DataSF Research Brief:

Ethics and Accountability in Algorithms

Author: Charlie Moffett

Table of Contents

1. Overall Summary

Scholars have called for critical approaches of the "mythology" of so-called Big Data - this 'widespread belief that large data sets offer a higher form of intelligence and knowledge'[1]. These beliefs stem from two purported characteristics of algorithms: access to more information (computed faster and more reliably), and value-neutrality[2]. What has become abundantly clear in recent years however is that algorithms are in fact not objective.

For this project, we investigated the issues of accountability, transparency, and fairness as they relate to algorithms. In this document, we review for each of these:

- What it is and why we care
- Summary of the research, including recommendations.
- Existing frameworks and major issues to be considered

This is a major topic of scholarly interest, being pursued by hundreds (probably thousands) of academics and researchers across multiple fields[3]. Government has two primary roles to play - regulating the algorithmic space, and employing them responsibly.

Below are some high level findings across all three areas.

Black boxes create cause for concern. The opacity and inexplicability of algorithms is a recurring critique throughout the literature. Algorithms will continue to secure influence over more parts of our lives, and in more impactful ways. But they are largely invisible to the average user, and promise very serious consequences if the surrounding issues are not addressed upfront. ProPublica's story on "machine bias" in an algorithm used for sentencing defendants is a posterchild example. We also see these issues spring up in controversies that are closer to our daily lives, like with Facebook's news feed, or Google's search results, or Twitter's trending topics. Most importantly for us, however, is the notion that black box algorithms pose a threat to the legitimacy of governance. Left unchecked, they diminish the public's ability to participate in and challenge governance-related decisions, which contradicts salient normative principles of liberal democratic governance.

Your data might be used against you in ways that you wouldn't expect. As Rutgers Law Professor Ellen Goodman puts it, privacy is like a weed. It comes up at every turn, within every sub-issue. Seemingly innocuous datasets serve as a digital footprint for individuals when combined. Depending on the domain, algorithms can be tuned to be more or less stringent with regard to privacy. But although privacy weaves like a common thread throughout the various areas we touch on here, it is "to the side a bit" and falls outside of the scope of the relevant literature and consequently outside of this research summary. Fairness and transparency, which we do explore in the following pages, will prove to be "much messier affair[s] than privacy"[4].

---

[1] boyd d and Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication, & Society 15(5): 662–679
[2] Daston L and Galison P (2007) Objectivity. New York, NY: Zone Books
[3] Algocracy as Hypernudging: A New Way to Understand the Threat of Algocracy, by John Dahaner, Institute for Ethics and Emerging Technogies, 1/17/2017
[4] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic

What are we comparing automated decision-making to? Recent developments in "Big Data" represent a transition from collective details to individual details, and from prediction to decision-making. Says leading thinker Nicholas Diakopoulos, "government agencies seem unprepared for inquiries about algorithms and their uses in decisions that significantly affect both individuals and the public at large." As our traditional notions of morality collide head-on with algorithms, the challenge becomes instilling societal norms into these new sorts of systems. The traditional way of addressing this concern was with human oversight (watchdog groups, regulators, the law, etc.), but the scale at which algorithms have come into play simply cannot be dealt with by human organizations alone.

2. Accountability

2.1 What and Why

Accountable: "subject to having to report, explain or justify; being answerable, responsible."

More and more jurisdictions are using algorithms to direct resources and investments and to deliver services. As algorithms replace human decision-making, the link between decision-making and those responsible becomes fuzzy. Given the real and well-documented consequences of such systems - already rampant in critical domains such as criminal justice, workforce development, education and finance - it will be paramount to identify and hold responsible the right actors in a logical and just manner. "The algorithm did it" is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes .

2.2 Research Summary

Literature and Practices

Algorithmic accountability has recently become a hot topic for professors and private sector parties alike. Though a host of resources have been helpful in teasing out the leading thinking on this issue, there are two frameworks in particular that distill the key components in a pragmatic manner and bring to surface the major considerations for jurisdictions. "Principles for Accountable Algorithms" was developed as part of a Dagstuhl seminar held last summer at the Leibniz Center for Informatics. "Principles for Algorithmic Transparency and Accountability" derives from the Association for Computing Machinery's Code of Ethics and was formalized in a statement from the ACM US Public Policy Council earlier this year.

The jury is out about who to blame when things go wrong. Part of the motivation behind this project is to avoid mistakes that lead to media catastrophes and potentially set back the progress in made by data-driven decision-making movement (O. Wise, personal communication, July 17, 2017). Though "technology advances are most effective when they are accompanied by increased clarity about responsibility for failures and liability for damages," disagreement is widespread about how to ascertain liability in the planning process (Schneiderman, 2016). Existing software

---

and contextual bandits. In Advances in Neural Information Processing Systems, 2016.

contracts typically protect developers from legal repercussion through “hold harmless” provisions.

Government should have stronger accountability standards than what the private sector has demonstrated to date. Ultimately what you want in this burgeoning domain is for city government to remain accountable to its citizens, and for city government to hold vendors accountable along those same lines. Current practice by private sector vendors is deny access to any information about the algorithm in defense of “trade secrecy”, with a few promising exceptions (Wexler, 2017). Even requests for such details facilitated by the Freedom of Information Act (FOIA) are regularly shut down (Goodman, 2017). When it comes to the important decisions about these systems and analyzing their performance, we have to bring the right stakeholders to the table. Affected groups should have representation, meaning diversity across gender, race, technical expertise, etc. Guiding principles are a helpful first step, but creating institutional structures to consistently advise on the ethic issues will be critical. Luckily we aren’t starting from scratch in this endeavor - universities have been doing this for awhile with IRBs and the like.

Cities that are doing contracting have a lot of leverage they aren’t using. In order to be proactive about accountability, cities shouldn’t enter into overbearing contracts with vendors that restrict information sharing regarding their products. If the City (in the analog world) decided to close certain schools, it would ultimately come up with some rationale that would be subject to public scrutiny, and we would be able to know if mistakes had been made and decide whether we agree with its normative traces. Algorithms are already starting to govern in cases like these, but without those deliberative processes.

Recourse. Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions. Develop a plan for what to do if the project has unintended consequences. This may be part of a maintenance plan and should involve post-launch monitoring plans. Some have suggested appointing an ombudsman whose task it would be to communicate about and field inquiries into government automation[5]. It should be easy and obvious for wronged parties to find retribution for any harm caused to them by a decision made entirely or in part by an algorithm.

Proactive reporting about key metrics can be a sure path toward accountability. Many have focused on retrospective checks to ensure that the algorithm is performing as promised. The burden instead should fall more squarely on the vendor or department who created the algorithm. These parties will be in a better position to reveal how the algorithm works (even if only partially). We explore the practice of disclosure more closely in the following section.

References

Below is a list of sources used.

- E. Goodman, personal communication, July 24, 2017
- R. Shroff, personal communication, July 19, 2017
- Ravi Shroff. Predictive Analytics for City Agencies: Lessons from Children's Services. Big

[5] http://www.cima.ned.org/blog/algorithm-ombudsman/

Data, 2017. (PDF)
- Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (link)
- ACM US Public Policy Council. Statement on Algorithmic Transparency and Accountability. 2017. (PDF)
- Diakopoulos, Nicholas. 2016., "Accountability in Algorithmic Decision Making" Communications of the ACM, February (59) 2: 56-62. DOI: 10.1145/2844110 (PDF)
- Shneiderman, B. (2016). Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. Proceedings of the National Academy of Sciences, 113(48), 13538-13540. (link)
- Code of Silence, by Rebecca Wexler, Washington Monthly, 6/13/2017 (link)

3. Transparency

3.1 What and Why

Transparent: "characterized by visibility or accessibility of information especially concerning business practices"

A general characteristic of algorithms is their obscurity. Transparency, and in particular explainability, helps us to understand the limits of an algorithm and the boundaries within which it should be trusted. The private sector has recently become quite active in addressing this topic. Apple recently joined Amazon, Google, Facebook and other leading technology companies in a 'Partnership on AI' research project. The Defense Advanced Research Projects Agency (DARPA) has begun work on a 5-year program to develop "explainable artificial intelligence (XAI)". Big time players are recognizing that as algorithms continue to advance and exert broader impact, it is important that AI evolves with trust and transparency in mind.

3.2 Research Summary

Literature and Practices

Auditability

In order to meet the definition of auditable, algorithms (as well as associated data and decisions) need to be comprehensively recorded so that third parties can probe, understand, and review its behavior. Both the ACM and FAT/ML frameworks specify this concept as a key principle, and "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms" from Sandvig et. al. provides detailed insight into the logistics and nuisances around this best practice.

Keep a detailed record of decisions and assets (models, algorithms, and data). By establishing your own, internal auditing processes, you're inviting others to monitor, checking, and criticize your algorithms. Designing for auditability also provides a mechanism for double-checking your work and forcing yourself to be explicit about decisions, increasing understandability and replicability. For example, internal auditing mechanisms help you maintain important details about which factors may be responsible if outcomes become problematic. Keeping snapshots of your algorithm throughout its development (versioning) is an important component of this

process. My favorite resource for specifics about avenues for transparency about data and models is Nick Diapolous' "Algorithmic Transparency Standard" from his articled *Accountability in Algorithmic Decision Making*.

Confidentiality is a major obstacle to transparency. In situations where the private sector is involved (typically as a vendor), algorithms often remain opaque to the interested public because their details are protected as 'trade secrets' out of fear of competition and abuse. By far, the most popular example comes from criminal justice - risk-assessment software providers invoke a "trade secret evidentiary privilege" to withhold its source code, and courts refuse to order that it be disclosed[6]. This is not a problem with the algorithm; it's an institutional/legal problem. Audit trails could help mitigate this issue, but guidelines need to be developed and implemented in regulation for when government use of an algorithm should trigger an audit trail.[7]

Design terms of service that welcome audits of your practices. Most websites today post terms of service that prohibit the types of activities needed robust audit testing[8]. Best practice dictates that these terms allow the research community to perform automated public audits. If your algorithm makes use of sensitive information, trustworthy third parties can be designated from the outset to conduct the audits. As touched on with regard to accountability, it's also important to prepare communication with additional outside parties that may want to run audit tests on your algorithm. Adopting an attitude of transparency nurtures ethical behavior around algorithms in the community as a whole.

Not everyone agrees about how much you need to reveal. Many argue for complete source-code transparency, while others believe only certain pieces of information (like benchmark datasets and aggregate results) are necessary. With BlightSTAT in New Orleans, for example, both the source code for the algorithm and the datasets used to train the tool are published online, though results from the program could have been strengthened by something akin to a pre-deployment equity analysis (see 'Fairness' section for more) (O. Wise, personal communication, July 17, 2017). Even in instances where everything about the algorithm is known, some will feel that the algorithm is so complex that they cannot understand it.

Explainability

An algorithm's decision must be explainable to anyone who it affects. Explanation has long been deemed a crucial aspect of transparency. It helps us to reduce risks of error, abuse, and arbitrariness. Decision-making processes employing algorithms and similar data-driven approaches complicate this equation. Algorithmic approaches promise to refine and improve accuracy and efficiency, but the logic and rationale behind each decision remains opaque to human understanding. This is particularly important in public policy contexts.

Procedures and decisions need be accompanied by explanations. Users, and more importantly subjects, of the decisions made or influenced by an algorithm should have access to non-technical terms describing the product in place. Both the procedures followed by the algorithm as well as

[6] https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

[7] http://www.slate.com/articles/technology/future_tense/2016/02/how_to_hold_governments_accountable_for_their_algorithms.html

[8] http://www.mlandthelaw.org/papers/goodman1.pdf

the specific decisions that are made require explaining. A truly interpretable algorithm allows stakeholders to understand the outcomes, not merely the process by which outcomes were produced (E. Goodman, personal communication, July 24, 2017). A number of useful methodological recommendations have already been set forth to address explainability. Data visualizations might also be leveraged to succinctly communicate the workings of an algorithm[9]. More fundamentally, we should question the use of an algorithmic process if its effects are not meaningful or if it cannot be explained.

Input variables and their respective weights should be disclosed. The variables used in an algorithm, how they are defined and datafied, and their weighting are essential design decisions that deserve careful consideration and scrutiny. For example, Oregon discloses the 16 variables and their weights in a criminal justice model used there to predict recidivism[10]. Other players in this space, like the for-profit company Northpointe, have been widely criticized for protecting how inputs are weighted.

Users should be made aware when and why algorithms are being employed. At a high level, transparency around human involvement might involve explaining the goal, purpose, and intent of the algorithm.Jurisdictions should disclose if and when an algorithm is being employed at all, as well as the degree to which human agency is being exercised in such situations. Public policy and law would also say that the public needs to understand (in broad brush) what the politics are.

Efficacy

Once the goals are clear as to what the algorithm was designed to accomplish, its effectiveness can be assessed. Effectiveness relates to how accurately you find answers to the question(s) proposed. Pollsters for example disclose a margin of error so that consumers can have an understanding of how much precision they can reasonably expect. These issues extend to the realm of algorithms.

Document and make available the methods and results of validation efforts. Out of all the ink spilled on these topics in recent years, very few are talking about transparency around algorithmic efficacy (A. Nicklin, personal communication, July 25, 2017). Certain key statistics can alleviate questions about the accuracy of the algorithm. These include details like margin of error, accuracy rates, and measures of uncertainty. Expected and worst case implications can inform mitigation procedures and public communications. And again, make sure contracts with vendors have clauses for transparency - vendors needs to be on the hook the same way the government is with regard to disclosing these key statistics.

Make clear your data quality confidence and which shortcomings may be present. Algorithms designed to recognize patterns in big data often produce a confidence value that can serve as a measure of uncertainty in the outcomes. As algorithms are employed more broadly, the data that feed them gain a premium. Gaps and errors run the risk of being amplified and accordingly require detailed inspection. A description of the way in which the training data was collected, for example, should be maintained, along with information about the data collection

[9] https://bost.ocks.org/mike/algorithms/

[10] https://apnews.com/1efa2f2f5f004295a35d175f58149e67/we-need-know-algorithms-government-uses-make-important

process. Some disclosure could be made about whether the data was private or public, and if it incorporated dimensions that if disclosed would have personal privacy implications. The most important step you can take to improve and make clear your data quality confidence is to treat your data like a strategic asset[11].

References

Additional readings considered.

- Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. (link)
- C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry, 2014. (PDF)
- http://theconversation.com/we-need-to-know-the-algorithms-the-government-uses-to-make-important-decisions-about-us-57869
- https://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/
- https://freedom-to-tinker.com/2017/05/31/what-does-it-mean-to-ask-for-an-explainable-algorithm/

4. Fairness

4.1 What and Why

Fairness: “impartial and just treatment or behavior without favoritism or discrimination”

The algorithm should never preferentially favor (i.e. choose with higher probability) a less qualified individual over a more qualified individual. Proactive measures should be taken in order to recognize and minimize bias. Responsible practice in this area directly takes into account provisions for fairness and anti-discrimination.

4.2 Research Summary

Literature and Practices

The data fed into an algorithm is a common source of unfairness. This is part of the reason why it is important to disclose what the model actually uses as inputs. Though some companies have used the idea of trade secrecy to protect themselves, others still (competitors at times) have opted to reveal their training datasets. In the name of fairness, some data should remain uncrunched and some variables should be left out your model. Consider the following example from Cathy O’Neil’s *Weapons of Math Destruction*:

“From time to time, people ask me how to teach ethics to a class of data scientists. I usually begin with a discussion of how to build an e-score model and ask them whether it makes sense to use “race” as an input in the model. They inevitably respond that such a question would be unfair and probably illegal. The next question is whether to use “ZIP code.” This

[11] https://govex.jhu.edu/three-ways-data-quality-challenges-arise/

seems fair enough, at first. But it doesn't take long for the students to see they're codifying past injustices into their model. When they include an attribute such as "ZIP code," they're expressing the opinion that the history of human behavior in that patch of real estate should determine, at least in part, what kind of loan a person who lives there should get."

When we look at how to make algorithms more fair, choosing the right inputs is high on the list of effective measures. Other sources of unfairness in data include different populations with different properties (ex: SAT scores favor those with means to pay for tutors) and less data (by definition) about minority populations.

Even when bias is eliminated from data, algorithms can create destructive feedback loops. A negative outcome shouldn't engender additional negative outcomes down the road. A feedback loop in algorithms is where the results are somehow fed back into the model (sometimes intentionally, other times not). A harmless example might be an ad placement model. But in the case of recidivism risk models, the consequences can be brutal. If you are at a higher risk of recidivism according to the algorithm, then the judge tends to sentence you for longer. When you are put in jail for longer, then by the time you get out of jail, you typically have fewer resources and fewer job prospects, and often more likely to end up back in jail as a result. By being labeled high risk, you become high risk. In this way algorithms can create their own reality in a sense.

There are numerous and competing notions of fairness relevant to the discussion. Equal treatment of groups, unbiased treatment of groups, and fair treatment at the individual level are all relevant considerations, but unfortunately we can't achieve them all simultaneously. Accordingly, decisions need to be made about which are more important for a given scenario. We can also make choices about the cost we associate with errors made by an algorithms. When testing for accuracy as previously mentioned, you can calculate the error rates and types for different sub-populations and assess the potential differential impacts.

Fairness concerns necessarily creates a tradeoff with performance. If attaining the right level of fairness means we don't use certain variables, we lose out on certain amount of predictive accuracy. Thus, imposing fairness constraints on an algorithm does not come without a cost[12]. How we choose between different points along the tradeoff curve should be considered on a case-by-case basis.

References

Additional readings considered.

- O. Wise, personal communication, July 17, 2017.
- A. Nicklin, personal communication, July 25, 2017.
- https://medium.com/@technickle/three-ways-data-quality-challenges-arise-fb2c4cafe9a9
- https://www.newscientist.com/article/mg23431195-300-bias-test-to-prevent-algorithms-discriminating-unfairly/
- https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html
- https://datasociety.net/events/databite-no-101/
- https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/
- Obama Whitehouse - Preparing for the Future of AI

[12] https://www.youtube.com/watch?v=lRIg_oMNF_c

5. Appendix

Additional documentation:

- Project charter
- Exhaustive list of resources
- Notes from interviews
- Unstructured research notes

Pew Research Center's "Seven major themes about the algorithm era" from their large-scale canvassing of technology experts, scholars, corporate practitioners and government leaders:

Seven major themes about the algorithm era

INEVITABLE ALGORITHMS	**Theme 1**	**Algorithms will continue to spread everywhere** • The benefits will be visible and invisible and can lead to greater human insight into the world • The many upsides of algorithms are accompanied by challenges
	Theme 2	**Good things lie ahead** • Data-driven approaches to problem-solving will expand • Code processes will be refined and improved; ethical issues are being worked out • "Algorithms don't have to be perfect; they just have to be better than people" • In the future, the world may be governed by benevolent AI
CONCERNS	**Theme 3**	**Humanity and human judgment are lost when data and predictive modeling become paramount** • Programming primarily in pursuit of profits and efficiencies is a threat • Algorithms manipulate people and outcomes, and even "read our minds" • All of this will lead to a flawed yet inescapable logic-driven society • Some fear people could lose sophisticated decision-making capabilities and local intelligence • As code takes over complex systems, humans are left out of the loop • Solutions should include embedding respect for the individual
	Theme 4	**Biases exist in algorithmically-organized systems** • Algorithms reflect the biases of programmers and datasets • Algorithms depend upon data that is often limited, deficient or incorrect
	Theme 5	**Algorithmic categorizations deepen divides** • The disadvantaged are likely to be even more so • Algorithms create filter bubbles and silos shaped by corporate data collectors. They limit people's exposure to a wider range of ideas and reliable information and eliminate serendipity
	Theme 6	**Unemployment will rise** • Smarter, more-efficient algorithms will displace many human work activities • Some seek a redefined global economic system to support humanity
SOCIETAL CHALLENGES	**Theme 7**	**The need grows for algorithmic literacy, transparency and oversight** • It starts with algorithm literacy – this goes beyond basic digital literacy • People call for accountability processes, oversight and transparency • Many are pessimistic about the prospects for policy rules and oversight

PEW RESEARCH CENTER