

ACCOUNTABLE ALGORITHMS

Joshua A. Kroll

Postdoctoral Scholar, UC Berkeley School of Information

Respectfully submitted in response to Question 9 of the FTC's Docket #FTC-2018-0056 on "Competition and Consumer Protection in the 21st Century", regarding the consumer welfare implications of artificial intelligence and automated decision making.

(Abstracted from Joshua A. Kroll , Joanna Huey , Solon Barocas , Edward W. Felten , Joel R. Reidenberg , David G. Robinson & Harlan Yu Accountable Algorithms, 165 U. Pa. L. Rev. 633 (2017). Available at: http://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3)

Many important decisions that were historically made by people are now made by computer systems: votes are counted; voter rolls are purged; loan and credit card applications are approved; welfare and financial aid decisions are made; taxpayers are chosen for audits; citizens or neighborhoods are targeted for police scrutiny; air travelers are selected for search; and visas are granted or denied. The efficiency and accuracy of automated decision-making ensures that its domain will continue to expand. Even mundane activities now involve complex computerized decisions: everything from cars to home appliances now regularly executes computer code as part of its normal operation.

However, the accountability mechanisms and legal standards that govern decision processes have not kept pace with technology. The tools currently available to policymakers, legislators, and courts were developed primarily to oversee human decision makers. Many observers have argued that current frameworks are not well-adapted for situations in which a potentially incorrect, unjustified, or unfair outcome emerges from a computer. Citizens, and society as a whole, have an interest in making these processes more accountable. If these new inventions are to be made governable, this gap must be bridged.

My work describes how authorities can demonstrate—and how the public at large and oversight bodies can verify—that automated decisions comply with key standards and policy goals. I describe two categories of solution to this problem: *ex ante* approaches aiming to establish that a decision process works as expected (which are commonly studied by computer scientists); and *ex post* approaches which are used after decisions have been made, such as review and oversight methods (which are commonly used in existing governance structures). I show how the tools of *ex ante* analysis can be used to guarantee that *ex post* analysis can operate and be useful and effective.

I challenge the dominant position in the legal literature and in policy discussions that transparency can address these problems. While often vaunted as at least ameliorating concerns with computer systems, transparency alone can only be effective when it is total, covering a system's full source code, data, and operating environment. Yet, disclosure of the source code and data underlying a computer system is often neither necessary (because alternative technical approaches provide better evidence) nor sufficient (because of issues analyzing disclosures). Furthermore, full transparency can be undesirable, such as when it discloses private information or enables strategic gaming of systems (e.g., by tax cheats minimizing their likelihood of audit or terrorists or smugglers minimizing their chance of search).

Instead, I argue that technology is creating new opportunities—subtler and more flexible than total transparency—to create evidence that automated systems align with legal and policy

objectives. Systems must be designed with such alignment in mind. Doing so will improve not only the current governance of automated decision-making, but has the potential to improve—in certain cases—the governance of decision-making in general. The implicit and explicit biases of human decision makers can be difficult to find and counteract, but we can investigate how an algorithm arrives at a decision, declaring the decision-making process, purpose specifications, and data governance and privacy policies ahead of time and verifying their veracity afterward. That is, computer systems are accountable when they are designed to allow introspection of their operation and this can be verified by a skeptical public.

Transparency has its place, but only when the disclosures that are made further the goal of convincing outsiders that they system is operating correctly, however that may be defined in a particular application. Instead of full transparency, I advocate for partial transparency combined with technical evidence that disclosures relate to and account for actual decisions and the applied policy of a decision maker. This sort of disclosure enables the analysis of compliance of automated decision-making systems with key governance requirements including *procedural regularity*, the property that all decisions are made according to an announced set of rules consistently applied in each case, as well as more complex requirements. For example, a data subject may want to know that a system is operating *correctly* (e.g., that a lottery really chooses fairly from among the submitted entries) or consistent with some principle of *fairness* (e.g., that a credit scoring system doesn't use irrelevant or legally proscribed data such as race or gender).

Machine learning and artificial intelligence present further complications for accountability; with these technologies, rather than having a decision policy written and designed by a human programmer, the policy is discovered using a learning algorithm and (sometimes vast quantities of) data that together form a trained model. Often, these models do not lend themselves to easy interpretation and so it is not possible to read back what decision policy is in force. As a result, the usual process of setting requirements and validating compliance of the code with those requirements becomes significantly more complex. Fortunately, the framework described above still applies: new technologies from computer science allow the training of models which are known to have desirable technical properties that can speak to fairness as conceived in the law, and we can design systems to make use of these new methods while also producing sufficient evidence to allow the fairness of the design to be disclosed. This can be accomplished without requiring the disclosure of the model, which may be a trade secret, or the underlying data, which may be personal and private. However, it is still necessary to have a conversation in the open about what makes a system fair or unfair and how compliance will be established.

For example, it is common in many U.S. jurisdictions for people arrested for a crime to be assessed for their risk of recidivism upon release using one of a few common models trained on the historical crime data of populations of convicted offenders from around the country. These risk assessment models have in some cases been shown to be subject to different levels of error for racial minorities. Unfortunately, because at least some of these models are produced commercially, their exact operation is a trade secret and the specific cause of the differences in errors (as well as whether any individual defendant is negatively affected by shortcomings in the model) is hard to investigate. This has led to public arguments between journalists, who argue that minority should not be subject to high risk scores when they will ultimately not re-offend, and the model's developers, who argue that an equal fraction of minority and majority defendants who are given a certain score will re-offend and therefore the model is not biased. In a sense, both sides are correct: the model itself acts neutrally in the face of a world which is

unfair to minority defendants (and made more unfair by the use of the model against those defendants). And statisticians have shown that it is mathematically impossible to satisfy both of these concerns at once. Of course, that is not to say that the unfairness of the world should not be corrected, nor to say that the model should not be designed to provide equity in outcomes. Indeed, the extent to which equity for minority defendants should be a design constraint of the model is a policy question that should be resolved publically. The public then has an interest in receiving sufficient disclosures about the operation of the model to understand whether the chosen policy is properly implemented by the model in practice.

I conclude by noting that the way forward on the governance of automated decision-making systems is through increased collaboration between computer scientists and policymakers. Computer Scientists have or can develop the tools to implement a wide variety of policies and have proved adept at investigating real-world fairness concerns and discovering bad or unusual behavior in large and complex systems such as ad networks, search engines, and social network post ranking systems. However, computer scientists and technologists are generally in a poor position to determine what policy should be in effect, especially *ex ante*. Policymakers, on the other hand, are well positioned to run governance processes to determine what rules should be in force and to flag important systems of interest for extra scrutiny and review. And to the extent that policies cannot be specified ahead of time, policymakers can structure oversight and review processes that allow individual cases to be considered *ex post*. However, policymakers operate processes that are inherently slower than the development of technology, and so will always need input from computer scientists as to how to mediate policy with technology. More concretely, computer science and programming reward specificity, while the policy process requires ambiguity to function. Only by collaborating can technology and policymaking professionals create truly accountable algorithms