information, the glitch permits us a rare opportunity to investigate

* Indicates equal contribution.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04.
http://dx.doi.org/10.1145/3313831.3376638

real world decision making in the presence of misestimated risk.

Before proceeding, we pause to make an important point. These types of technical issues are not uncommon. What is uncommon is for organizations to choose to be transparent about their occurrence. We recognize Allegheny County for their transparency and hope that this approach will become the norm in the deployment of algorithmic systems in sensitive societal domains.

The remainder of this paper is organized as follows. We begin with a discussion of related work in which we provide an overview of phenomena such as algorithm aversion and automation bias that come to bear on human decision making in the presence of algorithmic decision support tools. We then describe the child welfare decision making context, the risk assessment tool deployment setup, and our available data. Our analysis of the data begins by demonstrating that there was a marked change in workers' screening decisions in the post-deployment period. Having established that an overall change in behavior did occur, we then investigate the extent to which call workers deviate from recommendations based on a misestimated risk score. We show that workers are able to appropriately override the tool in many such cases. We also probe questions of potential disparities in adherence to recommendations across racial and socio-economic groups, finding that the deployment of the tool neither significantly mitigates nor exacerbates disparities observed at the given level of analysis. Lastly, we conclude with a discussion of human and system factors that we believe may have contributed to the observed results, and outline opportunities for further research to better understand relevant factors.

BACKGROUND AND RELATED WORK

Prior research has attempted to answer *whether* and *how* the deployment of algorithmic risk assessment tools affects users' decisions. While many have advocated for the adoption of these tools on the basis of their superior predictive accuracy, findings are mixed on whether integrating prediction tools into decision making significantly improves decision quality. Indeed, research in the field suggests that the outcomes of decisions taken by a human aided by a decision support system are often no better than those taken by the human alone.

Recent work has paid special attention to the introduction of risk assessment in the context of pretrial decision making in the criminal justice system. Although the integration of the risk assessment tools was, ex ante, expected to lead to a sharp and persistent decrease in incarceration rates, recent findings suggest that there is no impact at all [15] or find there is a decrease but of much smaller magnitude than initially hoped [44]. There is consensus that these lackluster results are due at least in large part to the wide heterogeneity in judges' compliance with the tools' recommendations [9]. Notably, differential compliance has been shown to be a factor driving increased poor-rich [42] and black-white [46, 2] disparities in the post-deployment period. For instance, [2] found that the increased racial gap in incarceration rates post-deployment was due both to inter-variation—judges in whiter counties showing higher compliance—and by intra-variations—overrides of low and moderate risk being more frequent for black than for white defendants.

More broadly, there are two competing tendencies that have been observed in studies of human compliance with algorithmic recommendations: *algorithm aversion* and *automation bias*. *Algorithm aversion*–the tendency to ignore tool recommendations after seeing that they can be erroneous–originates from a lack of agency [29, 12] and lack of transparency of the algorithm [49]. Studies have shown that users will knowingly sacrifice accuracy in favor of gaining some control over the algorithm's output [14]. Similarly, [18] reports an experiment in which humans override the machine's predictions when these are highly reliable. Users' reliance on the system is known to vary with the observed [51, 52] and stated accuracy [50] of the system. However, even if the recommendations of more accurate systems are followed more often, agents affected by algorithm aversion may nevertheless prefer human judgment over algorithmic predictions even when evidence known to both the designer and the user clearly indicates that the algorithmic predictions are more accurate than human assessment [13].

Users affected by *automation bias*, on the other hand, will follow tool recommendations despite available (but unnoticed or unconsidered) information that would indicate that the recommendation is wrong. *Automation bias* consists of two classes of errors. Omission errors refer to instances where humans fails to detect problematic cases (or fail to act) because they were not flagged as such by the system. A prominent example is that of pilots in high-tech cockpits, who are prone to relying blindly on automated cues as a heuristic replacement for vigilant information seeking [35]. Commission errors refer to instances where humans take action on the basis of an erroneous algorithmic recommendation, failing to incorporate contradictory external information into the decision process. In the clinical decision support context, commission errors may result in patients being subjected to unnecessary, potentially invasive testing or treatment.

Studies analyzing factors contributing to automation bias have found that complex tasks and time pressure may increase over-reliance on decision support [41, 17]. The users' experience level and their confidence in their own decisions have also been found to be causes of automation bias [31, 33]. Social accountability has been found to reduce automation bias [43], an important result when considering decision support systems used by experts with high public visibility or who are publicly elected, such as judges. Meanwhile, studies focused on the causes of algorithm aversion have found that repeatedly seeing the algorithm make the same mistake leads to decreased reliance of the agent on the system [13], while giving some control over the algorithm can counter this phenomenon [14].

Automation bias and algorithm aversion are opposing phenomena. While automation bias degrades decision quality by driving over-compliance with algorithmic recommendations, algorithm aversion does so by driving under-compliance. There are two characteristics of the decision context that are indicative of which form of bias is likely to dominate: the type of task, and the level of automation. A significant portion of

Title search: A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores

First appeared as part of the conference materials for the
34th Annual Technology Law Conference session
"Artificial Intelligence and Machine Learning: Legal and Ethical Considerations"