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