Algorithmic Accountability

by Alex Rosenblat, Tamara Kneese, and danah boyd

A workshop primer produced for:
The Social, Cultural & Ethical Dimensions of "Big Data"
March 17, 2014 - New York, NY
http://www.datasociety.net/initiatives/2014-0317/

Brief Description

Accountability is fundamentally about checks and balances to power. In theory, both government and corporations are kept accountable through social, economic, and political mechanisms. Journalism and public advocates serve as an additional tool to hold powerful institutions and individuals accountable. But in a world of data and algorithms, accountability is often murky. Beyond questions about whether the market is sufficient or governmental regulation is necessary, how should algorithms be held accountable? For example what is the role of the fourth estate in holding data-oriented practices accountable?

Detailed Topic Description:

Algorithms can be hugely beneficial in sorting through vast troves of information to deliver what is potentially the most useful sort. Automated algorithms can use a sequence of well-defined steps and instructions to generate categories for filtering information based on a combination of motives about a desirable outcome. In the final expression of that combination, the elements of uncertainty, subjective interpretation, arbitrary choice, accidents, and other ingredients in the mix are rendered invisible, and what is displayed to the end-user who interfaces with the algorithm's product is just the functionality of the technology. For instance, Google, Yahoo, and other search engines can effectively create "filter bubbles" for the results people see when they query items, which can be problematic. Some information is more visible to one individual versus another based on the user profiles that the search engine has on them, and how its algorithms predict what might be most relevant to the users according to their profiles. How might algorithms affect the flow of educational materials or other types of information? Who or which networks of stakeholders are the arbiters of algorithmic power that strongly influence information flows?

Designing software to mobilize and unlock the supposed power of the "big data" phenomenon is often focused on the best technical ways to achieve a particular outcome, like personalizing search results so that a user gets information that is tailored to their interests. Algorithms break down information into certain constituent parts, and

reconfigure it into a new production of information to fulfill particular goals. This can have huge societal benefits. For example, a Microsoft Research team has come up with an algorithm to help medical researchers sort through data on 120,000 individuals in a few hours, in contrast with current algorithms that cannot make computations on such large datasets. This algorithm has the potential to identify a patient's risk for diseases, and even which drugs might be best suited to them.

Can any problem be resolved using an algorithm? The social, cultural, and political impact of an algorithmic solution has consequences that play out far beyond the technical innovation behind the restructuring of information. Does an algorithmic-orientation to solving technical problems mitigate or downplay the social, political, or ethical issues at stake? For example, one study found that Google's AdSense algorithm, which automates targeted advertising to serve users the adverts that are most relevant to them, is more likely to suggest possible arrest records for racially associated names that are being queried, like Trevon Jones, than for Caucasian names like Geoffrey. How can this association affect someone's job prospects, or their application to rent an apartment? Negative adverts linked to a person's name are likely to get more clicks than neutral or positive adverts. Higher click-rates increase the value of those adverts, which makes them appear more often, thereby reproducing the prejudicial impact. The discrimination produced in the search results is unintentional, but do companies have an obligation to correct for prejudice? If an algorithmic solution is being presented, it is useful to examine how the problem is being framed? Is it possible to build an algorithm that is either bias-free, or has corrective measures built in for explicit biases? What kind of mechanisms should exist for evaluating discriminatory or prejudicial outcomes, and what criteria would they use? Are there methods other than reverse engineering for evaluating allegedly prejudicial outcomes?

Part of the difficulty in determining algorithmic accountability for a wide range of issues, including discrimination, is confusion about how to look at an algorithm. The logic of an algorithm is not immediately visible, nor would that logic be available if one had the source code. Many algorithms are too complex for any one developer to understand the mechanisms at play. Learning algorithms, such as those that underpin everything from recommendation engines to filters, rely on particular data sets and allocate different weights to each variable. There are ingrained biases in selecting data sets, variables, weighting, and test cases. Different kinds of algorithms - like associative, regressive, or sequence analysis algorithms - perform different functions, and some might be easier to 'correct' than others, particularly because no one, not even the designers, are quite certain what tweaks will create the desirable result. Do the biases become obvious in the consistency of results generated by algorithmic logic? For example, the Staples website used an algorithm that generated different discounts on prices for the same products to people based on their location data; in effect, people in higher-income areas received higher discounts than people in lower-income area. While that is not illegal, users may not wish to reveal personal information if they do not trust commercial enterprises to treat them fairly. Can algorithms exacerbate existing

disparities between socio-economic classes by using proxies, like geo-location data, to enable subtle price discrimination? What would best practices in this area look like? Can companies build trust with their customers by participating in an accountability program, thus creating a competitive advantage in their marketplace?

The complexity of algorithms - and the limited computational literacy of the public - pose barriers to making networks of algorithmic power relations readily transparent. Does software code, promulgated as the technical salve to many of the world's ills, have the potential to legitimately compete with more democratically-enacted forms of power, like legal codes? More specifically, can the law keep up with technological innovation in a way that subjects complicated algorithmic systems to the usual process of checks-and-balances that is generally imposed on powerful items that affect society on a large scale? Do we examine the institutions that use algorithms to make decisions, or the designers, or multiple stakeholders? Does crediting or blaming an abstract 'algorithm' for some particular outcome become a way for institutions or organizations that use the algorithm to avoid being held accountable for bad or unethical practices? In this type of scenario, how is the individual disempowered or empowered to address institutions that cite algorithmic powers or rationales for the individual's experience?

One way of understanding the practice of journalism, and of the fourth estate generally, is as a mechanism for holding institutions of power - including governments and corporations - accountable by making visible problematic practices, corruption, and abuses. But what does it mean for the fourth estate to hold an algorithm accountable? What kinds of tools do journalists have to engage with the impact of algorithms? Do they need technical expertise? Journalists and academics can examine the social impact of an algorithmic effect when those effects are evidently discriminatory or negative, but how else can they investigate precisely what an algorithm is doing, or who is affected by it? Who might be a credible inspector? What kinds of transparencies would make these issues easier to address? Who, besides journalists, should play an important role in holding algorithms accountable? What mechanisms need to be put into place for that to occur?

Determining algorithmic accountability has real consequences for understanding and regulating who or which entities control flows of information in public and private spheres. What would governing, policing, or even designing ethical algorithms look like? Algorithms operate within the contexts of specific datasets, interfaces, use situations, business models, cultural expectations, etc. By focusing uniquely on algorithms as the source of both pros and cons in a data-driven world, it's possible that algorithms are being fetishized. Algorithms are also designed to learn and shift, and to be tweaked. How do we hold a moving target accountable? Is a conversation about algorithmic accountability more usefully extrapolated to a conversation about holding institutions accountable for the outcomes of their methods, regardless of what the methods are? Should different sectors use different approaches to accountability, based on the different types of trust that the public has in those entities? What kind of

watchdog system would be useful for diagnosing any ills or breaches of trust within those organizations? What happens when institutions are not the relevant actors?

Case Study 1: Predictive Policing

In Philadelphia and Baltimore, police adopted a software designed to predict which parolees were more likely to commit murder, to help inform them of the supervisory level that should be allocated to each released prisoner. The predictive variables used by software designers create focal points for policing authorities to use that renders some groupings of people, or crimes, more visible than others and thus more suspicious, which can undermine due process. For instance, John Doe was young when he committed a simple drug possession crime. The algorithm that predicts whether John Doe is likely to commit a violent offense in future has nothing to do with whether or not his offense was violent; his age, gender, and the time between his last and any later offenses are the predictive variables that matter.

Is it problematic that John Doe the Parolee is subject to police supervision on par with someone who has already committed a violent offense? In one sense, the predictive variables embedded in algorithms are just a more refined version of how decisions about whom to police, and with which levels of attention, are already made, but it is important to consider how transparent these decisions are, or can be, when 'the data' is being cited as a rationale. Do such mechanisms undermine due process? Will the data be used as 'proof' in ways that reinforce such surveillance? How does an algorithm subtly shift the way that individuals are evaluated to have paid their dues to society for crimes they committed?

In New York City, longstanding concerns have surrounded the "stop-and-frisk" program. Although individual law enforcement officers consistently reported that they do not target black and Latino men, these groups of people are stopped at unrepresentative rates in New York. When journalists used data to reveal the prejudicial nature of the program, heated debates ensued. In many ways, this was made possible because journalists could access and interpret the relevant data. But how can a journalist challenge the authority of an algorithm? While the ultimate decision-maker may be a human, the technology that produces information to inform the decision-maker is often challenging to investigate.

Case Study 2: Visibility/Invisibility of Online Content

In Germany, Google is required to tweak its algorithm to remove autocomplete suggestions that are defamatory or libelous, like auto-completing a search engine query of the name "George Wilson" with "is a White Supremacist." Bettina Wulff, a publicist and the wife of former German President Christian Wulff, sued Google for algorithmically generating search results on queries to her name indicating that she once worked as a prostitute. People who share the same name as someone who does have

infamous political leanings, a history of criminal activity, or a particular risk of disease are vulnerable to negative associations that are calculated by algorithms.

Is an algorithm responsible for the potentially harmful inferences it imputes to people? What if a specific George Wilson is, in fact, a White Supremacist, but another person shares his name, and is vulnerable to the associations with the more nefarious George? How do you appeal that sort of algorithmically generated association to reduce false positives?

While lawsuits may force companies to alter aspects of their algorithm, the storm of media coverage increases the visibility of the issue, and the negative associations inherent to that issue. For example, Bettina Wulff's Wikipedia entry now includes her fight against rumors that she worked as a prostitute, and news of Wulff suing Google is among the top articles that appear on Google's search. This phenomenon of making things more visible by asking them to be less visible is often referred to as the Streisand Effect, referring to what happened when Barbra Streisand attempted to suppress photographs of her house, generating further publicity for the photos. What does it mean to hold systems accountable when the act of journalism makes the issue more visible? Is there a more private route that we can envision for reporting and resolving problematic issues on the data that is available on us?

Just as algorithms - and the reporting of them - can make things more visible, they can also make things less visible. News items from different sources appear in the Facebook newsfeed, and Facebook has become a main homepage for news items for a lot of its vast amount of users. According to Facebook, the goal of Facebook's newsfeed is "to deliver the right content to the right people at the right time." Because of its position in the ecosystem, Facebook can influence the likelihood that traffic will go to different websites in positive or negative ways. When Facebook tweaked its newsfeed algorithm to reward content that it deemed to be of higher quality, the company fundamentally altered the flow of traffic from its site to other sites. Needless to say, users seeking to get their content in front of as many people as possible had long gamed Facebook's system. But what does it mean that Facebook - or any company, for that matter - can easily alter the flow of traffic by altering its algorithms? What kind of power do these companies have over other companies that rely on being linked to, including journalistic enterprises?

Case Study 3: Discriminatory Black Box

Insurance providers, like many other companies, are not allowed to discriminate on the basis of protected classes. In other words, people cannot be denied insurance because they are black or Muslim. And yet, sometimes insurance is fundamentally a mechanism of discrimination. Insurers try to minimize risks and maximize profits. Marginalized populations, including protected classes of people, are often more risky to insure, in part because of how discrimination has historically made it harder for people

in these groups to get access to high quality medical care, favorable mortgages in low-risk communities, and educational opportunities. Thus, if insurers want to minimize their risks, they would often benefit by not covering many marginalized populations.

As insurance determinations are increasingly computed algorithmically, it is more difficult to determine whether or not a person is being discriminated against inappropriately. The designer of an algorithm may have no intentions of producing discriminatory results. For example, algorithmically inferring race with a high degree of accuracy without actually knowing race is relatively easy. Unless an analyst is testing to make sure that race is not a factor, the correlates that enable such discrimination to occur can often go unnoticed.

Who is responsible for holding insurance providers responsible for not discriminating? Does this require assessing inferences made algorithmically? Does it require offering test cases to make certain that inappropriate discrimination is not accidentally taking place? Must we simply assess whether or not discrimination is occurring by looking at the outcomes? What is the appropriate way to hold such institutions accountable?

Questions to Consider

- What are the major social, cultural, and ethical tensions that emerge when thinking about algorithmic accountability? What needs to be better understood to address what's happening?
- What conflicting values and tradeoffs are at stake? How do we understand relevant actors, stakeholders, and "camps"?
- How are the opportunities and challenges of algorithmic accountability different in different domains (e.g., criminal justice vs. healthcare vs. marketing)?
- What are additional salient case studies that highlight the tensions, tradeoffs, and issues?
- Who should be holding algorithms accountable? What is the role of the government? Of data providers? Of technologies and tools? Of educational institutions? Of media institutions?
- Who should serve as a data caretaker? What is the role of the government in supporting, regulating, protecting data caretakers?
- Who can challenge algorithmic systems, and what kinds of expertise might they need to do so? What is the role of the fourth estate?
- Do algorithms affect the flow of information in new ways, and who is affected by them?
- How can the need for transparencies be balanced with the proprietary nature of some algorithms?