ALGORITHMIC FAIRNESS AND OPACITY WORKING GROUP (AFOG)

AFOG workshop panel 1: What a technical ‘fix’ for fairness can and can’t accomplish

By Daniel Kluttz, Joshua A. Kroll, Jenna Burrell, Deirdre Mulligan

Corresponding author: dkluttz@berkeley.edu

Published August 13, 2018

Berkeley SCHOOL OF INFORMATION

Google

INTRODUCTION

As the use of algorithmically based decision-making systems has grown across all sectors of social and economic life, so too have concerns over whether the outputs given by such systems are unfair to certain individuals or groups. What are the implications of researchers and engineers developing technical solutions to problems of "fairness" in algorithmic systems, particularly those that incorporate machine learning? The opening panel of the 2018 AFOG Summer Workshop wrestled with the question of what it means to define, address, and measure the abstract value of fairness in a sufficiently technical way that it can be implemented in a software system. It also moved beyond proposed technical solutions to discuss real-world examples of situations that evoked fairness concerns and debate whether, how, and under what circumstances technical 'fixes' fall short and non-technical interventions should apply.

NO SINGLE DEFINITION. NO SINGLE TECHNICAL "FIX."

At the outset of our panel, all panelists agreed that automated decision-making (e.g., classification, risk-assessment) tools should aim to produce "fair" outcomes, or at least minimize "unfair" outcomes as much as possible. However, the panelists rejected the idea that accounting for such an abstract, complex, and contested value of fairness can be accomplished with any single mathematically based, technical solution. More fundamentally, not only is there no single definition of fairness that we can rely on and implement, a narrow focus on what is technically tractable regarding 'algorithmic fairness' risks distracting us from questions about the fairness, ethics, or justice of the broader systems within which such algorithms are embedded. For example, we might ask how algorithmic applications in criminal justice support or undermine broader reform movements. Or we might consider how improvements in facial-recognition accuracy of minority group members might smooth the path for privacy-violating surveillance applications. This is a theme we revisit throughout this piece.

From a technical standpoint, an algorithmic system can exhibit bias--in the sense that it systematically discriminates against particular individuals or groups in favor of others in some socially undesirable way--because of *bias present at any point along the technical pipeline*. In other words, bias could be present in the data inputs (i.e., the raw, observed data), in the logic or process used to map raw data to constructs of interest (i.e., the (perhaps unobserved) variables that are relevant for the model), and/or in the outputs (i.e., the analysis of the data, which yields a decision or prediction). Friedler et al. 2016 refer to these as the observed space, the construct space, and the decision space, respectively. Friedler et al.'s formalization introduces an important element of skepticism about human behavioral and demographic data to research on fairness in machine learning (e.g., that an IQ score is an unbiased measure of intelligence, that arrests and convictions accurately measure criminal activity). But it is even more complicated than that—researchers and software engineers should be aware of and distinguish between fairness concerns emanating from different sources within each of the three spaces. For example, just within the observed space, there could be a variety of culprits, including selection, sampling, and reporting bias. Particular definitions and assumptions of "fairness," much less the technical strategies and metrics with which to address them, may be more or less applicable depending on where in the pipeline one is focusing one's attention.

DESIGN AND METHODOLOGICAL ISSUES

One point stressed by panelists is that data scientists and engineers often do not exhibit an adequate *appreciation for how fairness considerations of their model outputs may be affected by matters of system design and research methodologies* (e.g., data documentation, distinguishing and addressing different types of bias, reliability, validity). This is especially the case in the corporate world and particularly among those who develop machine-learning algorithms. Measurement issues, which come up in the "construct space" described above, are particularly problematic for machine-learning practitioners. The panelists observed that social scientists, compared to machine-learning practitioners, generally are more adept at thinking critically about their constructs (operationalizations of theory-driven concepts into observed variable) and evaluating construct validity (how well a measure captures the concept it is intended to capture). Panelists felt that machine-learning practitioners rarely consider such questions, noting that they tend to focus more on model evaluation, such as comparing their models' performance on training data vs. test data, error metrics (as with the confusion matrix), and cross-validation.

ALGORITHMS EMBEDDED IN BROADER CONTEXT

None of the above includes sources of unfairness or bias coming from *outside* of an algorithm's immediate pipeline. Panelists agreed that we should think of *fairness as a contextual property of broader socio-technical systems* and not simply as an instrumental property of the technological components. This focus on context and systems was a major topic revisited throughout the discussion. Algorithms and their outputs are embedded within multiple technical and societal systems, all of which can have an impact on fairness. First, from a technical perspective, and in practice, not only is there the immediate pipeline we discussed above (itself a system), algorithmic systems often interact with other software and technical systems. Even within a single organization, an algorithm could be "fair" when conceived of or deployed by itself but could raise concerns when used in the context of company's broader technical infrastructure. Second, from a societal perspective, the long shadows of discrimination, bias, and inequality, loom over every aspect of algorithmic decisions. Racial stereotypes, for example, can be encoded in the raw data that train machine-learning models.

NON-TECHNICAL INTERVENTIONS (IF NOT SOLUTIONS)

Our panel discussion of fairness concepts and the limitations of technical solutions led us to turn to *non-technical* interventions that can address fairness. Panelists noted that the drive to formalize and define fairness in ways that can be implemented technically tends to focus our attention in particular ways and leads us to tackle fairness with "fixes" of certain forms and not others. *Technology, on its own, is not good at explaining when it should not be used or when it has reached its limits*; sometimes, the right answer is not to aim at "fixing" the technology, but to return to the drawing board and reconsider the structure of the broader social system into which the technology fits.

Indeed, as one panelist argued, we can't simply introduce a "fair" algorithm into an unfair system and expect a fix. To illustrate the point, the panelist used the example of pretrial detention algorithms, which aim to quantify the risk of an arrestee failing to appear for court or re-offending and are increasingly used by judges when setting bail terms. The panelist argued that what is "unfair" has much more to do with a cash bail system's unjust effects on poor and racial-minority arrestees than with the accuracy of any software's risk-assessment score (on risk assessment tools and pre-trial detention generally, see Koepke and Robinson 2018). Bail reform, the panelist argued, is better served by policy interventions, not technical fixes. *The government, academic, and technology sectors should engage with one another more frequently and educate one another on their respective areas of expertise*. Doing so would allow each to have better understandings of when and how policy and technical interventions could work in tandem to address fairness concerns more effectively.

The panel also agreed on the importance—in any application of an algorithmic system—of *clearly articulating an organization's goals and using technology to serve the best interests of users (who should also be clearly defined)*. Ideally, even for commercial applications, an organization's goals should entail prioritizing the safety and well-being of an organization's overall set of users and refraining from perpetuating biases or harms, even if at the expense of short-term profits. However, returning to the theme of contextualizing algorithms, panelists acknowledged that the question of setting goals depends heavily on placing the algorithmic system into its proper context. The scale of a system can matter, systems can have different uses to different people, and determining whether a system aligns with socially desirable values or some "ground truth" can depend on the application. Highly contested social issues are particularly challenging.

Take, for example, the case of Holocaust-denial content and Google's search algorithm. As covered extensively by the media in late 2016, when first reported on, search results after a query for "Did the Holocaust happen?" were returning a link to Holocaust-denying content on the first page of search results, often the first or second result listed. As tech journalist Danny Sullivan observed in a piece summarizing the incident, evidence indicated that Google ultimately tweaked its algorithms, such that denial content was less prioritized (i.e., moved off of the first page of results) or didn't appear altogether after the query. This example shows the kinds of difficult, fairness-related questions that can be raised when it comes algorithms. In this case, an algorithm was optimized in such a way as to serve the interests of promoters of Holocaust-denial content. But that also had the effect of promoting a historical falsehood and offending the sensibilities of those who recognize the atrocities and horrific legacy left by the Holocaust. Manually tweaking the algorithm to demote or hide the denial websites would promote historical accuracy but would go against the interests of denial-content publishers and users seeking Holocaust-denial content (whatever the motivation). Here, we can assume that the vast majority of users—and society as a whole—preferred not to promote such content and instead prioritized the "ground truth." But other cases aren't as clear; one need only consider the current state of American politics and news media to realize that "ground truth" is often a fuzzy, contested concept.

Getting away from the issue of what is "ground truth" or not, a workshop participant compared the Holocaust-denial example to Google's behavior in providing authoritative answers or clear nudges for other socially important search queries, such as conspicuously displaying contact information for suicide-prevention or domestic-abuse aid centers after subject-specific queries. Building on this discussion, another participant pointed to Google's 2016 decision to ban payday-loan advertisements from being highlighted to users of its search services. (Connecting us back to the pre-trial detention and bail-system discussion above, Google also recently banned ads for bail-bond services.) With all of these examples, panelists agreed that the ultimate question comes down to how to best achieve the goal of promoting the best interests of users as a whole. However, as the examples illustrate, situations are rarely clear-cut, users' interests are not always aligned, algorithmic tweaks can raise concerns over corporate power, and determining clear organizational goals can be tricky, especially with regard to socially contested issues.

RECOMMENDATIONS

While this first panel served primarily to frame discussions for the rest of the day, we conclude with some recommendations for researchers and practitioners that emerged from the panel.

1) ***Gain insights into how technical experts develop algorithmic systems in practice***. The vast majority of public scholarship on fairness and algorithmic systems evaluates these systems from an external position, with limited information about the engineering and design teams that implement them. While not in any way abandoning these lines of research, researchers should also develop theories of, and empirical research into, how technology frameworks interface with the technical experts (data scientists, engineers, etc.) and organizations that design and deploy them. Do those experts have sufficient knowledge of sound research-design and methodological principles that they should use to inform and interrogate their work? Do they have an appreciation for how those principles relate to issues of fairness in their work? If not, why not? To what extent are different teams (e.g., engineering, design, product, policy) involved in the development, testing, and implementation of these systems? How are domain experts involved in the development of algorithmic systems or algorithms themselves? Finally, we emphasize that the dearth of research on such experts and their organizations is less a function of researchers not having thought to look into these questions but rather because many of the sites for these sorts of investigations are corporate/for-profit and not readily open to outside researchers. We urge the for-profit sector to facilitate the advancement of knowledge by making more concerted efforts to provide such researchers access to its work practices. And we urge researchers to be persistent and creative in findings ways to observe and collect data on technical experts in practice.

2) ***Think systematically and consider context.*** To better understand fairness implications, we implore practitioners, as well as more technically inclined researchers, to situate algorithmic software systems within broader systems and social contexts. A few concrete recommendations emanated from this broader, more abstract imperative.

(a) Develop clearer, more systematic language and standards for thinking about system design and research methodology. For example, **develop guidelines and better documentation around measurement** and incorporate them into the machine-learning pipeline. Perhaps more importantly, recognize that the data we can manage to collect are generally highly imperfect proxies for the things they are meant to represent (e.g., re-arrests as a measure for recidivism).

(b) **Draw on broader historical and sociological insights** to understand the domain of interest. For example, study reform movements to understand how fairness and justice problems are constructed by those with deep domain expertise and participation. Understand how algorithms fit into this (it may be that they are part of the problem) and whether they directly address those problems or redirect focus away to other, less critical problems.

(c) Ask **how and why people and organizations that design, profit from, and use algorithms conceptualize the domains in which algorithmic systems are applied**. For example, how do organizations and departments that contract with software companies to build or deploy these systems understand their capacities and limitations? When a system is initially developed for one application, market, or urisdiction, how transferable is it to other domains (i.e., how might changes in target jurisdictions or populations affect the form, function, and interpretation of a system)? What society-level histories, norms, values, and biases affect data inputs and decisions? How do these histories affect the reactions of users to being “classified,” particularly those who are members of under-represented and disadvantaged groups?

ACKNOWLEDGMENTS

We would like to thank all AFOG Summer Workshop participants for their insights, collegiality, and critical engagement during the workshop. In particular, we thank the panelists on Panel 1, “What a technical 'fix' for fairness can and can't accomplish:” Lena Z. Gunn (Electronic Frontier Foundation), Moritz Hardt (UC Berkeley Department of Electrical Engineering and Computer Sciences), Abigail Jacobs (UC Berkeley Haas School of Business), Andy Schou (Google), and moderator Sarah M. Brown (Brown University Division of Applied Mathematics). We would also like to thank Andrew Smart and event co-organizers Jen Gennai and Allison Woodruff for helpful feedback and suggestions on prior drafts of this report. Finally, we acknowledge the intellectual and financial support of our workshop sponsors, Google Trust & Safety and the University of California, Berkeley School of Information.

ABOUT AFOG

The Algorithmic Fairness & Opacity Working Group (AFOG) is made up of UC Berkeley faculty, postdocs, and graduate students at UC Berkeley. It is housed at Berkeley’s School of Information. AFOG is co-directed by Professors Jenna Burrell and Deirdre Mulligan and is funded by a research gift from Google Trust & Safety to support cross-disciplinary academic research and conversations between industry and academia to explore and address issues related to fairness and opacity in algorithms.

For more information visit: https://afog.berkeley.edu