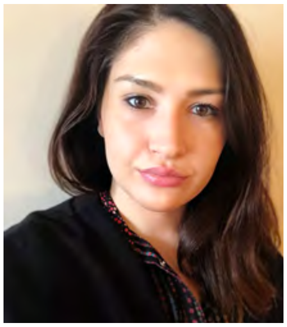
Algorithmic Auditing: The Key to Making Machine Learning in the Public Interest

By Sara Kassir

As the burdens for collecting enormous amounts of data decreased in recent years, advanced methods of analyzing this information rapidly developed. Machine learning, or the automation of model building, is one such method that quickly became ubiquitous and impactful across industries. For the public sector, artificially intelligent algorithms are now being deployed to solve problems that were previously viewed as insurmountable by humans. In international development, they are working to predict areas susceptible to famine; in regulation, they are detecting the sources of foodborne illness; in medicine, they are adding greater speed and precision to diagnostic processes.

The advancements presented by big data and machine learning are undeniably promising, but the technology also poses significant risks, particularly when algorithms are assumed to be infallible. While it may be true that these applications process any information they are given "objectively," human-generated data invariably reflects human biases. Therefore, automated tools can end up entrenching problematic simplifications about the world. Both government and private sector industry players have experienced calls to proactively address this issue.

This article argues that the risks of machine learning applications are best mitigated through a process of "algorithmic auditing," which institutionalizes accountability and robust due diligence in the technology. By assessing the ways in which bias might emerge at each step in the development pipeline, it is possible to develop strategies



***Sara Kassir** is a recent Master of Public Policy graduate of the Harvard Kennedy School.*

for evaluating each aspect of a model for undue sources of influence. Further, because algorithmic audits encourage systematic engagement with the issue of bias throughout the model-building process, they can also facilitate an organization's broader shift toward socially responsible data collection and use.

What is an algorithmic audit?

Algorithmic auditing is an effort to ensure that the context and purpose surrounding machine learning applications directly inform evaluations of their utility and fairness. Stephen Hawking wrote about the limitations of abstraction in his *A Brief History of Time:* "The usual approach of science of constructing a mathematical model cannot answer the questions of why there should be a universe for the model to describe. Why does the universe go to all the bother of existing?"[1] Admittedly, the esteemed theoretical physicist was not writing about machine learning applications in the public sector, but his message on intentionality in analysis is nonetheless salient. Data, models, algorithms, and other means of simplifying the world cannot be separated from the context in which they are produced. Through audits, machine learning tools are examined with the appropriate frame of reference in mind.

What principles from auditing can be translated to machine learning?

With the professional practice dating back to the Industrial Revolution, an audit is defined as "a formal examination of an organization's accounts," initially with the intent of protecting a firm's investors from fraud.[2] Over the past century, these examinations have diversified to encompass goals much broader than identifying financial risk. Today, auditors may examine an organization in terms of its regulatory compliance, process efficiency, environmental impacts, or ethical standards. But regardless of the precise focus, the procedure is directed toward the establishment of legitimacy. According to sociologist Mark Suchman, this "generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions" is crucial to an entity being able to function within society.[3]

The degree of assurance that can actually be achieved through an audit obviously varies depending on the industry. For example, in recent years, the field of safety engineering has consciously attempted to signal that no audit can ever deem something like an airplane 100 percent safe. As one report from the UC Berkeley School of Information points out, "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."[4] Machine learning applications, with their diverse consequences and potential for bias to emerge, are similarly impossible to ever deem "100 percent risk-free," and this spirit of imperfect assurance should inform how they are tested. In particular, three tenets of general auditing theory map well to the complexity of auditing algorithms specifically. These are: (1) the notion that an auditor must exercise judgement to explore the relevant details of a case; (2) the need to assess the inner-workings of process, rather than only examining its outputs; and (3) the expectation that an organization, subject to auditing endeavors, document its activities for the purposes of evaluation.

Principle 1: Marrying structure and judgment

The first auditing principle that is relevant for machine learning processes relates to the "steps" an auditor is expected to follow in completing their examination. Auditing, like any profession, is subject to ongoing debates about best practices. Scholar Michael Power, who explores the field as a principle of social organization, describes one of the industry's greatest tensions as the "structure-judgment" problem, or the notion that a tradeoff exists between auditing procedures that rely on prescribed techniques and those that give greater weight to individual judgement. Power uses the metaphors of "mechanism" and "organism" to describe the debate.[5] Mechanism names an aspiration for an integrative formal approach to audit, which holds out the promise of an algorithmic knowledge base. Organism assumes that the whole is always greater than the parts, and that the specificity of knowledge places limits on the mechanistic world view. In recent years, auditing firms have been increasingly pulled toward the former approach as they seek to standardize their offerings and manage human resources.[6]

"Structure" and "judgement" may appear to be at odds from the perspective of a company that performs external financial audits, but the dichotomy presented by Power and other scholars actually proves useful in the context of algorithmic auditing. For an individual attempting to evaluate a machine learning algorithm for bias, both approaches have merits. In framing the systematic investigation of bias, risks present in machine learning tools as an "audit." It is worth noting that organizations may initially view the exercise as intrusive or punitive. As the above discussion of auditing reveals, productive evaluations require collaboration from the people who built and use the model, so it is important to actively combat this perception. Rather, the "audit-ready" organization is one that understands there are genuine benefits to having an objective "extra set of eyes" look for bias risks in a model. A few cultural aspects can help facilitate this productive exchange, including clear consensus around goals and cross-disciplinary inputs into the development process.

Principle 2: Examining outputs, as well as inputs

Another auditing principle that proves useful for evaluating algorithms is the notion that a system must be comprehensively assessed for integrity. In some ways, this framing actually contradicts the training of data scientists. As statistician Leo Breiman once wrote, "Predictive accuracy on test sets is the criterion for how good the model is."[7] The focus on reducing test error rate of a model—which represents performance on data that was not included in the training set—has shaped much of the progress made in machine learning over the past two decades. As one publication from the UC Berkeley School of Information notes, "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."[8] Accordingly, audits cannot assume that a seemingly correct output from a model is sufficient evidence that the appropriate inputs were used, particularly when the goal is to minimize systematic bias.

Indeed, an over-emphasis on test error rates is particularly problematic from the perspective of mitigating algorithmic bias. Consider what happens if a developer is building a model to predict the likelihood that an individual will repay a loan they are issued. During the testing process, the developer splits the population by demographic background, and notices that the model is more likely to predict a positive outcome for certain groups. One possible reason for this discrepancy could be that the training data, which is primarily comprised of the credit records of individuals and their demographic characteristics, disproportionately represents on group. To improve the error rate for other subgroups, the developer might mechanically adjust some of the model's parameters so that it performs better across all groups. While this practice works as a "band-aid" solution for the existing population, it also means that the developer never has to question the structural features of the data that are feeding the biased results. This means such bias could reemerge as the model is deployed over new demographic subgroups.

The preceding example does not imply that the process of adjusting model parameters is inappropriate in and of itself. Rather, the practice draws attention to the fact that algorithmic audits are meant to identify sources of bias that might not be paid much attention during development. As such, it is just as important for an audit to examine inputs and their processing as it is to measure outputs (predictions) for accuracy.

Principle 3: Relying on robust internal documentation

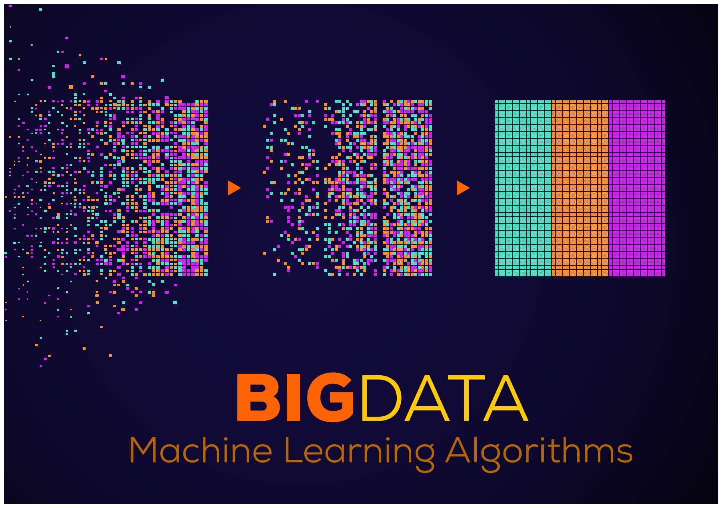
Finally, the tenet of auditing that is perhaps the most important for evaluating algorithms is also the most difficult and controversial. Namely, this is the idea that in order for audits to occur, the organization in question must make an effort to document its activities for the purposes of later review. Power describes this process quality as "verifiability," or the "attribute of information which allows qualified individuals working independently of one another to develop essentially similar measures or conclusions from an examination of the same evidence, data or records."[9] With respect to machine learning applications, "verifiability" might require keeping track of everything ranging from how the data is cleaned to how individuals are trained to interpret and act upon the model's results.

At a high level, there are two types of challenges related to producing "auditable" machine learning applications. First, the most sophisticated and accurate algorithms in use today (e.g., neural networks) are exceedingly complicated and cannot be effectively described through language. In short, these types of algorithms come at the cost of constraining an auditor's ability to parse out the inner-workings of a model, though providing the benefit of improved accuracy. Second, as often claimed by companies that rely on algorithms for revenue, open models allow for the possibility of "gaming" by the constituent population. While it is not necessarily inevitable that the results of an algorithmic audit are made public, organizations with this worry may hesitate in fully documenting internal procedures.

Notably, even in cases where the exact contours of an audit are subject to debate, the idea that machine learning applications should be developed with a certain degree of formal documentation is crucial for risk mitigation. Regardless of the algorithm's complexity, organizations can still commit to transparency around factors such as optimization criteria, data inputs, sampling processes, and feedback loops, all of which can limit the potential for unintentional bias to become entrenched.

Key takeaways

To summarize how the above-mentioned principles of auditing translate to practice, an algorithmic audit involves examining each part of a model's lifecycle, using a combination of standardized best practices and discretionary judgment calls, all of which are informed by the available documentation and social context. In the words of one article from *Harvard Business Review,* the process "must be interdisciplinary in order for it to succeed," relying on "social science methodology and concepts from such fields as psychology, behavioral economics, human-centered design, and ethics."[10] Such an approach is necessary in light of the fact that no complete list of "wrong" practices exists in machine learning. Rather, the goal for an auditor must be to ask if any steps of the development process are approached in a manner that does not give sufficient attention to the issue of bias, with the definition of "sufficient" obviously varying across contexts.



The public sector's potential for leadership

As society begins to grapple with the potential drawbacks of machine learning, perhaps with some of the initial fervor surrounding big data subsiding, public sector organizations are presented with an opportunity. Rather than waiting for the issues of bias to be solved by technology companies or relying on legislators to push regulation, algorithmic auditing serves as a middle ground, balancing progress with caution. Governments should pursue innovative data analysis methods that will empower them to better serve and understand their constituencies, but in a manner that promotes accountability and equity.

Footnotes

1. Hawking, S. (2018). *A Brief History of Time.* Random House USA.
2. Audit. (n.d.). Retrieved from https://www.merriam-webster.com/dictionary/audit
3. Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *The Academy of Management Review,* 20(3), 571. doi:10.2307/258788
4. Burrell, J. (2018, August 13). Report from the first AFOG Summer Workshop (Panel 3). Retrieved from http://afog.berkeley.edu/2018/08/13/report-from-the-first-afog-summer-workshop/
5. Dirsmith and Haskins (1991). Inherent risk assessment and audit firm technology: A contrast in world theories. *Accounting, Organizations and Society,* 16(1), 61-90.
6. Power, M. K. (2003). Auditing and the production of legitimacy. *Accounting, Organizations and Society,* 28(4), 79-394. doi:10.1016/s0361-3682(01)00047-2
7. Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science,* 16(3), 199-231. doi:10.1214/ss/1009213726
8. Burrell, J. (2018, August 13). Report from the first AFOG Summer Workshop (Panel 4). Retrieved from http://afog.berkeley.edu/2018/08/13/report-from-the-first-afog-summer-workshop/
9. Power, M. (1996). Making things auditable. *Accounting, Organizations and Society,* 21(2-3), 289-315. doi:10.1016/0361-3682(95)00004-6
10. Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., & Katyal, V. (2018, December 01). Why We Need to Audit Algorithms. Retrieved from https://hbr.org/2018/11/why-we- need-to-audit-algorithms