Report from the first AFOG Summer Workshop

Jenna Burrell, Deirdre K. Mulligan, and Daniel N. Kluttz
Algorithmic Fairness and Opacity Working Group (AFOG)
School of Information
University of California, Berkeley
Workshop date: June 15, 2018
Workshop location: South Hall, UC Berkeley campus, Berkeley, CA
Published: August 13, 2018

Berkeley SCHOOL OF INFORMATION

Google

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 and moderators for each session. We also 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 the AFOG website at https://afog.berkeley.edu.

INTRODUCTION

On June 15, 2018, the Algorithmic Fairness and Opacity Working Group (AFOG) held a summer workshop with the theme, "Algorithms are Opaque and Unfair: Now What?" The event was organized by Berkeley I School Professors (and AFOG co-directors) Jenna Burrell and Deirdre Mulligan and postdoc Daniel Kluttz, and Allison Woodruff and Jen Gennai from Google. Our working group is generously sponsored by Google Trust and Safety and hosted at the UC Berkeley School of Information.

Inspired by questions that came up at our biweekly working group meetings during the 2017-2018 academic year, we organized four panels for the workshop. The panel topics raised issues that we felt required deeper consideration and debate. To make progress we brought together a diverse, interdisciplinary group of experts from academia, industry, and civil society in a workshop-style environment. In panel discussions, we considered potential ways of acting on algorithmic (un)fairness and opacity. We sought to consider the fullest possible range of 'solutions,' including technical implementations (algorithms, user-interface designs), law and policy, standard-setting, incentive programs, new organizational processes, labor organizing, and direct action.

BACKGROUND

Researchers (e.g., Barocas and Selbst 2016; Kleinberg et al. 2017), journalists (e.g., Miller 2015), and even the federal government (e.g., Executive Office of the President 2016) have become increasingly attuned to issues of algorithmic opacity, bias, and fairness, debating them across a range of applications, including criminal justice (Angwin et al. 2016, Chouldechova 2017, Berk et al. 2017), online advertising (Datta et al. 2018), natural language processing (Bolukbasi et al. 2016), consumer credit (Waddell 2016), and image recognition (Simonite 2017; Buolamwini and Gebru 2018).

There has been recent progress especially in understanding algorithmic fairness as a technical problem. Drawing from various formal definitions of fairness (see Narayanan 2018; Corbett-Davies and Goel 2018; Kleinberg et al. 2017), researchers have identified a range of techniques for addressing fairness in algorithm-driven classification and prediction. Some approaches focus on addressing allocative harms by fairly allocating opportunities or resources. These include fairness through awareness (Dwork et al. 2012), accuracy equity (Angwin et al. 2016; Dieterich et al. 2016), equality of opportunity (Hardt et al. 2016), and fairness constraints (Zafar et al. 2017). Other approaches tackle issues of representational harms which occur when a system diminishes specific groups or reinforces stereotypes based on identity (see Crawford 2017). Proposed solutions include corpus-level constraints to prevent the amplification of gender stereotypes in language corpora (Zhao et al. 2017), diversity algorithms (Drosou et al. 2017), causal reasoning to assess whether a protected attribute has an effect on a predictor (Kilbertus et al. 2017, Kusner et al. 2017), and inclusive benchmark datasets to address intersectional accuracy disparities (Buolamwini and Gebru 2018).

These new approaches are invaluable in motivating technical communities to think about the issues and make progress on addressing them. But the conversation neither starts nor ends there. Our interdisciplinary group sought to complement and challenge the technical framing of fairness and opacity issues. In our workshop, we considered the strengths and limitations of a technical approach and discussed where and when hand-offs, human augmentation, and oversight are valuable and necessary. We considered ways of engaging a wide-ranging set of perspectives and roles, including professionals with deep domain expertise, activists involved in reform efforts, financial auditors, scholars, as well as diverse system users and their allies. In doing so, we considered models that might be transferable looking to various fields including network security, financial auditing, safety critical systems, and civil rights campaigns.

THE PANELS

Below is a brief summary of the panel topics and general themes of the discussion. Full write-ups for each panel are linked. Our aim in these write ups is not to simply report a chronological

account of the panel, but to synthesize and extend the panel discussions. These panel reports take a position on the topic and offer a set of concrete proposals. We also seek to identify areas of limited knowledge, open questions, and research opportunities. We intend for these documents to inform an audience of researchers, implementers, practitioners, and policy-makers.

Panel 1 was entitled "What a technical 'fix' for fairness can and can't accomplish." Panelists and audience members discussed specific examples of problems of fairness (and justice), including cash bail in the criminal justice system, "bad faith" search phrases (e.g., the question, "Did the Holocaust happen?"), and representational harm in image-labeling. Panelists noted a key challenge that technology, on its own, is not good at explaining when it should not be used or when it has reached its limits. Panelists pointed out that understanding broader historical and sociological debates in the domain of application and investigating contemporary reform efforts, for example in criminal justice, can help to clarify the place of algorithmic prediction and classification tools in a given domain. Partnering with civil-society groups can ensure a sound basis for making tough decisions about when and how to intervene when a platform or software is found to be amplifying societal biases, is being gamed by "bad" actors, or otherwise facilitates harm to users. [READ REPORT]

Panelists for Panel 1: 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). Moderator: Sarah M. Brown (Brown University Division of Applied Mathematics).

Panel 2, entitled "Automated decision-making is imperfect, but it's arguably an improvement over biased human decision-making," describes a common rejoinder to criticism of automated decision-making. This panel sought to consider the assumptions of this comparison between humans and machine automation. There is a need to account for differences in the kinds of biases associated with human decision-making (including cognitive biases of all sorts) and those

uniquely generated by machine reasoning. The panel discussed the ways that humans rely on or reject decision-support software. For example, work by one of the panelists, Professor Angèle Christin, shows how algorithmic tools deployed in professional environments may be contested or ignored. Guidelines directed at humans about how to use particular systems of algorithmic classification in low- as opposed to high-stakes domains can go unheeded. This seemed to be the case in at least one example of how Amazon's facial recognition system has been applied in a law-enforcement context. Such cases underscore the point that humans aren't generally eliminated when automated-decision systems are deployed; they still decide how they are to be configured and implemented, which may disrupt whatever gains in "fairness" might otherwise be realized. Rather than working to establish which is better–human or machine decision-making–we suggest developing research on the most effective ways to bring automated tools and humans together to form hybrid decision-making systems. [READ REPORT]

Panelists for Panel 2: Angèle Christin (Stanford University Department of Communication), Marion Fourçade (UC Berkeley Department of Sociology), M. Mitchell (Google), Josh Kroll (UC Berkeley School of Information). Moderator: Deirdre Mulligan (UC Berkeley School of Information).

Panel 3 on "Human Autonomy and Empowerment" examined how we can enhance the autonomy of humans who are subject to automated decision-making tools. Focusing on "fairness" as a resource allocation or algorithmic problem tends to assume it is something to be worked out by experts. Taking an alternative approach, we discussed how users and other 'stakeholders' can identify errors, unfairness, and make other kinds of requests to influence and improve the platform or system in question. What is the best way to structure points of user feedback? Panelists pointed out that design possibilities range from lightweight feedback mechanisms to support for richer, agonistic debate. Not-for-profit models, such as Wikipedia, demonstrate the feasibility of high transparency and open debate about platform design. Yet participation on Wikipedia, while technically open to anyone, requires a high investment of time and energy to develop mastery of the platform and the norms of participation. "Flagging"

functions, on the other hand, are pervasive, lightweight tools found on most mainstream platforms. However, they often serve primarily to shift governance work onto users without the potential to fundamentally influence platform policies or practices. Furthermore, limiting consideration to the autonomy of platform users misses the crucial fact that many automated decisions are imposed on people who never use the system directly. [READ REPORT]

Panelists for Panel 3: Stuart Geiger (UC Berkeley Institute for Data Science), Jen Gennai (Google), and Niloufar Salehi (Stanford University Department of Computer Science).
Moderator: Jenna Burrell (UC Berkeley School of Information).

Panel 4 was entitled "Auditing Algorithms (from Within and from Without)." Probing issues of algorithmic accountability and oversight, panelists recognized that auditing (whether in finance or safety-critical industries) promotes a culture of "slow down and do a good job," which runs counter to the "move fast and break things" mindset that has long defined the tech industry. Yet corporations, including those in the tech sector, do have in-house auditing teams (in particular, for financial auditing) whose expertise and practices could serve as models. Generally, internal audits concern the quality of a process rather than the validity of the "outputs." Panelists pointed out that certain processes developed for traditional auditing might work for auditing "fairness," as well. A "design history file," for example, is required in the development of medical devices to provide transparency that facilitates FDA review. In the safety-critical arena, there are numerous techniques and approaches, including structured safety cases, hazard analysis, instrumentation and monitoring, and processes for accident investigation. But there are also particular challenges "fairness" presents to attempts to develop an audit process for algorithms and algorithmic systems. For one, and recalling Panel 1's discussion, there are numerous valid definitions of fairness. In addition, problems of "fairness" are often not self-evident or exposed through discrete incidents (as accidents are in safety-critical industries). These observations suggest a need to innovate auditing procedures if they are to be applied to the specific challenges of algorithmic fairness. [READ REPORT]

Panelists for Panel 4: Chuck Howell (MITRE), Danie Theron (Google), Michael Tschantz (International Computer Science Institute). Moderator: Allison Woodruff (Google).

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

AFOG workshop panel 2: Automated decision-making is imperfect, but it's arguably an improvement over biased human decision-making

By Deirdre Mulligan, Amit Elazari, Jenna Burrell, Daniel Kluttz

Published August 13, 2018

Berkeley SCHOOL OF INFORMATION

Google

INTRODUCTION

Among the goals driving the adoption of automated decision-making tools is the belief that they can protect against biased, inconsistent, and irrational human decision-making. Sometimes this belief stems from the assumption that human and machine processes are cognitively similar, while biases and other failings are positioned as uniquely attached to human cognition. Yet human and machines make decisions in fundamentally and qualitatively different ways. Hints of these differences can be seen in research on image identification. For example, humans hone in on the hidden objects in images quickly, while by contrast machine learning algorithms may classify images based on background data a human would identify as irrelevant--for example using the presence of snow in an image to classify images as wolves rather than huskies. Humans are good at recognizing individuals they know despite changes in hairstyle, glasses, makeup, and aging. Computers are often stumped by such changes, and even less subtle manipulations to inanimate objects like street signs can lead to surprising misclassifications. The distinct ways humans and machines construct knowledge reflect the different things they do well, as well as their respective blind spots.

Against these acknowledged differences, how should we evaluate automated decision-making systems? What metrics should we apply to determine whether machine learning systems or outputs are "fairer" than human processes? If machines and humans reason differently shouldn't we care not only about how well machine processes limit the biases attached to human cognition but also those biases that may be uniquely machine-made? Furthermore, in practice, automated decision-making tools are often positioned not to replace human roles, but to augment their decision making. How do we evaluate such hybrid decision-making processes? How might we ensure that the combination advances appropriate conceptions of fairness rather than compounding the biases each form of processing produces? What design or governance strategies could dynamically leverage the strengths of humans and machines?

This report from our AFOG Workshop panel, "Automated decision-making is imperfect, but it's arguably an improvement over biased human decision-making," is a call for more rigor in how

we evaluate the relative "fairness" of human and machine systems. Reducing unfair processes and outcomes requires systems that constrain human biases, address the distinct biases that may emerge from automated decision-making processes, and tailor decision-making to context-relevant fairness qualities. Doing so requires us to better understand the sources of bias in automated decision-making processes (calling to mind long-standing and rich research on human biases and heuristics in cognition and decision-making); methods to evaluate and compare human/machine/human + machine processes for "fairness"; and design and governance models that lead to progress on fairness, as well as traditional goals of improved predictive accuracy, efficiency, and speed. Fairness is context-specific, often contested by various stakeholders within that context, and typically includes both substantive rule(s) and the procedures for selecting and applying it.

Below, we consider the implications for advancing and measuring "fairness" in terms of:

- the differences in how machines and humans approach the construction of knowledge and develop intuition; and,
- the messy ways humans integrate, resist, and tinker with automated decision-making systems.

Finally, we offer a set of tentative strategies that reframe the question, from 'which is better' to how to create and configure more just systems in light of these complexities.

STARTING WITH THE DATA

Panelists and participants felt this problem had to be tackled from the bottom up, starting with close consideration of the data. There was a shared sense that addressing bias requires developing an intense focus and sensitivity to the contexts, methods, and instruments of data-collection. This emphasis on understanding data is common across the social and physical sciences, but panelists felt it was just being reckoned with in data and computer science. Claims that data is "objective" and speaks for itself, or mere lack of focus on the lived history of data, pose a barrier to discussions about fairness. The data may be self-reported and therefore

suffer from well-known biases (such as reporting bias), captured through processes or in geographies that limit generalizability, and obtained from processes or systems that embed specific societal biases, such as racism and sexism. The data or identified features may be a poor proxy for the phenomena stakeholders are hoping to measure or predict (i.e., they may lack ecological validity). Attempts to address bias will fail if researchers and practitioners fail to attend to the selection process, tools and methods of collection, and the social practices and institutions that shape the biases of data fed into machine learning systems.

THOUGHTS ON DISTRIBUTING DECISION-MAKING BETWEEN HUMANS AND MACHINES

Rather than asking whether machine learning can "free us" from human biases, panelists advocated a closer examination of the particular ways machines and humans embed and express bias. This sort of detailed analysis could provide roadmaps for allocating work in ways that minimize problematic bias, as defined by context.

The introduction of automated decision-support systems into work practices often shifts control over the kinds of data available and considered relevant as well as who and what makes such determinations. These shifts can displace professionals' or other users' ways of constructing knowledge and making decisions. These shifts can lead domain experts to contest the data, logic, and judgments of automated systems.

The introduction of newsroom metrics into journalism provides an example of tensions between algorithmic decisions and human judgment in a professional context. The algorithmic decision-support tool Chartbeat was introduced to help journalists succeed, but its value was contested by journalists who understood it to be an effort to displace their judgment about newsworthiness. Chartbeat supplied metrics based on a story's virality and other data indicating audience engagement, but it risked elevating profit over what journalists considered quality journalism. Assessing quality journalism is a complicated task, burdened by definitional problems as well as the money and time required to do holistic, generally qualitative,

assessments. Due to the difficulty of defining and measuring something akin to quality, automated decision-support tools such as Chartbeat rely on poor proxies, such as circulation on social media, likes, page views, etc.

Another example comes from the history of the consumer credit industry. Early processes used variables such as clothing, gender, and race to determine creditworthiness. We've progressively moved to a world where these explicitly discriminatory data have been replaced by data more facially neutral and more directly relevant to judging likelihood to repay (e.g., information about financial history). But because the facially neutral data mask unequal histories of market exclusion, devaluation of labor, and other manifestations of both individual and institutionalized discrimination, these systems reproduce the inequitable access to credit along race and gender lines they sought to alleviate. While algorithmic systems do not harbor the cognitive biases that plague human decision-making, when algorithmic systems ingest data rife with the detritus of those cognitive biases, they have the damning effect of perpetuating them with a new sheen of legitimacy born from claims of objectivity. Seen from this light, the new system launders discrimination under the guise of neutral indicators of creditworthiness. An automated system envisioned as meritocracy at scale merely makes the history of discriminatory practices harder to see and confront. Historically marginalized groups are consequently trapped in a web of so-called objective and fair measures that are decidedly unjust. Instead of eliminating bias it sublimates it.

Judgments on Data Relevance

Automated processes require agreement on what data is relevant to a given decision. The data considered relevant in a system may be vast--potentially tens of thousands of different categories of input data--but it is a closed set. In contrast, human processes may delineate the data relevant to a task but individual humans can spontaneously bring new information into the process in real-time. While humans cannot process the tens of thousands of data points an algorithm can, they can selectively pull data in and out of the decision making frame based on case-specific, situational knowledge. Constraining the decision maker's ability to expand, or

narrow, the data used to render a decision can upset context-specific or domain-specific perspectives on fairness (again, fairness may be defined differently by stakeholders other than the user in a given context).

In addition to constraining the data that can be considered, the shift to automated decision making or decision support systems can also replace professionals' logic with the choices of the engineer or system designer. The constraints automated decision making processes place on users can be particularly corrosive to context-specific definitions of fairness when relevant professional, regional, or site-specific experts are not consulted in the systems development. For example, criminal justice risk-assessment tools, which have been around for decades and are often simply logistic regressions, are almost uniformly created outside of the jurisdictions in which they are deployed. There are less than sixty tools used across the entire US. Research found that these, and other common automated decision-support tools, are generally acquired as commercial off-the-shelf products, rather than collaboratively developed or tailored for the conditions and context of use.

Human Engagement with Decision-Support Tools in Professional Contexts

While researchers have documented automation bias--deference to machine decisions--research by Angele Christin finds that in some instances professionals resist algorithms just as they do other tools that are introduced into the workplace. For example Christin found different kinds of resistance and tinkering with risk-recidivism tools in the justice system. Some of that resistance appears to be grounded in conceptions of fairness. For example, a senior judge whom Christin interviewed said of such tools, "I don't look at the numbers. There are things you can't quantify . . . You can take the same case, with the same defendant, the same criminal record, the same judge, the same attorney, the same prosecutor, and get two different decisions in different courts. Or you can take the same case, with the same defendant, the same judge, etc., at a two-week interval and have completely different decision. Is that justice? I think it is" (Christin 2017). Thus, justice (which may or may not always align with fairness) from

his perspective was served by discretion rather than rigidity. Christin found probation officers similarly resisting the rigidity by tinkering with the criteria to obtain the score they thought adequate for a given defendant. Legal professionals questioned why they should follow a completely opaque model, developed by a private corporation, over their own professional judgment.

The resistance to the risk-assessment tool rested on four distinct but connected claims: that it didn't capture professional judgment; that predetermined limits on what data could be considered in decision-making were inappropriate; that use of an opaque system was inappropriate; and that deference to a corporate system was inappropriate. The first two raise questions about what is fair: decisions on a fixed set of data guided by set rules versus variations in data and analysis to address subject and context specific circumstances across a population with diverse histories. The second two objections reflect commitments to procedural fairness including access to the information and decisional rules and constraints on delegation of decision-making authority to unaccountable profit-driven private parties. These points of resistance are grounded in fairness concerns that cannot be addressed by formalizing the "right" definition of fairness within the automated system.

These concerns with delegation may be intensified where machine learning is offered as a service with preconfigured defaults. For example, using the default "confidence threshold" of 80% in Amazon's company's face-matching technology, Rekognition, incorrectly matched 28 members of Congress with arrestees in the database--a 5% error rate among legislators--with a disproportionate number of false positives for African-American and Latino members. Amazon's system documentation contains some language (on page 131 of 433) suggesting law enforcement use a confidence threshold of 99%:

> All machine learning systems are probabilistic. You should use your judgment in setting the right similarity threshold, depending on your use case. For example, if you're looking to build a photos app to identify similar-looking family members, you might choose a lower threshold (such as 80%). On the other hand, for many law enforcement

use cases, we recommend using a high threshold value of 99% or above to reduce accidental misidentification.

However, this advice is not necessarily being carried out in practice. A guest blogpost by a Senior Information Systems Analyst for the Washington County Sheriff's Office on Amazon Web Services instructs law enforcement to use a confidence threshold of 85% when using Rekognition. This combination of preset defaults, buried recommendations, and conflicting advice undermine fairness by limiting the likelihood system users will identify and make configuration and data choices that align with contextually relevant definitions of fair treatment in high-stakes applications.

Dignity as/and Fairness

A final fairness concern that cannot be addressed by formalizing a correct definition is reflected in the EU General Data Protection Regulation, which views fully automated decision-making as presumptively unfair. According to the Working Party, Article 22 of the GDPR creates a "general prohibition" on solely automated decisions that have legal or similarly significant effects. There is no parallel to this prohibition in United States law. The GDPR prohibition reflects a distaste for machines judging humans--regardless of whether machine processes produce more fair or just outputs--grounded in dignitary interests. The unfairness isn't about unequal treatment or unfair processes but rather about reducing an individual to a set of data. Again, this aspect of fairness cannot be addressed by improving the data or reasoning of the automated decision-making system. As Meg Leta Jones explains, in Europe, automated processing has been used as a mechanism for oppression. National data protection frameworks, such as those of France and Germany, reflected this experience and connect data protection to dignity and personality. These member state regimes influenced developments in EU data protection law, and "[a] particular idea of dignity can be found in rulemaking processes across Europe that protected humans from being treated as data to be processed by machines" and in the references and nested frameworks in which data protection professionals position data protection. To preserve this dignity, the European approach seeks to ensure a "human in the loop human as a

regulatory tool to address the effects of automation[.]" As Jones explains while "...the person and people of Europe may be legally constituted as entities protected from automated decision-making and deserving of a human in the loop, those in the US are protected from the flaws of humanity through the computational neutrality of information systems."

Intuition?

Despite these differences, automated decision-making systems and seasoned professionals are similar in that both are, to some degree, opaque and cannot fully explain their decision-making processes. Professionals draw on their intuitions making quick judgments honed from exposure to thousands of cases. Computer decisions are perhaps not so different, also developing expertise from great quantities of data. But computer intuition does differ from its human counterpart because it arrives at its conclusions by computational (rather than neurological) processes. Its 'reasoning' is thereby comparatively harder for humans to comprehend. For example, the 'Go' playing AI built by Google's DeepMind developed strategies for the game that had not yet been discovered and that human Go players now learn from. This computer intuition raises a new type of fairness question: the circumstances under which it is acceptable to learn from, and/or act on, computer intuition that predicts things beyond human intuition.

"Intuition" may be the secret ingredient of seasoned professional judgment and rich machine-learning models, but the intuitions developed will be distinct. First, machine-learning models are often trained on data that represent professionals' decisions and related outcomes, rather than professionals' decision-making processes. Learning from actions and outcomes may build a machine intuition that bears little resemblance to the intuition that guides a professional's decision-making process. Second, machines and humans 'see' in different ways--machines can identify complex patterns and scan across massive data sets; humans can identify things they've seen (such as faces), despite a wide range of subtle and relatively extreme perturbations (changes to hair style, plastic surgery, aging etc.) The different intuitions developed by human and machine systems may produce similar outputs in some instances but

not in others. More generally, observational data about decisions and outcomes is unlikely to capture professional judgment, and even with the same data, machines and humans will find and miss different things within it. Rather than building on and improving professional logic and intuition, complex automated decision-making systems are likely to replace it with distinctly computational intuitions that "not only...depart from intuition, but...might not even lend themselves to hypotheses about what accounts for the models' discoveries."

Assuming that automated systems can improve on existing decision-making processes muddies the waters. We should think about when the ways that machines see, reason, and develop intuition can improve the fairness of a system and its outputs. Sometimes we may be better off without the automated process. Sometimes, automation may offer more of a Schumpeterian path to better outcomes, rather than an incremental improvement on human reasoning.

Is fairness the right question? On reflection panelists and participants questioned whether justice might be a more appropriate goal. Addressing fairness focuses on the comparative question of when it is "unfair to distinguish among people and treat them differently and why?" while a focus on justice would focus on whether and how the introduction of algorithmic systems advance human rights, and ultimately designing, using, and avoiding them toward that end.

RECOMMENDATIONS

1. **Draw from prior research on ways of arranging human-machine handoffs**. There are bodies of research that consider how technology enhances human capacities. For example, work on "distributed cognition" analyzes memory work in complex tasks (such as sailing or airplane navigation) as well as other cognitive processes as something distributed between humans and instrumentation. Do insights from this work translate into new domains of human-machine partnership? How relevant are they to the application of machine learning and other AI systems?

2. **Research real-world practices of augmented decision-making** to understand how human actors (particularly those with domain expertise occupying professional roles) reason about decision-support tools. What do humans do when their own decisions and that of the machines diverge? To what extent do humans even consult decision-support tools made available to them? Christin (2017) models this approach in her comparison of professionals working in web journalism and criminal justice. She finds that in criminal justice in particular professionals refer very little to scoring systems and algorithmically determined rankings. There are unanswered questions about the extent of "automation bias" (i.e the assumption that the machine must be 'correct') in different settings. There are also, so far, few guidelines for human actors about how to relate to these systems, when to trust or mistrust them, or how to contest or question them in different domains of application.

3. **Build interpretable machine learning models when appropriate and useful, and conduct user studies** (for example see Lakkaraju et al) to systematically evaluate user understanding of model behavior. This is the foundation for understanding possibilities for human/machine partnership and its limits.

4. **Consider alternative metrics**. In particular, what are the alternatives to "accuracy" in comparing human and machine decision-making? Often arguments about the superiority of prediction or classification in machine learning rest on accuracy measures. Yet accuracy can be artificially produced through overfitting when training machine learning models. To achieve an effective comparison between humans and machines, particularly when questioning decision-making fairness, requires considering the specific differences in classification they produce. How did a machine arrive at classification A while a human arrived at B and what does this tell us about the differences in reasoning between these two forms of decision-making? Other comparison metrics could include fairness itself. In a narrow sense this might mean how error rates differ between classes (for example how men versus women are categorized as qualified to take out a loan). In a broader and

more ambiguous sense, it might mean comparing how "interpretable" a classification made by a machine vs. human is (with an eye on facilitating human partnership and oversight).

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 2, "Automated decision-making is imperfect, but it's arguably an improvement over biased human decision-making:" Angèle Christin (Stanford University Department of Communication), Marion Fourçade (UC Berkeley Department of Sociology), M. Mitchell (Google), Josh Kroll (UC Berkeley School of Information), and moderator Deirdre Mulligan (UC Berkeley School of Information). We would also like to thank Amit Elazari 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

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

AFOG workshop panel 4: From the black box society to the audit society — are algorithms auditable?

By Andrew Smart

Published August 13, 2018

Berkeley SCHOOL OF INFORMATION

Google

INTRODUCTION

One of the topics discussed at the AFOG workshop was the role that algorithmic audits might play in ensuring algorithmic fairness and accountability. The question is how to develop rigorous approaches that can determine whether algorithms behave in undesirable ways and may also introduce a higher standard of accountability for companies and public institutions designing and programming algorithms. In the ideal case, algorithmic audits should give a voice to all stakeholders. Given the steady stream of cases of algorithmic unfairness both in the media and in academic research, and the recognition that algorithmic decisions will become more pervasive and more consequential, there is an urgency around holding algorithms accountable for their actions. In keeping with this trend, a commercial race seems to be emerging between nascent startups and small consultancies and large established companies offering algorithmic audits for bias and unfairness.

As Kroll (2015) points out, "In general, accountability fosters important social values, such as fairness, transparency, and due process, each of which is a deep subject addressed by an enormous literature." The goal of having accountable algorithms is to make sure that the increasing use of algorithms in every domain furthers these social values. Algorithmic audits, like audits in more established industries, are techniques for verifying that technological and business practices are accountable to important social values. The ideal purpose of audits is to enable society to verify that its expectations of public and private organizations are met, and to offer a way to articulate what it means for organizations to behave in an accountable and responsible manner, and to provide evidence of performance to a standard (internal or external). Audits are also a technique for controlling risk in large organizations.

A recurrent theme from panelists on audits was how the culture of tech has traditionally favored a rapid product development cycle, whereas "auditing" requires slowing down to check that the concerns of many stakeholders are addressed, and that the standards against which one is being audited are met. Underlying this clash of cultures for algorithmic auditing for fairness are the problems of *what* and *whose* values get to decide what is fair? The challenge is to define

what the goals or standards are, and then how can companies and practitioners show outsiders that they are working toward or achieving those goals. Even if tech companies begin auditing their algorithms and development processes against fairness criteria, this is no guarantee that algorithms will become fair. The hope is that, at the least, they will be less unfair. There are also multiple sources of external unfairness in society and the question is to what extent do algorithmic systems reinforce existing social stratification, create new forms of stratification, and what algorithms can do to mitigate social injustice?

The following sections give an overview of what audits are, a brief history of audits and the concerns that motivate auditing in several areas: sociology, law, medicine, safety-critical industries, and finance. We also try to draw lessons for algorithmic auditing from these domains. Finally we examine future directions for research and how to move forward in a thoughtful way.

WHAT IS AN AUDIT?

Audits are tools for interrogating complex processes to determine whether these processes are compliant with company policy, industry standards or regulations. While there are many similarities across fields in terms of what an audit is, there are key differences, too. Power (1997) describes the myriad types of audits, “In addition to financial audits, there are now environmental audits, value for money audits, management audits, forensic audits, data audits, intellectual property audits, medical audits, teaching audits, technology audits, stress audits, democracy audits and many others besides.” And we now add to that list, “algorithmic audits.” Audits may be considered what Floridi (2014) calls *metatechnologies,* or second- and third-order technologies that operate on and regulate other technologies; metatechnologies include the sociotechnical conditions in which the technology is embedded (e.g., rules, conventions, and laws).

For clinical trials in medicine for example, an audit is defined as “a systematic and independent examination of trial related activities and documents to determine whether the evaluated trial

related activities were conducted, and the data were recorded, analyzed and accurately reported according to the protocol, the sponsors standard operating procedures, Good Clinical Practice, and the applicable regulatory requirements." The IEEE defines an audit for software as, "an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures." The idea of a *systematic* and *independent examination* or *evaluation* is a common ideal for audits in general. The audit is meant to check that a process and its result were done according to a value-based and agreed-upon set of rules or standards.

Ideally, audits are concerned with not only the output of a specific system, but also with the checks, controls, and quality of the system generating the output. A system with poor quality controls may produce good outputs by chance, but there may be a high risk of the system producing an error unless the controls are improved. The theory is that if an organization has sufficient control over its processes, whatever downstream or emergent property is desired--e.g., "airline safety", "pharmaceutical safety", "financial trust," and now "algorithmic fairness"--can be guaranteed (or at least the likelihood of a failure can be decreased) by an auditable (i.e., controllable) process. Auditing itself is its own industry with experts, literature, and conferences. And it's not that large tech platforms don't already audit many of their processes: financial auditing is a well-developed practice in the tech industry.

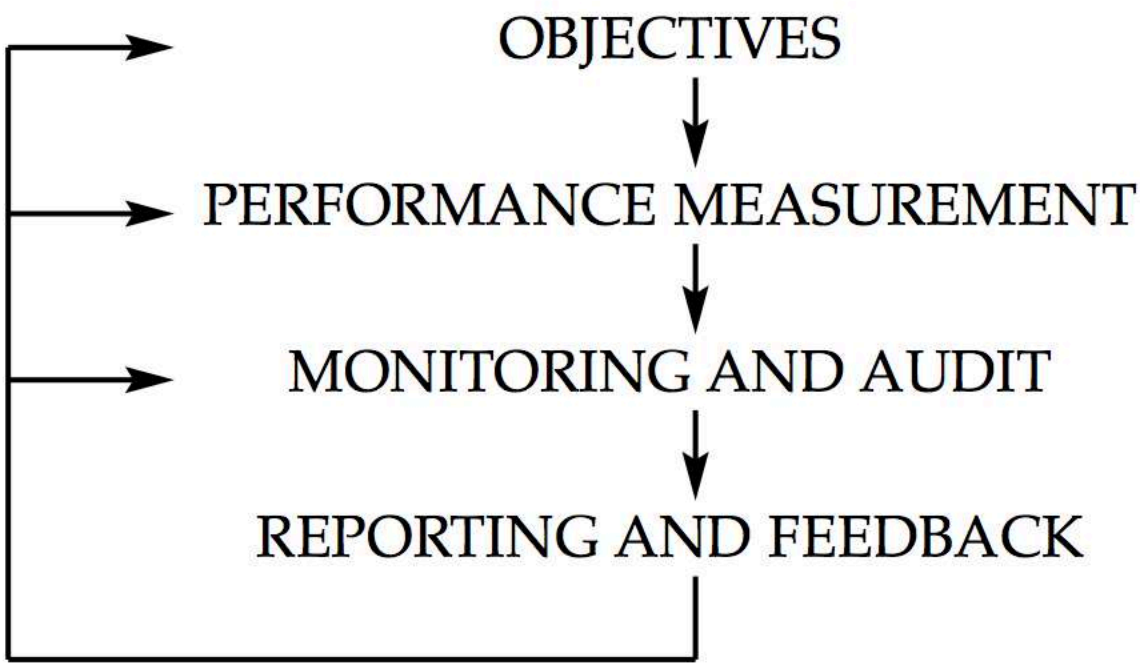

Figure 1: **A General Model of Self-Auditing and Control**

Figure 1 is taken from the ISO 9000 standard for quality assurance. The idea is to enable organizations to establish their procedures for ensuring quality, what Power (1999) calls "control of control." For machine learning, one might speak of "algorithms of algorithms."

Subjecting the machine learning technology development process to audits whether internal or external is not common practice. Are algorithms even the kinds of objects that can be audited? Burrell (2016) points out that algorithms and the organizational decision making that produces them are often deliberately opaque, and algorithms are protected as trade secrets and to prevent hostile actors from manipulating online services. However, these are justifications for resisting external audits, but are not practical reasons preventing audits. Then there is the inherent inscrutability of deep-learning models (Selbst and Barocas, 2018). Complete

transparency in terms of releasing code is has drawbacks and seems poorly suited to meeting fairness concerns. Transparency also has several serious disadvantages in terms of privacy and the risks of bad actors gaming systems. Using a technique known as “model inversion” it is already possible to infer private data from analyzing algorithms (Veale, 2018). Sandvig (2014) proposes five types of potential algorithmic audits which could answer questions about human values: 1) code audits; 2) noninvasive user audits; 3) scraping audits; 4) sock puppet audits; and 5) crowdsourced/collaborative audits. However as Deirdre Mulligan points out, in civil rights law items 2 & 4 from Sandvig’s list would be considered testing and distinct from auditing. “Testing” is a common practice when officials are looking for evidence of dicrimination. Kroll et al., (2017) propose a type of partial transparency using cryptographic techniques to evaluate if an algorithm meets certain independently-defined value criteria, such as fairness, without needing to explicitly check source code.

WHAT SHOULD WE AUDIT?

It is therefore important to distinguish what exactly might be audited when we discuss auditing algorithms, and what standards - processes or outcomes - they should be audited against. Normatively, an external auditor should inspect a machine-learning product development process to ensure that proper safeguards are in place to protect human values such as fairness. As mentioned previously, this would require process-documentation and outputs from the models to check against predefined and agreed-upon metrics. These metrics would have to be designed so as to capture the values we care about.

Sandvig (2014) and Kroll (2017) propose technical and computational solutions to auditing algorithms. This is likely more feasible given the aforementioned opaque way in which algorithms are designed and the secrecy surrounding the development process at private companies. Internal auditing could also take the form of evaluating the product development process itself, as is done in other regulated industries. This would require that these companies pursue “auditability” in terms of design plans and documentation. There are open questions around the lack of standards around algorithm development. What sorts of artifacts or

documents could an auditor demand that might satisfy some predefined criteria? However, even if an auditor were to find something "adverse" during the audit, who is accountable? In the case of aviation, medicine, and finance, there are regulatory bodies with varying degrees of legal authority to impose fines, take away certifications, or even demand documentation from companies.

In computer science there is a distinction between "black box" auditing where the auditor only has access to the output of the system, and "white box" auditing where an auditor knows the internal workings of the program and/or the processes involved in the development of the system. However, neither black or white box auditing guarantees that the root cause of a behavior can be discovered. The difference between black box versus white box auditing is mirrored in the distinction between internal auditing and external auditing. Most companies in regulated industries like pharmaceuticals perform internal audits, and they are subject to external audits. Often, these companies conduct internal audits in order to prepare for external audits.

A typical machine learning development process might be as follows: 1) product concept 2) product design 3) data collection 4) data processing 5) model selection or design 6) model training 7) model evaluation 8) product testing, and, finally, 9) putting the trained model into production systems. Importantly, the joint statistical distribution of the training and test data are not the same as the underlying distribution of new data from the real world after the model is in production. In other words, the model used in production is using a certain, more or less "representative" distribution, but that is different from the "real world" distribution about which the model is making predictions.

At each stage of development, there may be several organizations or teams involved in decision making. The goal is usually to make sure a product launches and "lands" successfully according to success metrics set out for the product. These can be any number of key performance indicators (KPIs); for example, an increase in Daily Active Users (DAUs) or number

of downloads. Many attempts are now being made to include fairness metrics in these objectives.

In safety-critical industries and finance, the governance of organizational, technological and managerial processes are constructed to be transparent and auditable. It's also important to note that appropriate documentation for auditing purposes is also intended to improve an organization's control over its product development. Finally, the concept of "tracability" is crucial in these industries: the behavior of the final built system, its components, their origin, and all key decisions should be traceable back to the original design requirements and intended use or function.

Larger research questions for industry and academia include the following:

1) What can (and can't) we borrow from other industries and practices and apply to the unique issues surrounding audits of algorithms?
2) What would an algorithmic audit look like in practice? What would we have to adapt from other types of audits to suit machine-learning algorithms? What if audits themselves are opaque?
3) Is the algorithm the right target for an audit, or would it be wiser to look at the larger system in which the algorithm operates?
4) How can we avoid some of the pitfalls of creating a mere check-box activity or an "audit society" with layers of bureaucracy?

A BRIEF HISTORY OF AUDITS

> "Wherefore in all great works are Clerks so much desired?
> Wherefore are Auditors so well fed…?
> Because that by number such things they finde,
> which else would farre excell mans minde."
>
> Robert Recorde (1540)

Concerns about transparency and accountability are nothing new of course. Health and safety, medicine, security, education, intellectual property, aviation, discrimination law and policy, and corporate finance each have long histories and interactions with audits. It is also interesting to consider that auditing is fundamentally driven by calculation and quantification, just as algorithms are. Yet there is a sense now that Big Data and predictive algorithms, which quantify on an unprecedented scale, have become unaccountable. It's useful to review the history of auditing in other domains to draw lessons that may inform algorithmic audits.

Audit studies in law and sociology

Audit studies, most often conducted in law and sociology to detect housing or job discrimination, date from the WWII era. An audit study to detect discrimination is a type of field experiment that stages randomized encounters between auditors and decision makers (e.g., landlords). The auditors are as closely matched as possible in all features and characteristics except the one under investigation (usually race or gender). In the case of racial discrimination in housing, for example, the theory is that if the landlord offers a home to a white person over a "matched-in-all-other-aspects" black person, then the researcher can infer that racial discrimination motivated the decision. Thus, "race" here becomes an isolated treatment effect, similar to the drug/placebo given in a randomized controlled clinical trial. Kohler-Haussman argues, however, that "audit studies do not measure the objective isolated treatment effect of race and race alone because there is no such thing to measure." In other words, treating race as an isolatable trait ignores the thick ethical and sociological underpinnings of race as a social construct. How social categories like race and gender are constituted is an important consideration for algorithmic fairness and auditing algorithms *qua* decision makers, as well.

For example in the criminal justice domain, the much-discussed COMPAS algorithm for recidivism risk was audited by its designers and found to be "good", but when ProPublica used different metrics based around criminal defendants' point of view, the COMPAS algorithm was

clearly unfair. The question is what metrics should an organization choose to audit against? And who should have a say in the choice of metrics? Echoing a theme that emerged during earlier panel discussions on evaluating algorithmic fairness, what is meaningful to measure and how can we know what those measurements are? For algorithmic fairness audits, we must define our objectives and how to measure performance. Defining and measuring fairness is a notoriously difficult task in machine learning (Kleinberg, 2016).

Finance audits

Financial auditing had to play catch up as the complexity and automation of many financial business practices became too unwieldy to manage manually, thus stakeholders in large companies, and government regulators desired a way to hold companies accountable. Concerns among regulators and shareholders that the managers in large financial firms would squander profits from newly created financial instruments prompted the development finance audits.

Additionally, as financial transactions and markets became more automated, abstract, and opaque, threats to social and economic values were answered increasingly with audits. But financial auditing lagged behind the process of technology-enabled financialization of markets and firms. Similarly today, as algorithms invade more and more aspects of life, audits are proposed as an answer to perceived threats to important values like privacy, autonomy, and ethical behavior - but a framework to audit algorithms is lagging behind.

However, the limits of auditing have been seen most notably after the financial crash of 2008. Financial auditing was widely criticized in the wake of the 2008 financial crash as most firms that suddenly failed had received positive audit results immediately prior to public declaration of financial difficulties (Sikka, 2009: 869 - cited in Styhre, 2015: 147).

Safety engineering

The fields of fault tolerance and safety engineering have developed a rich set of tools and techniques for analyzing high-assurance and fault-tolerant systems and their risks. These include analyses of systems like avionics on commercial aircraft and submarines. But it is generally recognized that audits are no guarantee of safety. Indeed, audits can provide a false sense of security and an organization can slowly drift toward unsafe decisions. If audits become a mindless check-box activity, then they might be used to claim an organization is doing something it is actually not. And once safety standards are relaxed, they typically don't snap back until there is a catastrophe.

As mentioned, the use of audits as a means of controlling risk in other industries reveals a fundamental cultural difference between safety-critical and the tech industries: risk-aversion, regulation and safety-first thinking using the precautionary principle for the former versus "move fast and break things" for the latter (or as one panelist put it, "slow down and do a good job" versus "permissionless innovation"). In contrast to the "ship it and fix it later" ethos that has defined the tech industry, safety engineering requires that the developer define what must be avoided (e.g., airplane crashes, patient death) and engineer backwards from there.

Medicine and pharmaceutical

Internal and external Quality Assurance audits are a daily occurrence in the pharmaceutical industry, and the documentation and audit trails are as important as the drug products themselves. But large pharmaceutical companies have become enormous bureaucracies with thousands of people devoted to documenting and auditing the drug development process (and then documenting and auditing the auditing process). Crucially, the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) are regulatory agencies with the legal authority to close down a company that routinely fails audits or cannot produce reliable documentation.

In medicine, the stages of product development are strictly defined. In fact, for medical devices, federal law (specifically, Code of Federal Regulations Title 21) mandates that medical-device makers establish and maintain what are called "design control" procedures in order to ensure

that design requirements are met. In other words, it is a legal requirement that medical devices be "auditable."

Medical-device makers must maintain procedures to ensure that these design requirements meet the "intended use" of the device. The intended use of a "device" (or, increasingly in medicine, an algorithm - see Price 2017 for more) determines the level of design control required: i.e., a tongue depressor (a simple piece of wood) is the lowest class of risk (Class I), while a deep brain implant would be the highest (Class III). The intended use of a tongue depressor could be "to displace the tongue to facilitate examination of the surrounding organs and tissues." The intended use is what differentiates a tongue depressor from a popsicle stick. This may be important when thinking about an algorithm that can be used to recommend movies or to identify tumors; depending on its intended use, the same algorithm might have drastically different risk profiles.

In legal regimes like this one covering medical devices, some software designers may get out of those documentation requirements altogether. For example, the 21st Century Cures Act, signed into law in late 2016, amended what gets classified as a "medical device" in the first place. So whether a certain "software function" (including clinical decision support software, which is one of the hot areas for machine learning in medicine) is regulated by the FDA as a "device" turns on statutory exclusions of the Food, Drug, and Cosmetic Act (FDCA) that were added by the Cures Act (see statutory language here and the FDA's draft guidance here and here). Within the context of decision support software, the FDA's initial interpretations of the amendments have elicited strong reactions from industry stakeholders, illustrating the complexities and contested nature of such a regulatory regime in the context of ML/AI.

In any case, for products classified as medical devices, at every stage of the development process, device makers must document the *design input, design output, design review, design verification, design validation, design transfer,* and *design changes.* All of this is kept in a Design History File (DHF), which must be an accurate representation of the product and its development process. Included in the DHF is an extensive risk assessment and hazard

analysis, which must be continuously updated if any new risks are discovered. When the FDA audits a pharmaceutical or device company, it will inspect the design history file. Finally, companies are required to proactively maintain what is called "post-market surveillance" for any issues that may arise with safety of a medical device.

Security audits

Finally, in the computer security domain a successful ecosystem has evolved between tech companies and individual researchers finding security vulnerabilities. There are firms that do security audits that are distinct from internal or external vulnerability red teams. The latter often participate in what are called "bug bounties" where hackers are paid by websites to find and report bugs or security vulnerabilities. Elezari Bar On (2018) suggests modeling algorithmic auditing on the "white-hat" hacker bug bounty model adopted by the security industry, "we need a market that will facilitate a scalable, crowd-based system of auditing to uncover 'bias' and 'deceptive' bugs that will attract and galvanize a new class of white-hat hackers: algorithmic auditors. They are the immune system for the age of algorithmic decision-making." Within security research, some tools and practices that are used to test algorithms for bias may raise questions under the Computer Fraud and Abuse Act (CFAA) (Berkeley Center for Law & Technology Workshop Report, 2015).

THE LIMITS OF AUDITS

An assumption underlying the justification for auditing is the idea that auditing simply produces objective facts. This is epistemologically problematic in itself, but also when combined with any potential financial interests among the auditors. Audits are not based on an Archimedian fixed point, but rather, like machine-learning models, they are based on the views and values of humans and organizations with particular interests. As discussed above auditing does not guarantee safety and can miss systemic risks, as the example of financial crisis spectacularly demonstrated.

Indeed as Kroll (2017) argues, even for highly technical evaluation of computer code, audits are limited in their ability to attribute cause to changes in system behavior, or explain why a particular change inputs had a meaningful impact on output.

From a social perspective, Power (2000) argues that it is "important to understand the growth and circulation of an idea of audit, a growth in which accountants have been powerful agents in selling their auditing capabilities but which cannot solely be explained in this way." Who are the agents selling their algorithmic auditing capabilities today? As mentioned, both large companies and startups are beginning to offer algorithmic audits as a service. Who are the algorithmic accountants? And who should hold these auditors accountable?

For a financial firm, for example, the auditing process is known and internal financial processes are designed to be "auditable." Not only is a deep neural network not designed to be auditable, it is difficult to see how its millions of weights and nonlinear functions could be auditible. Power (1996) says: "Making things auditable is a constant and precarious project of a system of knowledge which must reproduce itself and sustain its institutional role from a diverse assemblage of routines, practices and economics constraints." And even with access to "the guts of a system (the code, the architecture)", the policies and procedures that govern its development, and information about the use-environments and contexts, the values implications of design decisions may remain obscured (Mulligen & Bamberger, 2018). Ultimately, Power has argued audits risk becoming "performative rituals of verification," which might temporarily assuage our fears about inscrutable and uncontrollable processes or algorithms, but they are hardly cold, objective, independent producers of "facts". Therefore how we define what is to be audited becomes extremely important. Related to the fact that audit results too must be interpreted, Karl Popper pointed out, "Every observation and, to an even higher degree, every observation statement, is itself already an *interpretation in the light of our theories."*

Styhre (2015) describes the inherent embeddedness of audits and auditors within the very systems that should be audited, "The audit per se and its procedures and routines are never

isolated from the practices of the professionals conducting the work, and therefore the credibility of the professionals remains a key issue...the distinction between inside and outside, externality and interiority being of key importance for the legitimacy of the audit work is porous and fluid." There is no objective gaze from the audit. Financial auditing, for example, is fundamentally an *inferential* process which has to draw conclusions from a limited set of documents, budgets, oral testimony or direct observation (Power,2000). What might a similar sort of inferential auditing process look like for algorithms? Machine learning is also *inferential* and draws conclusions from limited datasets.

RECOMMENDATIONS

As calls for the regulation of tech grow louder and more pointed, there is an opportunity for a closer collaboration among stakeholders in our increasingly information-based and algorithmic society. Well-thought through and voluntary algorithmic auditing practices could be used to demonstrate the ethical foresight necessary to decrease the probability of what Mittelstadt (2015) terms "regulatory whiplash", where "overly restrictive measures (especially legislation and policies) are proposed in reaction to perceived harms, which overreact in order to re-establish the primacy of threatened values."

(1) **An ethical algorithmic audit could address questions of fairness and accountability at each stage of product development** outlined above. This would require documentation to be generated making the product development process "auditable" such as audit trails and audit logs.The question of interiority versus exteriority is again important to consider, as external auditors in safety-critical industries are given access to internal documentation, design plans, and risk assessments.

(2) **Further research on what would be required to make algorithms and their development "auditable".** As discussed, neither algorithms nor the development process within most companies are designed to be auditable. Audit trails or audit logs are not generated

in standardized formats, nor are they used consistently for purposes such as determining fairness.

(3) There is an opportunity to create a **more meaningful dialogue among a wider array of stakeholders** in algorithmic accountablity, e.g., vulnerable communities impacted by technology, nonprofits on the frontlines of fighting discrimination and other forms of injustice, legislators and regulators, politicians, and the tech industry as a whole. Conferences on algorithmic accountability such as FAT* are a good start, but are still not well-known outside highly specialized academic disciplines. A broader public-oriented effort should bring all stakeholders to the table.

(4) **Do not make algorithmic audits as opaque as the algorithms they're meant to audit.** In the other industries discussed the process of auditing has become a professionalized endeavor, with highly specialized auditors who produce lengthy audit reports. The validity of the audit rests largely on the credibility of those professionals conducting the audit. Thus there is a risk for algorithmic auditing that, rather than revealing issues and improving accountability, algorithmic audit reports become interpretable to only a small group of highly trained experts - which is exactly one of the problems currently with algorithms that algorithmic audits are meant to address.

(5) In view of the complexity of both the algorithms themselves and the processes through which they are built (hundreds of engineers working across geographies), further research **should develop hybrid auditing techniques** that allow both internal and external stakeholders to trust their values are protected. These techniques should evaluate both the organizational processes and work practices, as well as the technical functioning of the systems. This could include everything from data generation and collection and model selection, but also quality assurance checks as the project advances.

(6) What would a standard for **good algorithmic development practice look like?** In accounting and finance, medicine, and other industries there are clear standards such as Generally Accepted Accounting Practices (GAAP) and Good Clinical Practice (GCP) for clinical

trials or Good Manufacturing Practice (GMP) for drug makers. Internal and external auditors then use these guidelines to find violations or to make recommendations to the organizations being audited.

ACKNOWLEDGMENTS

We would like to thank the panelists on Panel 4: Auditing algorithms (from within and from without): Chuck Howell (MITRE), Danie Theron (Google), Michael Tschantz (International Computer Science Institute) and the moderator Allison Woodruff (Google). This blog is based on discussions from the panel, and on other input from many scholars and experts from industry who attended the workshop. We would also like to thank Daniel Kluttz, Jenna Burrell, Allison Woodruff, Deirdre Mulligan and Jen Gennai for helpful feedback and suggestions on prior drafts of this report. Finally, we acknowledge the intellectual support and organization from the AFOG group at UC Berkeley.

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