AFOG workshop panel 3: human autonomy and empowerment

By Jenna Burrell

Published August 13, 2018

Berkeley SCHOOL OF INFORMATION

Google

INTRODUCTION

A growing body of work, much of it case-based or ethnographic, illustrates the negative consequences to humans when automation is put in place as an opaque and unimpeachable authority. These tools are sometimes applied in domains where they can have significant consequences for the quality of life or life opportunities of individuals. For example, mathematician Cathy O'Neill describes the experience of a young man, Kyle Behm, who was continually screened out of job opportunities by a questionnaire misappropriated from mental health screening. Tammy Dobbs, who is disabled by cerebral palsy, saw her health benefits radically reduced by a new allocation algorithm. In both cases the decision-making process was wholly opaque and lacked any built-in appeal process.

With systems operating at massive scale there are also the inevitable exceptions, the cases that break the assumptions of the decision-making tool and require special intervention from humans. For example, Zeynep Tufekci describes the case of a Facebook user who was stuck in an endless loop trying to keep a Facebook account in her name, which in English appeared to be an obscene word, but, when using a nickname, ran afoul of Facebook's "real names" policy. There's also the case of Mr. Null. These are examples of algorithms operating in the way they were designed to and with correct data.

In addition, there are concerns about erroneous data and errors in the implementation of the algorithms themselves (see post on a school-ranking error or the case of a recidivism risk score error). More sophisticated tools built using machine learning models can exhibit quirky failures of "common sense" (see hairless head in a clueless photo booth, or deep neural networks are easily fooled, or predicting pneumonia risk). Empowering humans to identify errors and provide oversight can be an important failsafe procedure.

Therefore alongside (allocative) fairness, another laudable industry goal would be to ensure, as much as possible, the rights of individuals subject to automated decision outcomes to investigate and appeal those decisions. In a similar vein, the 'right to explanation' provision in the European Union's recently implemented General Data Protection Regulation (GDPR)

reflects a recognition that data collection and processing can be disempowering because of their opaqueness (see, e.g., Selbst & Powles 2017 and Edwards & Veale 2018). This is especially critical (as some of the examples above illustrate) when (1) such algorithms are implemented at massive scale, potentially impacting a global user base, (2) on platforms that are hard or impossible to opt-out of, (3) where they are employed in domains (such as lending, criminal justice, or employment) that affect the life chances of individuals, and (4) where the complexity of the decision-making tool precludes immediate understanding.

This report from our "human autonomy and empowerment" panel from the first AFOG workshop is a call to make the consideration of user autonomy and, more broadly, human autonomy part of efforts to achieve fairness. In what follows we consider the ways platforms already support mechanisms of user and stakeholder feedback, to what extent these mechanisms support user autonomy, their applicability to the algorithmic fairness problem, and the strengths and shortcomings of such approaches overall. Finally, we offer a set of recommendations that include organizational arrangements and processes, platform policies, and design elements.

It could be that the anxiety around algorithmic decision-making, and the introduction of machine learning and deep learning into new domains, stems not from algorithms that are *unfair*, but from the very fact of automation. If algorithmic unfairness is the singular source of concern, then defining what 'fairness' means (often in terms of the 'fair' allocation of a resource) and implementing an algorithm accordingly would address this problem. And indeed one research direction pursues such an approach. Implicitly, it is researchers, programmers, system builders, project managers and not *users* who are the ones granted an active role in such a process. Alternatively we may ask, how can those who are *subject* to algorithmic classification be better supported to understand how these systems work and the role they play within these systems? How might users be supported to *appeal decisions made by these systems and participate in improving them*, particularly around problems of fairness? This is the question of human 'autonomy' and 'empowerment' our panel explored.

FLAG FOR UNFAIRNESS?

The flagging function offers built-in user feedback in its most lightweight form. Some machine learning based classification tools already allow users to correct classifications, ultimately helping to improve the overall accuracy of the tool. For example, spam filters, such as the one built into Gmail, allow emails to be manually recategorized by users. We distinguish functionality that has the potential to reshape a platform overall from personalization functions and favor examining the possibilities of the former. Broadly, flagging allows users to "participate in how the platform content is organized, ranked, valued, and presented to others" (Crawford and Gillespie 2014). A common flagging feature implemented across most platforms allows users to report 'offensive' content. This approach crowdsources the role of governance and the enforcement of site policies. It is a strategic way to leverage the attentional resources of users on massively scaled platforms that handle mountains of user-generated content.

We discuss the flagging function as an entry point for fleshing out what "autonomy" might mean and what it means to design the user interface of automated decision-making systems to leverage the knowledge users possess. While flagging functions can be useful to consider, they are ultimately highly limited.

In defense of the flagging function, it does offer a couple of benefits. For one, it is lightweight and requires a minimal investment of time for the users who employ it. Flagging functions also constrain and structure feedback in a way that can be aggregated and managed efficiently, which is useful on a massively scaled platform.

Flagging functions are often implemented to channel user responses into a preset list of possibilities. As a consequence, as Crawford and Gillespie (2014) note, "they leave little room for the articulation of concern." Such functions may not, in fact, be primarily about empowering users. Rather, they are about putting users to work. The flagging function effectively distributes the work of site governance and policy enforcement, alleviating the burden on platform providers to review all content. Flagging provides little room for users to raise questions about

or force a review of site policies and may have no lasting effect on how the platform operates. In some cases, feedback is less narrowly channeled, but where open-response is offered (such as from Google's 'send feedback' link), this presents the great challenge of triaging, making sense of, and incorporating feedback into site operations.

Can there be empowerment without transparency? When users flag or report something on the platform, there may be only limited feedback about what action was taken in response. Feedback often comes as a simple verdict that some content does or does not violate site policy, but the response may not specify how. Withholding full transparency of operations is justified as a way to prevent users from 'gaming' the platform. However, gaming is also an expression of user autonomy and empowerment, even if it works at cross-purposes to the goals and desires of platform providers or other users.

If we relate flagging to questions of fairness, another problem of transparency emerges. Problems of fairness in the distribution of a resource (i.e., allocative fairness) may not be possible for individual users to perceive because they require access to information about overall patterns across the system. This problem has, for example, long delayed efforts to address gender-based pay inequality in the tech industry and other fields. You know your own salary, but typically not that of all of your colleagues or anyone else in your same role across the industry.

VARIED USERS

Users have a particular expertise that draws from their identities, life circumstances, and personal experiences. Platforms are frequently designed with assumptions that fundamentally don't work for certain groups, often minority groups and those who are not well represented or understood within engineering and design teams. For example, Facebook's "real names" policy has proven particularly fallible for transgender people, drag queens, political dissidents living in authoritarian regimes (see Tufekci 2017), and American Indians. Our panel also discussed challenges to successfully providing choices, autonomy, and empowerment that have to do

with different attitudes and orientations to the platform, degrees of technical literacy, and general beliefs about automation:

(1) **Disengaged users:** While functions for providing feedback may be available, this does not mean they will be used. Complicated functions may pose a barrier such that only very dedicated or engaged 'super-users' will use them. Wikipedia is an interesting alternate model to consider. It provides an example of deep but uneven engagement. A community of volunteer editors are heavily and actively invested in debating site policies and practices, and editors often commit many hours per week to the platform. But to become part of this class, and to use Wikipedia's many specialized backend site-management tools, means overcoming a huge barrier to entry. By contrast, a mistrust in online platforms to take a problem seriously, or general apathy about a societal problem (like racism) that is reflected in an online platform, may mean some groups of users do not engage available tools for giving feedback or addressing problems.

(2) **Burden on minority groups:** Another question of fairness relates to the labor involved in providing feedback. If minority groups that are poorly understood by design and engineering teams face more misclassification or forms of harm from a system, is it fair for them to also be burdened with the work of bringing this to the attention of platform providers? Preventing problems of algorithmic (un)fairness starts with anticipating and preventing them within design teams. After that, relying on allies or organized groups, rather than individuals from affected groups, may be a better approach to addressing both (1) and (2) and enhance the autonomy of users. One example is HeartMob, a volunteer collective organized to support women facing mob harassment on social media platforms.

(3) **Users who mistrust automation:** It is possible that for some, user 'mistrust' in a platform has to do with assumptions about automation. If users think their complaint is being dumped into the mechanics of an unthinking machine, they may not bother. Research suggests, however, that some populations of users assume there is far *less* automation behind platforms like Google search than there actually is.

Users as workers and as generating profit

Some specific concerns around how the autonomy of users is undermined by online platforms have become the focus of public concern and media coverage. These include concerns about (1) user interface designs intended to subtly manipulate users toward an addictive engagement with the platform to ensure that 'time on platform' figures are as high as possible. These fuel a site's profitability as a platform for advertising; (2) motivating users to contribute their labor (generally by generating content) to a platform but without remuneration and without transparency about how the platform benefits (and how much) from this labor.

On platforms such as Amazon Mechanical Turk (AMT), while "Turkers" are aware that they are workers (since they formally receive pay), there are other issues. For example, there are questions about autonomy over work assignments. Some AMT workers want a choice about the projects they wish to contribute to because they want to avoid participating in projects whose aims they consider unethical. However, the design of work allocation algorithms often intentionally obscures the nature of the project. Furthermore, metric-driven assessment of workers on online platforms (including AMT, Uber, etc) place employers (or customers) in the position to review work and deem it unsuitable with no oversight from the platform as to the appropriateness. As a result workers may be denied pay or even be banned from the platform.

NOT ONLY USERS

We've employed the word 'user' around 40 times already in this document treating it as a kind of stand-in for ordinary people, the broad group who use a platform but do not possess domain-specific knowledge or technical expertise regarding the functioning of those platforms. However, the term 'user' structures and limits our thinking in problematic ways. For example, many of the automated decision-making tools that are frequently in the news impose classifications on individuals who do not directly manipulate the tool. For example, court officials use bail calculators and risk-recidivism tools but the people 'scored' by these tools are

court case defendants or prison inmates in the criminal justice system. Generally speaking, when people are shut out of using a tool, the functions within the tool (for example, to report problems) are likewise unavailable to them. Autonomy must not be understood only as a user interface design problem. If an inmate or accused believes a risk score is erroneous, by what mechanisms can (s)he seek review? If there is a "flagging" function available, then how does (s)he compel the person responsible for operating the tool to engage this function? A broader term, such as "stakeholder" (rather than "user"), moves us toward including individuals and groups subject to a classification tool and its consequences but who do not manipulate it directly.

AGONISM

Enriched modes of feedback could better empower users to identify problems and instigate action over platform fairness. There are some key examples of public incidents that brought unfairness or harms wrought by automated systems to public attention. In each, other communication mediums separate from the platform were critical to publicizing and motivating attention and action from platform providers. In one case, a racist label was attached by Google's image labeling algorithm to an image of two African-Americans who were captioned as "gorillas." The story went viral when Jacky Alciné, a security engineer whose photo was miscaptioned, tweeted about it. The problem of inappropriate content served to children and the dangers of YouTube's autoplay function was the subject of a widely circulated Medium article. Traditional online mass media continues to play a key role in publicizing incidents, as exemplified by ProPublica's coverage of bias in criminal risk assessment tools. In all cases, these avenues offered a much richer discussion than 'flagging' functions ever could. They also were done with full transparency, and with significant debate and discussion. However, each case depended on individuals who were able to leverage certain privileges such as a large social media following, a gift with written expression and argument, and the time to commit to a process of evaluation and auditing.

There are also examples of separate counterpublics created outside of the platform in question to support forms of agonism, for example, as a way to organize Amazon Mechanical Turk workers. In some cases, disagreement with platform owners leads to the implementation in code of an alternate vision by users, such as the Blocktogether tool that uses Twitter's API to implement collective blocking of Twitter accounts responsible for harassment. Some forms of productive agonism appear to be fundamentally at odds with the orderliness sought by algorithmically driven systems of allocation, classification, and decision-making. Many such systems are generally premised on the need to choose (rather than to present options), to predict and to thereby smooth over struggle or uncertainty, to show the winners (and hide the losers once in competition). While support for human autonomy that is built-in to online platforms is a way for platform owners to take responsibility for supporting this right, a space independent of the platform may also be necessary to better facilitate dissenting views.

RECOMMENDATIONS

What follows are some proposals for how to build systems that provide automated decision-making while ensuring human autonomy is a central goal. These proposals include organizational arrangements and processes, platform policies, and design elements.

(1) In for-profits, **organizational structures should be set up to enhance user empowerment and challenge autonomy-denying designs**. Although an employee may occupy a role that makes them responsible for user interface design and evaluation, such roles are not automatically positioned to facilitate the autonomy of users. For example, designs that seek to induce addictive engagement may serve business ends but not the interests or well-being of users. What specific organizational structures most effectively serves this goal is an open question and would benefit from research. Who do teams within the organization report to and can they be made more accountable to users rather than shareholders, for example? What metrics are used for determining a team's success (i.e. not 'daily active users')? Rather than creating separate teams, could a user advocacy function embedded in business teams help steer decisions in ways that preserve autonomy.

(2) **When possible make an "appeal" process available. Provide a way to request human review for those impacted by platform decisions.** Users who are unhappy with one search engine (Google) may use another (Bing, DuckDuckGo). However, in systems where users and stakeholders have no way to opt-out and no alternative system to employ instead, this is especially critical. Virginia Eubanks documents how Indiana's automated welfare enrollment system was seemingly designed to trick recipients into losing their coverage and frustratingly allowed no way to trigger human review. While this was likely successful towards the stated aim of reducing welfare costs, it accomplished this goal through ethically questionable means and by undermining the autonomy of people within the system through automation. People who were, by law, entitled to access these government resources were nonetheless denied reasonable mechanisms to apply for them.

(3) **Don't conceal the humans behind the curtain.** Make the human labor within a system more visible, not less. Overclaiming the degree of automation may serve the firms interests in technical competition or may impress investors or shareholders; it works *against* users seeking insight into how a system functions. It also hides the aspects of system design that leverage human assessments or entailed human deliberation. Wikipedia gives an example of how to tie automation to human responsibility. Over 3000 tasks are automated on Wikipedia using 'bots.' However, anyone can message the bot owner through the bot account. If the bot owner is non-responsive, their account can be taken away and their bot disabled.

(4) **Consider the "features" (the categories of input data) used in the classifier algorithm and whether to exclude *some* or *all* that represent characteristics or behaviors that those subject to classification cannot control**. Machine-learning classifiers train on massive quantities of data and often treat all available data as "fair game" to use, apart from any features that must be excluded by law, such as "protected classes" (gender, race, religion, sexual orientation, and age). That said, researchers have found that simply omitting "protected class" information can simply shift biased algorithmic decisions onto proxy variables due to "redundant encodings" as Hardt et al point out. Furthermore, excluding categories of

data can reduce classifier accuracy. Still a consideration for the 'fairness' of data used in classification might offer some limited progress toward autonomy goals while not solving the problem entirely.

(5) **Find ways to incentivize external review and reporting of "fairness" problems with the platform**. Don't allow the burden of reporting to fall solely on the shoulders of those who suffer the most from bias. Find ways to enroll allies to this work. Create financial incentives. For example, Amit Elazari looks to network security as a model and suggests offering "bug bounties" for fairness. This could complement the existing efforts by journalists and academics who are incentivized towards novelty (and publication) but may not, for example, be as dedicated to exposing examples of an already-identified class of algorithmic unfairness.

(6) **For-profit organizations should look at the alternative approaches to concealment / transparency provided on non-profit platforms and in other industries.** In particular, it is worth reconsidering what and how much of a system absolutely must be concealed to prevent the platform from being manipulated or 'gamed' or to protect proprietary secrets. Full transparency platforms, such as Wikipedia, demonstrate how preventing 'gaming' may not necessarily mean maximizing concealment. Some of the data used in automated decision-making cannot be easily 'gamed.' The example of the Fair Credit Reporting Act also suggests an alternative mindset. Disclosing 'reasons' for being denied credit (as is mandated by the FCRA) allows individuals to make changes to their money management practices that will allow them to receive credit in the future. Providing greater transparency will generally empower users.

(7) **Support open-ended feedback but also agonism.** Open-response feedback creates more "room for the articulation of concern" (Crawford and Gillespie 2014). Bidirectionality in the flow of feedback can be added to this so that users know what effects their efforts have had. Humans perceive fairness as a process of negotiation and of seeking compromise. Platforms support debate around their practices and policies to varying degrees. Wikipedia models this intentionally while Twitter seems to facilitate it unintentionally. Yet as already noted, handling and

making sense of open-ended feedback at scale is extraordinarily difficult if not impossible. Exploring the range of possibilities between highly-structured feedback and open debate is a promising area for future research to determine the most effective approaches. Such research could seek to enumerate the varied practices on different platforms (as Levy and Barocas model by evaluating how various platforms manage discrimination by users through design and policy choices). Research could take on the form of user studies to better understand experiences of providing feedback, or where participants walk through feedback processes in different forms.

Fairness definitions have been dominated by a concern with equal allocation, particularly in high-stakes domains, such as employment and criminal justice. Interest in representational fairness draws attention to media depictions, denigration, and whether minority groups are even visible within algorithmic systems. There is also fairness as a matter of process or as the experience of debate, discussion, and negotiation. The autonomy of humans within these systems is, in part, about allowing users to recognize and report on problems of bias and unfairness. More fundamentally, depriving humans of their autonomy (as a pattern that frequently follows from automated decision-making) is a concern that is fundamentally a matter of justice. Populations are unequally subject to automation. In domains where some populations (and not others) retain their capacity to negotiate, seek an appeal, or make choices for themselves while others see it undermined or eliminated by automation, then it also becomes a matter of fairness.

ACKNOWLEDGMENTS

We thank the panelists on Panel 3, “User Autonomy and Empowerment:” Jen Gennai (Google), Stuart Geiger (UC-Berkeley Institute for Data Science), and Niloufar Salehi (Stanford University) for contributing their time and insights at the workshop. We would also like to thank Deirdre Mulligan, Daniel Kluttz, 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