Model Behavior: Mitigating Bias in Public Sector Machine Learning Applications

Sara Kassir

July 2019

M-RCBG Associate Working Paper Series | No. 130

The views expressed in the M-RCBG Associate Working Paper Series are those of the author(s) and do not necessarily reflect those of the Mossavar-Rahmani Center for Business & Government or of Harvard University. The papers in this series have not undergone formal review and approval; they are presented to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s). Papers may be downloaded for personal use only.
model behavior

mitigating bias in public sector machine learning applications

Client: IBM Center for the Business of Government

Sara Kassir
Master in Public Policy Candidate, 2019

Advisor: Professor Jim Waldo
Seminar Leader: Professor John Haigh

2 April 2019

This PAE reflects the views of the author(s) and should not be viewed as representing the views of the PAE's external client(s), nor those of Harvard University or any of its faculty.
table of contents

chapter one: bias in public sector algorithms
   i. introduction - 1
   ii. the basics: defining bias in public sector data applications - 2
   iii. the challenge: how machine learning complicates the issue of bias - 5
   iv. the impact: real-world cases & their consequences - 10
   v. summary - 15

chapter two: regulation & risk mitigation
   i. introduction - 16
   ii. the landscape: proposed strategies to mitigate algorithmic bias - 16
   iii. the framework: modern theories of managing ambiguous risks - 22
   iv. the choice: continuing with the inertia of the Big Data revolution - 26
   v. summary - 26

chapter three: algorithmic auditing
   i. introduction - 27
   ii. the theory: auditing principles, applied to algorithms - 28
   iii. the practice: deriving a framework for ML audits - 30
   iv. the culture: an audit-ready organizational mindset - 38
   v. summary - 40

chapter four
   recommendations & conclusion - 41
executive summary

As the burdens for collecting enormous amounts of data decreased in recent years, advanced methods of analyzing this information rapidly developed. Machine learning (ML), 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 ML 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, though under the unfortunate guise of neutrality. Concerns about “racist robots” and “sexist machines” have created mounting pressure for government intervention on artificial intelligence, but the form such action would take is unclear. Private sector industry players have also experienced calls to proactively address the issue.

This paper seeks to make the challenges surrounding machine learning actionable by contextualizing the technology alongside other modern innovations that have generated ambiguous risks. Like the harms of nuclear radiation or cyberattacks, the consequences of algorithmic bias are imprecise and context-dependent, often surprising organizations with unforeseen consequences. As history has shown, environments driven by such uncertainty tend to be misinterpreted by “top-down” regulatory frameworks. Instead, ambiguous risks are better handled through internal organizational structures that promote accountability and robust due diligence.

In the case of machine learning applications, this paper argues that accountability and robust due diligence are best achieved through a process known as algorithmic auditing. By assessing the ways in which bias might emerge at each step in the technical development pipeline, it is possible to develop strategies 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.
In 2013, the researcher Kate Crawford argued that the technology industry – with its purported objectivity and certainty – was leading society astray. She coined her concern "Big Data fundamentalism," or the idea that with larger datasets, decision-makers can be brought ever closer to absolute truths. “Data and data sets are not objective; they are creations of human design,” she cautioned. “Hidden biases in both the collection and analysis stage present considerable risks, and are as important to the big-data equation as the numbers themselves.”

Five years later, the potential for data and algorithms to encode – and even exacerbate – human biases is widely acknowledged. Silicon Valley is under mounting pressure to demonstrate control over this problem – over the course of a single month in late 2018, Amazon scrapped a secret recruiting tool that was showing bias against women, Google removed gendered pronouns from its latest Smart Compose tool for Gmail, and Salesforce shut down a project for reading human emotions due to unrepresentative training data. Clearly, effective strategies for curbing bias remain nebulous, despite the extensive resources being directed toward this issue.

The challenges presented by an information-rich environment by no means suggests that public sector organizations should abandon these tools. On the contrary, the latest methods of analyzing government data sources present incredible opportunities to promote efficiency, equity, and progress in many parts of society. **It is exceedingly important, however, that organizations seeking to make use of new technology do so with a ready understanding of the ongoing challenges and risks.** This is particularly true in the context of machine learning (ML) applications, which are increasingly being used to make decisions on topics ranging from healthcare to employment to education.

Before discussing strategies for appropriately implementing such tools, this chapter serves as an overview of the problem of bias, which government organizations inevitably encounter in contemporary approaches to data analytics.
When it took the U.S. government nearly a decade to compile the results of the 1880 census, a former Census Office employee named Herman Hollerith responded with the public sector’s first foray into Big Data. Using coded punch cards and an electromechanical tabulation machine, Hollerith created a system that is often regarded as the foundation of modern automatic computation. The wild success of his invention – decreasing the time to process the census to only three months – led him to eventually start his own company, with governments and insurance providers around the world leasing the equipment. In 1911, Hollerith’s firm was one of four that were merged to create the Computing-Tabulating-Recording Company, which would eventually become International Business Machines (IBM) in 1924.5

Today, Big Data is defined as “extremely large datasets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.”6 While applications may have existed at the federal level for some time, such efforts are more recent for local governments. When New York City launched its 311 Citizen Service Management System in 2003, the main purpose was articulated as providing residents with easier access to city information. By 2005, however, Gotham Gazette reported an interesting shift in the perceived benefits of the system, stating, “311’s real potential lies in the immense amount of data it has accumulated.”7

Initiatives like 311 have certainly increased the amount of information available to government officials, but it is problematic to assume that the findings produced by this information are inevitably objective. On the contrary, data-driven insights can be “biased” for a variety of reasons, and it is crucial that public sector entities seeking to use these insights are aware of their shortcomings. While the concept of “bias” is associated with a wide variety of words – unfairness, favoritism, subjectivity, skew – it is difficult to pin down a precise meaning. The definition also varies across disciplines, with statisticians describing it as “the tendency of a sample statistic to over- or under-estimate a population parameter,” and legal scholars describing it as “an unfair act or policy stemming from prejudice.”

In the context of public sector applications that can impact people’s lives, bias is best defined as a failure to distinguish between information that is meaningful and mere “noise” in the data. Such mix-ups readily occur in human decision-making – for example, studies have found that employers deciding whether to extend an interview to an applicant incorporate the “whiteness” of their name in the calculus, despite the fact that this trait is clearly unrelated to professional competence.8 When decisions are automated, computer programs read all variables as equally meaningful – until humans make decisions about which types of “noise” are most important to avoid. While the tradeoffs involved are occasionally obvious (e.g., for a hiring algorithm, discrimination on the basis of names is likely more harmful than discrimination on the basis of extracurricular activities), making such calls can be very challenging.

The type of bias that governments should be concerned with identifying and mitigating is a combination of several occurrences. First, in order to approximate a real-world phenomenon, a measurable proxy is established. In the case of New York’s 311 system, “count of calls to the center” might be treated as a proxy for “actual infrastructural service needs of New Yorkers.” Second, in order to claim that the metric is useful, an assumption is made that the distance between the proxy and the real-world phenomenon is consistent.
For 311, the assumption is that all New Yorkers express the same proportion of their actual needs to the call center, regardless of age, race, or class. (Notably, this is likely untrue, given varied degrees of trust in government by these groups, combined with the fact that people disagree over what qualifies as “damaged” infrastructure that the city is responsible for.) Finally, over time, the definition of the proxy shifts to become one in the same with the real-world phenomenon, allowing insights derived from it to carry greater authority. In other words, since there is no conceptual difference between “count of 311 calls” and “actual needs,” the calls become a basis for public initiatives like resource allocation. This basis, however, is biased.

The fact that the 311 data is not an ideal proxy for a nuanced concept like “infrastructural service needs” does not imply that the information is useless. On the contrary, it is important to note the role of context in determining if a metric is biased. In the words of one researcher, “A measurement can only be ‘biased’ insofar as it purports to measure something else.” If “count of 311 calls” were a proxy for a different variable, like “citizen outreach to government,” there would be less potential for bias, since there is less distance between these concepts. The reality that a metric that is perfectly fitting in one instance can be discriminatory in another is particularly relevant given the growth of interagency data-sharing efforts and open portals. With more information available, it is very easy to apply metrics to inappropriate contexts.

The “measurable proxy” mentioned above is the simplest type of model an organization could possibly employ, in which a single variable is used to explain the outcome of interest. Figure 1 depicts this relationship. Today, with the extensive availability of massive datasets, organizations seek more nuanced tools for understanding the problems they face. As the complexity of a model grows, however, so does the task of identifying and estimating the nature of its bias. Consider Figure 2, shown on page 4. In addition to the introduction of multiple sources of information, all of which have different relationships to their real concepts of interest, an organization must apply some function to combine this information. Whether this is done through statistical rules or algorithmic learning, the question becomes how close the estimates produced by this function are to the true concept of interest. If this distance is significant and/or variable for different segments of the population, bias may be present.
Practically speaking, some amount of distance will always be present between the “concept of interest” and the “model estimates.” Like in the single-variable case, the degree of this distance is determined by how closely the model’s outputs correspond to the intentions driving the analysis. There is an element of choice involved in this relationship. For example, if the concept of interest is “actual infrastructural repair needs of New Yorkers,” should the model seek to produce estimates that correspond most closely to (A) the count of 311 calls or (B) reports from the 59 community boards across the city? There is no universally correct answer to this question – option A might result in outputs that are biased against citizens who are disinclined to engage with the government, while option B might prove biased against neighborhoods with less effective community boards.

An important way in which bias emerges in Big Data applications then, is when organizations are not realistic about the conceptual assumptions driving their tools. If a city government builds a model that performs excellently at predicting which neighborhoods are likely to complain via 311 about infrastructural problems, but claims that it is also a robust basis for understanding the service needs of all New Yorkers, the resulting decisions will be problematic for major segments of the population. Scholars like David Weinberger are increasingly calling for organizations to be more transparent about what their analytical tools are “optimized” for, since this will call increased attention to the reality that “machines are imperfect and are often designed to serve inconsistent goals.” Frank conversations around optimization tradeoffs would also likely be useful in combatting the notion that data-driven decisions are inherently “objective” and do not need to be informed by clear human reasoning.

Clarity around goals should certainly be an imperative for any public sector initiative, but even the most well-intentioned of models sometimes result in biased conclusions that humans cannot predict. The next section will demonstrate how this occurs in the context of artificially intelligent approaches to data analysis.
The relative simplicity of the 311 example might beg the question – why is mitigating bias in everyday data analytics tools so hard? Over the past decade, there has been an explosion in a specific kind of data analysis known as “machine learning” (ML), or the deployment of algorithms that can improve their performance at a task over time as they receive more data. Because these applications are “artificially intelligent,” meaning they do not rely on an explicit set of rules programmed by a human, it can be exceedingly challenging to pinpoint the factors that lead them to draw particular conclusions. Further, given the speed at which ML algorithms process information, humans might not even be aware of the problematic nature of a model’s results – until they have been used to make a wrong-headed decision.

In general, the process of developing an ML model involves several steps, all of which can serve as opportunities for bias to enter the pipeline. First, a dataset for training is identified, which maps a set of predictor variables (e.g., the independent variables) to an outcome of interest (e.g., the dependent variable). After the data is cleaned so that it can be processed, a developer tests different algorithms to determine which best captures the complex relationships between the predictor variables and the outcome. To complete this testing, 70-80% of the existing data is set aside for training, while the remaining 20-30% is reserved for validation. Once a model is trained on the 70-80% portion, the developer assesses its accuracy on the outstanding 20-30%, comparing the generated estimates to the known outcomes. Finally, if the results of this test are deemed satisfactory given the established definition of “accuracy,” the model is deployed, allowing it to make new predictions as fresh input observations become available. Table 1 summarizes this process in six steps, along with the associated pitfalls that may introduce bias to the model.

---

1 Notably, the process of “learning” can either happen in real-time as new data becomes available, or through adjustments occasionally initiated by developers.

2 The subsequent explanation refers specifically to supervised, structured models. Unsupervised machine learning models explore unlabeled data, attempting to find relationships and patterns, without reference to known outcomes.

3 For a thorough discussion of the technical concepts of machine learning, see: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R*. New York: Springer.
<table>
<thead>
<tr>
<th>process step</th>
<th>definition</th>
<th>some potential sources of bias</th>
</tr>
</thead>
</table>
| training data         | Researching what metrics and supporting data are available and relevant to the question at hand | ▪ Sample is not representative of the target population  
▪ Predictor variables available are poor proxies for the real independent variables of interest  
▪ Data sources suffer from measurement error |
| identification        |                                                                             |                                                                                                |
| pre-processing        | Taking steps to clean, standardize, and otherwise transform data into a usable form | ▪ Values are missing or have errors  
▪ Thresholds for ordinal variables are arbitrarily assigned  
▪ Assumptions used to extrapolate variables (e.g., to fill in missing years) are flawed |
| model selection &     | Experimenting with various algorithmic approaches to maximize fit for the given dataset | ▪ Model is overfit or underfit on the training data (e.g., bias-variance tradeoff)  
▪ Cutpoints/ splines for polynomial functions are set at inappropriate thresholds  
▪ Feature selection fails to capture domain knowledge (e.g., important variables are left out, or unimportant ones are included) |
| training             |                                                                             |                                                                                                |
| accuracy evaluation   | Testing predictive performance of the model to optimize its utility in a real-world setting | ▪ Test errors are not generated for population subsets  
▪ Training sample is too small to meaningfully test performance on population subsets  
▪ Predictive performance is too narrowly defined (e.g., considering false positive rate, but not false negative rate) |
| insights presentation | Designing how estimates generated by the model information will be presented to the user | ▪ User is instructed to prioritize model estimates over personal subject matter expertise  
▪ User is not able to examine the input variables that generated a given prediction |
| deployment            |                                                                             |                                                                                                |
| feedback incorporation| Updating the model as new information becomes available | ▪ New data suffers from measurement bias (e.g., gaming)  
▪ New data collection does not allow for feedback loops  
▪ Model only receives selective feedback, leaving large gaps in performance |
|                       |                                                                             |                                                                                                |
To provide a more concrete illustration of the pitfalls summarized in Table 1, consider a government that has recently undertaken an effort to identify health care providers who are inappropriately prescribing opioids. Over the past two years, the agency behind this initiative randomly selected 2,000 providers to audit, using a very time-consuming and manual process. Having successfully determined that certain doctors were in fact inappropriately prescribing, the goal is now to expand this effort using a machine learning application.

1. training data identification: The training data for this algorithm will include information on the doctors that were manually investigated, as well as the results of the investigation. One way in which bias could enter the pipeline is if the developer building the algorithm was not involved in deciding which doctors to manually audit. The developer may not know, for example, that many of the audited doctors come from counties with high incidences of drug overdose. Two things happen to be true in this state – (1) overdoses are on the rise in rural counties, and (2) doctors in rural counties are more likely to be primary care physicians. While a physician’s specialty may be irrelevant to their prescribing practices, the sampling of the training data may result in a model “learning” such a connection exists.

2. pre-processing: Since data collection processes are rarely perfect, decisions on how to handle missing values are an important part of ML development process, and a common way for bias to emerge. After interviewing some of the caseworkers involved in the original audit, a developer might believe that the variable “average time between patient encounters” is a crucial input for finding bad prescribing practices. Therefore, the analyst drops all observations from the training data that are missing this value. A problem could potentially arise, however, if the analyst does not realize that the doctors who are missing this information are systematically different from the doctors who are not missing it. It could be the case, for example, that privately-owned hospitals do not track this information, and all of the doctors working for these types of organizations end up excluded from the sample.

![Figure 3: mock opioid prescribing audit results, used as training data](image)
model selection & training: A common misconception about machine learning is that it can identify universal truths – but in reality, humans must make important decisions about how a model operates. The developer working on the opioid prescription case could take a look at the training data, and because there are relatively few doctors with “extreme” prescribing figures, opt to select a model that is not particularly sensitive to outliers. This less flexible model could be the appropriate decision to make if the doctors in the audit sample behave similarly to the rest of the doctors in the state, but if this assumption is flawed, the model will produce biased predictions. Figure 4 illustrates one way in which this could happen.

Figure 4: bias emerging from decisions about the treatment of outliers

Training data (left): Since relatively few doctors in the training data prescribe a high proportion of opioids, the model might downplay the influence of these observations, viewing them as “outliers.”

Real-world data (right): In reality, a much larger number of doctors prescribe a high proportion of opioids than the audit sample suggests. Further, there appears to be a correlation between this information and inappropriate prescribing practices, which the model will not effectively learn, since it has been optimized to be less responsive to extreme values for this variable.

accuracy evaluation: When it comes to testing the performance of an ML model, developers can face a challenging situation if they are working with a relatively small sample. The analyst working on the opioid prescription case might try to assess the predictive performance of the algorithm by splitting the training data into two parts: 1,500 doctors to be used for training, and 500 to be used as a test set. The test results might be very promising when it comes to labeling each provider as “no impropriety” or “flagged for impropriety,” reaching the same conclusion as the human auditors 95% of the time.

The developer could unknowingly be missing significant incidents of bias, however, by failing to examine performance for subgroups of the population. For example, of the 5% of doctors that were misclassified, 80% of them were young women. Further, even if the analyst does realize this disparity in performance, if the training data does not include enough young female doctors to be a meaningful representation of the range of female doctors that exist in the real world, it will be extremely difficult to develop a model that performs well on providers within this demographic.
5. Insights Presentation: In order for a model to be practical, the predictions it generates must be presented to a user, who will then decide how to act. The final audience for the opioid prescription algorithm might be an auditing manager, who reviews each flagged case and decides if the doctor needs to receive a warning. To provide some context around the predictions, the analyst develops an interface that includes information on how the doctor’s records compare to the rest of the population. However, depending on how the developer elects to define “the rest of the population,” the doctor’s behavior might appear more or less extreme. Figure 5 demonstrates how the auditing manager’s perceptions might be unduly influenced by the interface.

6. Feedback Incorporation: While developers are working on an initial ML model, their universe is relatively confined to the subtleties of the training data and any conversations they have with subject matter experts about how to interpret it. The true range of outcomes is not really knowable until a model is deployed – in this case, collecting information on new doctors who were not previously audited and making determinations about their prescribing behavior.

One common practice for validating the performance of a live model is to incorporate feedback loops – that is, providing it with information about which of its predictions were correct and which ones were not. Bias can still emerge at this point in the process, however, if the model is only receiving feedback on subset of observations. For example, the agency now using the algorithm might elect to test its conclusions by having a team of human auditors independently investigate any flagged doctors. If the model turned out to be flagging some providers who are subsequently cleared by the manual audit, this data could help to reduce the “false positive” rate of the system. Unfortunately, this feedback strategy will not work to reduce the “false negative” rate of the system, so providers who are systematically slipping under the radar will continue to do so, contributing to ongoing bias.
While the above case focuses on many of the technical ways in which bias can find its way into the ML development process, machine learning is also plagued by major gaps in cross-sector communication, which exacerbate these problems. At many points in the opioid prescription example, the developer must make calls about tradeoffs, and these decisions will likely depend on conversations with subject matter experts (SMEs). Important questions could include:

- Is there anything systematically different about the doctors in the audit sample versus the rest of the providers in the state?
- Is there any reason to be worried about the quality of particular data sources?
- Is maximizing the number of doctors who are identified as inappropriate prescribers the only goal, or are policymakers concerned about the disruptive effects of false alarms?

Even the most engaged developer cannot answer these questions without context-specific domain knowledge. Unfortunately, the popular misconception that machine learning algorithms can provide a one-size-fits-all solution to data-driven questions decreases the odds that such conservations are prioritized.

### iv. The Impact

**Real-world cases and their consequences**

Thus far, this chapter has used hypothetical examples to define algorithmic bias and illustrate the breadth of ways in which it can emerge. To provide greater context around the real-world impacts, this section reviews several public sector cases in which contemporary data analytics tools resulted in problematic outcomes for their constituencies.

Before delving into these examples, it is crucial to underscore that, when properly deployed, algorithms can help solve problems that humans have previously found insurmountable. 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. These successes, however, must be balanced against the risk that can emerge when technology is not carefully considered and monitored.

### U.S. State and Local Judicial Systems: Assigning Risk Scores to Criminal Offenders

In 2016, non-profit newsroom ProPublica published a report on a criminal risk assessment tool called COMPAS, providing the first major investigation into the potential ramifications of algorithmic bias in public sector technology. COMPAS, or Correctional Offender Management Profiling for Alternative Sanctions, was created by a for-profit company called Northpointe to provide justice systems with the ability to “take a ‘retrospective’ look at risk and needs factors for making treatment and supervision decisions” as well as the ability to examine “‘prospective’ risk and needs for persons transitioning back into the community from incarceration.”

The purported benefit of the tool is to combat the systemic racism, both implicit and explicit, that has contributed to the disproportionate number of black people in the U.S. prison system.
COMPAS relies on 137 questions to arrive at predictions about which individuals are most likely to recidivate, thereby informing decisions about the awarding of bail, lengths of sentences, and terms of probation. While race is not one of the questions, the system does rely on information like a defendant’s education level and employment status, along with factors like whether their parents were ever incarcerated and if their friends use or sell illegal drugs. Individuals are also asked to agree or disagree with questions such as “If someone insults my friends, family, or group they are asking for trouble” and “The law doesn’t help average people.” According to ProPublica’s analysis using 7,000 risk scores of people arrested in Broward County, Florida during 2013 and 2014, “The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.”

Given the claim that COMPAS is a useful tool for mitigating human bias in the judicial process, ProPublica sought to examine its predictive power across different racial groups. Whereas an algorithm that had a similar error rate across racial groups would have signaled less bias than human decision-makers, Northpointe’s software proved to be racially inconsistent for both false positives and false negatives. In other words, black defendants who did not go on to recidivate were more likely to be labeled as “high risk,” while white defendants who did go on to recidivate were more likely to be labeled as “low risk.” The inevitable consequence of this disparity is biased sentencing practices, particularly among judges who are inclined to translate scores directly to their decisions.

While the questions used by COMPAS are known, the precise nature of the algorithm is not, making it essentially impossible to determine the source of bias in the predictions. In the U.S. context, any number of systemic injustices could be contributing to the racist
conclusions drawn by the software, but Northpointe’s claims of proprietary rights make the necessary examination unlikely. Despite this problematic opacity, courts across the country have been using similar technology since as early as 2002, with its effectiveness largely untested.

PredPol is a for-profit company that sells crime prediction software to police departments around the U.S., beginning with the LAPD in 2011. According to its website, “PredPol has a precise definition of predictive policing. For us and our customers, it is the practice of identifying the times and locations where specific crimes are most likely to occur, then patrolling those areas to prevent those crimes from occurring.” The company also claims an impressive geographic spread, noting that its products are “currently being used to help protect one out of every 33 people in the United States.”

In terms of its research origins, PredPol is the brainchild of anthropology professor Jeffrey Brantingham at the University of California Los Angeles, who adapted the technology from his Pentagon-funded research on predicting military casualties in Iraq. Algorithms are trained on 2 to 5 years of data from a given police department, using only information on the time, place, and type of crimes committed. Notably, the algorithm is not trained on any information regarding the people involved in the incidents. The resulting output provides law enforcement with a city map of 150x150-meter square boxes, coded by risk level for criminal activity on given days and at particular times of day. During the initial rollout of PredPol, Los Angeles’ Foothill Division saw a 13% reduction in crime over four months, compared to a 0.4% increase in crime in other parts of the city that had not deployed the technology.

While PredPol rather quickly spread to cities like Atlanta and Philadelphia, the technology was never widely adopted in other parts of California. Law enforcement officials in Oakland earmarked $150,000 for the software during budget planning in 2015, but ultimately opted out of implementing it, largely due to the belief that it would erode public trust. In the words of Tim Birch, Head of Research and Planning at OPD, “Maybe we could reduce crime more by using predictive policing, but the unintended consequences are even more damaging...and it’s just not worth it.” About a year after the OPD’s decision to renounce PredPol, a study by the non-profit, Human Rights Data Analysis Group, used Oakland’s publicly available crime data to fully demonstrate the extent of such consequences.

In an article published in *Significance* in October 2016, researchers Kristian Lum and William Isaac test the utility of a publicly-available version of the PredPol algorithm in predicting drug crimes across the city in 2011, using crimes from 2010 as training data. Data from the U.S. Census and the Centers for Disease Control and Prevention indicate that drug users reside throughout Oakland, and an ideal model would successfully map to this information. However, according to Lum, “We find that targeted policing would have been dispatched almost exclusively to lower income, minority neighborhoods. This is because, in records from the time prior to running the algorithm, the majority of the drug crimes recorded were in these same neighborhoods.” In other words, the model simply reinforced the notion that the department should continue dedicating resources to policing parts of Oakland that are already over-policed.
Because a history of racially-biased profiling is not unique to the Bay Area, many of the cities that rely on PredPol to inform their law enforcement strategies are likely experiencing the type of algorithmic bias identified by Lum and Isaac. Unfortunately, the ongoing practice of over-policing neighborhoods that are mostly black and poor only creates more data to support the notion that these are the neighborhoods in need of higher scrutiny, leaving crimes elsewhere underreported.

Arkansas Department of Human Services: Automating Needs Assessments of Disabled Beneficiaries

In 2012, Arkansas applied for a grant from the Centers for Medicaid & Medicare Services (CMS) that would increase the state’s funding for non-institutional, long-term services and supports (LTSS), or home-based care for individuals who cannot perform basic tasks for themselves. The state’s written application included the following statement: “Arkansas recognizes the benefits of a standardized, automated process, both for the consumers and for state policymakers...One of Arkansas’s strengths is the recent progress the state has made toward implementing automated standardized assessment instruments for individuals with LTSS needs.” The system ultimately implemented by the Arkansas Department of Human Services (DHS) is called the interRAI Home Care assessment, versions of which are used by about half of U.S. states.
Prior to the introduction of interRAI, Arkansas Medicaid recipients in need of LTSS care were interviewed by assessors about the nature of their disabilities. After answering questions about how much difficulty they had getting out of bed, feeding themselves, using the bathroom, and so on, patients would receive a personalized recommendation from a nurse or social worker on the number of hours they were eligible for an in-home aide. With the introduction of the interRAI tool, an algorithm was instead used to allot hours by sorting beneficiaries into “resource utilization groups,” or RUGs. The algorithm relied on about 60 characteristics, such as fever, weight loss, and ventilator use. When it recommended dramatically reduced hours of care for some severely disabled patients, many went without food, were forced to remain in soiled clothes, and saw their health deteriorate even further.25

An investigation into the technical nature of the interRAI system became inevitable when a group of Medicaid beneficiaries sued the Arkansas DHS. As one Legal Aid attorney working on the case noted, some of the underlying problems were a result of data collection errors: “One variable in the assessment was foot problems. When an assessor visited a certain person, they wrote that the person didn’t have any problems – because they were an amputee.” The model also employed a decision tree, which functions like a flowchart to sort individuals into the so-called “RUGs”. Because a single variable can result in a beneficiary being placed into an entirely different category, such algorithms can have high variation for patients close to the margins of the groups. In these cases, patients with the same conditions might receive wildly disparate amounts of LTSS support, simply based on how their symptoms are described.26

In May of 2018, a judge ordered the Arkansas DHS to terminate its use of the interRAI algorithm to determine individuals’ LTSS eligibility. While the court case had revealed several problems with the nature of the algorithm itself, the ruling focused on the implementation of the tool, particularly because the department had failed to notify beneficiaries of the changes to the assessment methodology.27

Outside the U.S., the United Kingdom is considered a hub for artificial intelligence research and development, with government officials working to implement predictive technologies to allocate services and improve outcomes. In its 2018 report “AI in the UK: ready, willing and able?”, the House of Lords noted the potential for artificial intelligence to cut time and costs in the public sector if deployed properly.28 One way in which this belief has manifested is in the Evidence Based Investigation Tool (EBIT), which is currently being used by police in Kent.

Developed by researchers at Cambridge University, EBIT “supports allocation decisions for certain public order offenses and common assault” by using “evidence-based solvability and public interest/vulnerability factors to determine whether a crime should be allocated, filed, or reviewed further.”29 To summarize, the idea behind the tool is that law enforcement can best serve communities by focusing on crimes that they have a chance of solving. In 2017, for example, the BBC reported that over 70% of burglaries that occurred in England and Wales in the four years prior were closed without a suspect. According to Deputy Chief Constable Ian Pilling, “We cannot do everything we would want to do...This means that we will assess every crime reported to us and that we will make decisions based on the threat and risk involved and on the likelihood of arresting and convicting the offender.”30
The initial version of EBIT implemented in Kent was trained on thousands of assault and public order cases, using eight factors related to solvability, including the presence of witnesses and surveillance footage. As summarized by News Scientist from an interview with the tool’s creator, Dr. Kent McFadzien, “Since these factors could change over time, EBIT always recommends one or two crimes with low solvability for investigation each day. The officers involved aren’t aware of this score, so this is a blind test of the algorithm’s effectiveness.” According to Dr. McFadzien, “It’s a permanently ongoing trial.”

In terms of how readily the program is being used, more than two-and-a-half times as many cases were dropped by law enforcement in 2017 due to being “screened out” than in the year prior. Such an outcome may not be troubling if the rate were consistent across crime types and victim demographics, but according to The Guardian, “police are increasingly dropping investigations into serious crimes such as sexual offences, violent attacks and arson.” Particularly concerning was the finding that 49 cases of sexual assault were dropped on the same day they were reported in 2017. While these outcomes reflect the unfortunate reality that these offenders have historically evaded prosecution, the resulting algorithmic bias could allow this trend to persist into the future.

More broadly, the notion of using EBIT has also presented law enforcement officers in the UK with a messaging problem. In the words of Police Federation Chair John Apter, who was elected in mid-2018, “I can think of nothing more insulting for someone who has been a victim of a crime than to discover that a computer algorithm has told a police force not to investigate because there is little chance of catching the culprit.” Accordingly, the technology appears to be sparking a larger debate on the human toll of public service automation in the name of cost-cutting.

V. Summary

- Big Data presents great potential for understanding complex problems, but only if analysis is conducted with greater awareness about the outcomes
- Bias in data applications can occur any time there is distance between a proxy metric and a concept of interest (e.g., “crime reported” vs. “real crime”), particularly if this distance varies across populations
- The nature of machine learning models presents significant opportunities for bias to manifest; they are artificially intelligent, meaning humans do not specify which relationships and trends the machines should “learn”
- Machines could end up learning and encoding trends that are functions of institutional behavior, human prejudice, measurement problems, etc., rather than “objective” truths
- Governments (and other organizations) can still reap incredible benefits from Big Data, so long as due attention is paid to issues like equity, fairness, and accountability
chapter two
regulation & risk mitigation

"Every technology, every science that tells us more about ourselves, is scary at the time." - Rodney Brooks (1954- )

i. introduction

The challenges associated with building unbiased machine learning models are in no way slowing interest in the field - according to Forbes, machine learning patents grew at 29.3% CAGR between 2010 and 2018, and worldwide spending on the industry is expected to grow to $77.6 billion by 2022. In the words of Alphabet Inc. president Sergey Brin, “The new spring in AI is the most significant development in computing in my lifetime.”

Unfortunately, the proportion of new investments in this technology that are being targeted specifically at the goal of ensuring that algorithms do not create or exacerbate inequalities is insufficient, particularly when contrasted with the pervasiveness of these applications in daily life. MIT president L. Rafael Reif recently emphasized this reality while announcing a $1 billion investment in machine learning at the university. “[AI] is rapidly enabling new researchers in every discipline and new solutions to daunting problems,” he stated in October 2018. “At the same time, it is creating ethical strains and human consequences our society is not yet equipped to control or withstand.”

Efforts to mitigate algorithmic bias may be underdeveloped and deprioritized, but the dramatic growth of the space has certainly led to an increasing diversity of players working to respond to the problem. This chapter will review some of the proposed strategies for approaching the risks of machine learning, particularly examining their respective utility for navigating the ambiguous trajectory of artificial intelligence.

ii. the landscape
proposed strategies to mitigate algorithmic bias

Marvin Minsky, co-founder of MIT’s Computer Science and Artificial Intelligence Laboratory, once wrote, “Every system that we build will surprise us with new kinds of flaws.” In many ways, this sentiment captures the challenge of applying a systematic approach to the risks of machine learning technologies. Despite the broad consensus that algorithmic bias is a problem that needs to be addressed for the public good, the context-dependent definition and infinite ways in which it can manifest present no universal solution. Further, given that an organization’s “bad” algorithms are only revealed as such
once the harmful effects are observable, it is unclear how proactive action to prevent such harm can be incentivized.

**framing the debate:**
**an analogue from cybersecurity**

To frame the debates around mitigating algorithmic bias, it is worth examining how another complex technological risk has been adjudicated by society. Consider the analogue of cyberattacks on Electronic Health Records (EHR) systems. While only 40% of all medical providers used EHR in 2012, this rate had increased to 67% by 2017.\(^{18}\) The technology's rapid spread brought with it benefits like efficiency and cost-cutting, but also involved the inevitable risk of patients' data being stolen by malicious actors. As a guide for small health care practices published by the U.S. Department of Health and Human Services (HHS) in 2008 noted, EHR "fundamentally changes your practice's health IT environment, and introduces risks to health information, specifically electronic health information, that you might not have considered before."\(^{39}\) There was no debate that these risks existed, but because hackers' methods are constantly evolving and vulnerabilities vary by organization, the precise ways in which they would manifest could not be fully predicted.

While the risk of major harm from cyberattacks was always present with the adoption of EHR, the calls to mitigate it became far more urgent once society felt the consequences. In January of 2015, insurance company Anthem, Inc. revealed that nearly 80 million patients' records had been stolen by hackers.\(^{40}\) A month later, President Obama spoke before the Cybersecurity and Consumer Protection Summit at Stanford University, stating that "it is one of the great paradoxes of our time that the very technologies that empower us to do great good can also be used to undermine us and inflict great harm."\(^{41}\) The following year brought about a series of executive orders, task forces, and legislative bills directed at bolstering cybersecurity across U.S. industries.

At a very broad level, the task of allaying cyberattacks has inspired two kinds of reactions. **On the one hand, there is the notion organizations should be required to take certain precautions when handling data.** This regulatory approach is best exemplified through the Health Insurance Portability and Accountability Act of 1996 (HIPAA), which established "a national set of security standards for protecting certain health information that is held or transferred in electronic form."\(^{42}\) **On the other hand, there is the notion that organizations should be given guidance to manage their unique security situations.** This idea is captured by the NIST Framework, which was created by the U.S. National Institute of Standards and Technology in 2014 as a "voluntary" means of "promoting the protection and resilience of critical infrastructure and other sectors important to the economy and national security."\(^{43}\) While the strategies were never intended to contradict one another, the "top-down" versus "bottom-up" dynamic makes adhering to both simultaneously a challenge. Indeed, HHS went as far as to create a "crosswalk" for HIPAA and the NIST Framework in 2016, seeking to help healthcare organizations understand how to apply them at once.\(^{44}\)

Cyberattacks and algorithmic bias share similar tensions. When these problems occur, they do so with pervasive and harmful externalities, unleashing their consequences on unsuspecting victims who are simply participating in contemporary society. This dynamic might indicate the need for sweeping intervention from policymakers. At the same time, the ways in which the harm might unfold are innumerable and constantly changing, making it difficult to pass laws governing the technology. Like those in the cybersecurity debate then, experts on machine learning are presented with a challenging question: If
there are thousands of applications of machine learning already in use, spanning nearly every industry and sector imaginable, is the “top-down” strategy (e.g., broad regulation) or the “bottom-up” one (e.g., organization-specific solutions) more feasible? Advocates for both approaches exist in the current landscape.

**Top-down:**

regulatory approaches to avoid the harms of ML

During a 2017 speech before the National Governors Association, Tesla founder Elon Musk made his position on artificial intelligence clear. “AI is a rare case where we need to be proactive about regulation instead of reactive,” he stated. “Because I think by the time we are reactive in AI regulation, it’s too late.” Others have echoed Musk’s sentiment, particularly in light of growing concerns over the massive influence of technology companies like Google and Facebook, but the exact form such laws should take remains unsettled.

In general, regulations can either adopt a “command and control” position (legislating standards and fining those who do not meet them) or an “incentive-based” approach (providing inducements that encourage socially responsible behavior). While incentive-based proposals might be appealing for their flexibility at first glance, it would be extremely difficult for a regulatory agency to induce “bias reduction” in this manner, since no universal definition of “bias” exists. Instead, proposals for regulating machine learning have focused on the “command and control” strategy, prescribing certain algorithmic features or safeguards in decision-making.

**Ensuring a “right to explanation”:** When the European Union implemented the General Data Protection Regulation (GDPR) in the summer of 2018, the law was mostly noted for its new standards around data privacy, but it also included an attempt to regulate artificially intelligent algorithms. Described as the “right to explanation,” the guiding principle is that individuals are entitled to understand how automated decision about them are made. If interpreted literally, this law could challenge the use of some “black-box” algorithms that are notoriously uninterpretable, despite having excellent predictive power. Some scholars caution, however, that the law’s requirement to provide “meaningful information about the logic of processing” is quite vague and may not actually be useful in mitigating harm. As an analogue, while the U.S. requires credit bureaus to provide consumers with access to their personal report, this increase in transparency does not translate to an understanding of how various items are weighted.

**Preserving human decision-making:** In the U.S. context, the heavily regulated health sector has taken some of the most proactive action in mitigating the risks of machine learning tools. In December of 2017, the Food and Drug Administration (FDA) released a statement on its updated approach to regulating clinical decision support (CDS) software, which sometimes use machine learning algorithms to diagnose conditions based on symptoms and identify possible treatments. According to the FDA, any automated recommendation produced by these technologies must be accompanied by its underlying logic, so that physicians can “independently review the basis for such recommendations.” A further implication of this regulation is that the models cannot be trained on any proprietary data, since it would obscure the logic to the provider.

**Evaluating developers, rather than software:** Because of the enormous potential benefits presented by digital health technologies, the FDA has also attempted to explore a regulatory framework that reduces the burden on innovators,
while simultaneously protecting patient safety. Launched as a pilot in early 2018, the Digital Health Software Precertification (Pre-Cert) Program evaluates the producers of machine learning tools, rather than examining each tool individually. In its initial statement, the agency explained the program:

Because software products can be adapted to respond to glitches, adverse events, and other safety concerns quickly, the FDA is working to establish a regulatory framework that is equally responsive when issues arise to help ensure consumers continue to have access to safe and effective products. In the Pre-Cert program, the FDA is proposing that software products from precertified companies would continue to meet the same safety and effectiveness standard that the agency expects for products that have followed the traditional path to market.50

The notion that cleared companies would be able to use their best judgment to flexibly address the challenges of machine learning models is appealing from a theoretical perspective. However, it is worth noting that very few existing medical device companies would qualify for this clearance, making its pragmatism questionable.51

In considering the tradeoffs of the top-down approaches to mitigating bias, it is clear that no universal benchmark or test can be applied to all machine learning algorithms. Instead, policymakers are left to focus on rules that promote explainability and accountability. These concepts are valuable for fostering the norm that machine learning tools cannot be left to their own devices without human input, which may be reasonable when dealing with life-or-death scenarios like healthcare or national security. Additionally, a top-down emphasis on interpretability may increase investment in technologies that garner human trust. According to one IBM survey of 5,000 businesses, 60% of executives who are interested in adopting artificial intelligence systems are concerned about opaqueness, viewing it as the main barrier to implementation.52

Regulation of artificial intelligence technology is opposed by diverse voices, with the most obvious rationale being that legislators will inevitably stifle innovation by getting involved. However, in terms of algorithmic bias, a more relevant concern is that regulation will take the place of more organic efforts to mitigate risk. Looking to the cybersecurity industry, some experts have observed that “attaining regulatory compliance is offering organizations a false sense of security on many levels.”53 To combat the emergence of a similar trend in machine learning, some organizations are advocating more “bottom-up” approaches to mitigating bias, with regulation becoming a possibility in the longer term.

The notion of defining algorithmic bias strategies as “bottom-up” is somewhat subjective. In contrast to regulations that are initiated by entities with enforcement power and applied universally to all relevant actors, these solutions are decentralized and uniquely defined by the respective needs of organizations. With this greater flexibility, “bottom-up” strategies are hard to meaningfully enforce, since there is no broad concept of what compliance looks like. In machine learning, these solutions are directed at empowering organizations to better identify incidents of bias in their unique applications.

fostering social awareness of developers: The lack of diversity in the machine learning workforce is often cited as a major contributor to the development of biased algorithms. Unrepresentative teams can result in technology that only caters to certain segments of the population, such as when Apple released an app for “comprehensive” access to health information that failed to include major health
problems faced by women. The ideal solution is to increase the diversity of perspectives present in Silicon Valley, but according to a study done by Montreal startup called Element AI, only 12% of today’s AI researchers are female. As efforts to recruit more women and minorities to the field are ongoing, some organizations are simultaneously working to make all ML developers more conscious of the social implications of algorithms. While tech companies are making some headway with training materials, such as a 60-minute module on fairness developed by Google, the majority of these efforts are occurring with universities integrating ethics courses into computer science curricula.

**decreasing the burden of finding bias:** In response to the notion that algorithmic bias often goes undetected by developers, technology companies are increasingly investing resources in open-source “toolkits” that check models for fairness. IBM’s AIF360 and Google’s What-if Tool, released one week apart from each other in September of 2018, essentially streamline the process of looking for disparities in a model’s predictive performance on subsets of the population defined by legally protected features. The tools offer functionality like letting users compare how a model’s predictions change if a particular predictor is swapped out for a similar variable. While there is utility in reducing the burdens of checking for algorithmic bias, organizations must still determine the relevant data segments.

**increasing data supply chain accountability:** The vast majority of machine learning research emphasizes the sophistication of new algorithms in predicting outcomes, with significantly less attention paid to questions of data collection and processing. As somewhat of an industry norm, researchers often opt to use the same very large training datasets, such as ImageNet, which includes 14 million hand-labeled pictures. However, bias is very present in this data, with more than 45% of images comes from the U.S., and only 3% coming from China and India combined. In the words of two Stanford professors, “This lack of geodiversity partly explains why computer vision algorithms label a photograph of a traditional US bride dressed in white as ‘bride’, ‘dress’, ‘woman’, ‘wedding’, but a photograph of a North Indian bride as ‘performance art’ and ‘costume.’” According to research centers like The AI Now Institute, organizations need to move away from the notion that pre-existing datasets can be universally recycled. Doing so requires allocating additional development costs, such as the human labor required for apt data collection.

The overarching goal of the “bottom-up” strategies for mitigating bias can be seen as efforts to shift the culture of machine learning toward being more mindful of social consequences. An obvious benefit of this approach is that it is informed by innovation and technological progress, rather than rigid legislation. However, the “bottom-up” strategies are notably devoid of enforcement mechanism, so organizations that do not take moral issue with contributing to systemic harms are left to their own devices. Further, it may be difficult to differentiate between those entities that are actually making concerted efforts toward mitigating unfairness, and those that are simply creating the optics of social responsibility.
<table>
<thead>
<tr>
<th>approach</th>
<th>example</th>
<th>overall goal</th>
<th>drawbacks</th>
</tr>
</thead>
</table>
| Demanding explainable  | GDPR                        | Protecting people from being subject to decisions that cannot be logically   | • Interpretable algorithms may have less predictive power  
| algorithms              |                             | reproduced by humans                                                          | • Algorithms may be considered proprietary information                                                              |
| Preserving human        | Medical technology          | Ensuring that appropriate information is being used to arrive at particular   | • Interpretable algorithms may have less predictive power  
| human oversight         |                             | decisions                                                                      | • No black box methods can be used                                                                                  |
| Evaluating organizations | FDA review process          | Expediting the process of evaluating algorithms, while also promoting public  | • Requires regulatory agencies to predict development of relevant technologies  
| for responsible         |                             | safety                                                                        | • Firms unlikely to be able to meet compliance standards                                                           |
| practices               |                             |                                                                              |                                                                                                                     |
| Educating developers    | Academia; technology       | Establishing a firm connection between the development of technology and its   | • Will likely need to take place over the course of a generation  
| to be socially aware    | ethicists                   | social consequences                                                            | • Also requires industry shift, with firms valuing ethics training in hiring                                        |
| Bias checks             | Technology companies; risk  | Automating the process of identifying common types of bias                    | • Organizations must know which types of bias should be checked for  
|                        | assessment firms            |                                                                              | • Involves testing for the symptoms of bias, without necessarily addressing underlying causes  
|                        |                             |                                                                              | • Requires established definition of fairness                                                                      |
| Supply chain            | NGO’s; technology ethicists | Disrupting institutional sources of bias in existing data and analytical      | • Complicates process of broad data-sharing efforts  
| accountability          |                             | systems                                                                      | • Might create additional burdens for organizations to collect and manage data                                    |
As is summarized in Table 2, significant debate remains as to the best way to systematically address algorithmic bias in machine learning applications. The speed at which algorithms process decisions, the dynamism of the risks, and the technological nuance involved in machine learning all contribute to the complexity of navigating proposed solutions. However, given the rate of human innovation over the past several decades, artificial intelligence is far from the first field to disrupt conventional processes for understanding and mitigating risks; the earlier discussion of cybersecurity is just one example. As such, a path forward can be informed by considering the features that algorithmic bias shares with other modern technological challenges, and the management processes that have proven effective in these spaces.

### Problems that share key complexities with algorithmic bias

The urgent need for public sector organizations to mitigate algorithmic bias comes from the disconcerting reality that the full consequences are unknowable. In traditional theories of management, risk is understood as having three parts – the event that could occur, the likelihood of it occurring, and the impact if it happens. Such a framing is useful for threats that are specific, like a natural disaster striking a town, because humans can imagine all of the ways in which harm might be inflicted. With this information in mind, statistical probabilities and cost-benefit analyses can be used to inform the appropriate mitigation tactic. When it comes to mitigating the harm caused by ambiguous threats, however, the basic three-pronged model of risk proves less actionable.

Philosophers, policymakers, and scientists have grown increasingly aware of the need for models that can preempt nebulous risks over the past several decades. In 1986, German sociologist Ulrich Beck famously wrote about the “risk society,” arguing that technology inevitably brings more dimensions of uncertainty than the Industrial Age, the vast majority of which are the product of human inventions. The so-called “manufactured risks” that occur as mankind pursues new innovations are particularly troubling because they result in “a denial of responsibility from organizations and individuals for creating these risks,” further complicating the preemption of emerging threats.

Beck’s theory notably gained traction because it was published at the same time as the Chernobyl disaster, which left large parts of Europe exposed to highly radioactive fallout after a nuclear reactor exploded outside of Kiev. However, as several Oxford University professors writing in *Journal of Cybersecurity* note, “Whilst the risks that Beck and Giddens describe in risk societies are inspired by advances in nuclear, chemical, and biomedical technologies, advances in information technology and cyberspace share the same characteristics.” Particularly, these modern challenges often involve humans failing to appreciate the magnitude of unintended impacts, electing to focus on short-term benefits over long-term costs, and spreading negative consequences through networked systems, to name a few. These broad characteristics are mapped in Table 3, depicting four types of modern hazards.

---

iv Anthony Giddens is a British sociologist who wrote on modernity contemporaneously to Ulrich Beck, developing similar ideas.
modernized strategies for managing risks

With the notion in mind that algorithmic bias shares many of the same problematic features of other modern technological challenges, the question becomes: **what has proven effective in these spaces?** It is important to recognize that society has not succeeded in fully mitigating the risks associated with advancements in nanotechnology, cyber networks, and nuclear power, but significant research has been dedicated to the
best possible courses of action. Most notably, scholars have increasingly called for hybrid mitigation tactics that combine top-down guidance with bottom-up analyses to best navigate ambiguous threats. The case of the U.S. government’s calls for context-dependent cybersecurity analyses via the NIST framework was previously discussed in this chapter, but similar trends can be observed in both nuclear energy and nanotechnology.63

In the context of nuclear accidents, the 2012 disaster at the Fukushima Daiichi Power Plant in Japan resulted in increased attention to the government’s strategy regarding risks. With a history of being dependent on fossil fuel imports for energy, the Japanese government has viewed nuclear energy as a means of increasing self-sufficiency since the 1960s. The irony of this position is not lost on scholars like psychiatrist Robert Jay Lifton, who studies the effects of conflict on human thought. “One may ask how it is possible that Japan, after its experience with the atomic bombings, could allow itself to draw so heavily on the same nuclear technology for the manufacture of about a third of its energy,” he notes. “But there was also a pattern of denial, cover-up and cozy bureaucratic collusion between industry and government.”64

According to a publication by the East-West Center at the University of Hawaii, the Japanese government was very aware of the public’s “nuclear allergy” in the aftermath of World War II, and intentionally devised top-down policy directives to overcome the fears of local politicians and their constituents. Following the events at Fukushima, decades of centralized control over the nuclear power debate was disrupted by increased activism on the part of survivors and a new trend of “citizen science” leading volunteers to participate in collecting and analyzing relevant data in their environment.65 Additionally, the infrastructural failures associated with attempting to evacuate and relocate thousands of people following the disaster highlighted the need for bottom-up input from a logistics perspective. In one nuclear risk-reduction framework published by the World Health Organization in 2015, there is a strong emphasis “that response to major disasters should include social mobilization and empowerment of local communities,” highlighting the limitations of centralized directives to address hazards.66

Turning to the case of nanotechnology, while the use of tiny particles has potential to revolutionize fields such as chemistry, biology, and medicine, scientists first became vocal about the risks to human health around the turn of the twenty-first century. One article published in Scientific American in 2008 described the use of nanoparticles made of clay, iron, and silver in common items like beer bottles, children’s formulas, and cleaning products, despite the limited information on the safety of such engineering. In the words of scientist Andrew Maynard at the Woodrow Wilson International Center for Scholars, “It all comes down to the need for more research. We can’t fly blind here. We need to know what’s going on.”67 While environmental and health advocates have called for top-down management from the U.S. FDA and other government agencies, the ambiguity surrounding the nascent technology’s impacts, feasibility, and applications complicate the regulatory debate.

Though frustrating to some, the public sector’s uncertainty regarding effective regulation of nanotechnology is not unwarranted – experts have yet to establish clear definitions for “nanoparticles”, insufficient data exists to weigh relative costs and risks, and the industry is developing far more rapidly than legislative processes allow.68 Accepting the limitations of conventional forms of government risk management, a growing scholarship has emerged that hazard mitigation must be approached from the bottom-up. For example, one group of science and law professors writing for Nanoethics advocate a “flexible evolutionary approach to risk ‘regulation’ especially in the immediate, near and medium terms,” calling for both “product stewardship and the professional ethics of researchers” as well as “subsidiarity and decentralization” to ensure that the most relevant stakeholders are contributing to discourse on norms and
procedures.\textsuperscript{69} From a practical standpoint, this strategy has taken the form of cities like Berkeley, California requiring all nanotech companies and university laboratories to report on their waste disposal procedures, with other local municipalities following suit.\textsuperscript{70} As robust data is collected through local initiatives, it is possible that the appropriate direction for federal oversight will become apparent.

Considering the respective discourses surrounding nuclear power, cybersecurity, and nanotechnology, an interesting pattern has emerged (see \textbf{Figure 8}). Initial excitement surrounding the potential benefits of new technologies result in rapid progress, leading society to temporarily ignore or minimize associated risks. Once the risks become more visible, calls for government to step in and protect the public via stringent regulations becomes the norm, as had been the case with traditional industrial hazards, such as automobiles or acid rain. Unfortunately, the significant ambiguity surrounding modern sociotechnical systems is not well managed with conventional theories of risk and tradeoffs. Because our society is dependent on technology and directed toward progress, it would never be realistic to simply abandon new innovations until their consequences are better understood.\textsuperscript{v} Instead, the acceptable course of action is to provide guidance to those who are closely positioned to the risks, so that they may use their context-specific knowledge to evaluate tradeoffs and react to ambiguity.

\textbf{Figure 8:} how society responds to modern technological risks

---

\textsuperscript{v} In general, the United States is thought to be less risk-averse regarding the regulation of science and technology than the European Union. According to The Institute on the Environment at the University of Minnesota, while the U.S. federal government “sets a very high bar for the proof of harm that must be demonstrated before regulatory action is taken,” the EU aligns more closely with the so-called “precautionary principle.” Under this framing, “When there is substantial, credible evidence of danger to human or environmental health, protective action should be taken despite continuing scientific uncertainty.”
iv. the choice
continuing with the inertia of the Big Data revolution

The above diagram highlights an important decision point for any society faced with the realization that a technology it has enthusiastically pursued may involve dramatic consequences. Essentially, progress can either be abandoned for fear that the risks are too great, or it can be continued with explicit attention to understanding and mitigating harms as they manifest. In the case of algorithmic bias, the pervasiveness of machine learning in virtually every sector and industry makes the former option impractical. Rather, the challenge for organizations seeking to implement artificially intelligent applications responsibly is to take on the task of engaging in “bottom-up” risk mitigation strategies, with the possibility of “top-down” frameworks being informed by these efforts down the road.

Earlier in this chapter, efforts to develop the social awareness of developers, bias checks, and analyses of supply chain accountability were all mentioned as proposed “bottom-up” strategies for mitigating unfairness in machine learning. The next chapter will examine practical means of implementing these tools in an auditing framework, particularly in the context of public sector organizations.

v. summary

- Algorithmic bias has been broadly identified as a problem that needs to be mitigated as machine learning becomes ubiquitous, but there is no established consensus on how this should be done
- Scientists and policymakers have spent significant time considering how to deal with modern technologies that present ambiguous risks (e.g., nuclear power, cyberspace, and nanotechnology), and the appropriate strategy for machine learning can be informed by these efforts
- Proposed solutions for mitigating algorithmic bias can broadly be divided into “top-down” strategies that focus on regulation and central management versus “bottom-up” strategies that attempt to bolster organizational capacity to deal with tradeoffs in a particular context
- “Bottom-up” strategies have emerged as a more flexible means of dealing with technological consequences, particularly when centralized authorities do not have adequate knowledge to intervene across all context
i. introduction

The previous chapters of this paper aim to, first, articulate the problem of algorithmic bias and, second, drive toward particular strategies for mitigating its harms. “Bottom-up” risk management is a logical framing for dealing with the inherent uncertainty of artificially intelligent data applications, but further elaboration is needed on how to carry it out. Algorithmic auditing, in short, is an effort to ensure that the context and purpose surrounding such applications directly inform evaluations of their utility and fairness. The goal of this chapter is to bring this strategy to life.

Stephen Hawking writes 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?”71 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. In short, 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.

In an ideal world, organizations would only deploy algorithms that are specifically designed to account for the nuances of a particular environment – but cost and efficiency concerns dictate otherwise. Mathematician Cathy O’Neil, for example, describes the use of credit reports to vet prospective job applicants: “Bad credit has grown to signal a host of sins and shortcomings that have nothing to do with paying bills,” she writes. “For certain applications, such a proxy might appear harmless...But if you’re looking for a job, there’s an excellent chance that a missed credit card payment or late fees on student loans could be working against you.”72

It is crucial to note that identifying the type of unfairness mentioned by O’Neil requires more than merely looking for computational or data errors. To quote another famous scientist, Nobel Prize winner Manfred Eigen once said, “A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant.”73 Accordingly, the concept of an “audit” discussed in this chapter extends beyond a mechanical examination of an application’s inputs and outputs, encompassing broader questions of ethical measurement, constituent engagement, and institutional use.
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. 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.

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% 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.” Machine learning applications, with their ambiguous consequences and diverse opportunities for bias to emerge, are similarly impossible to ever deem “100% 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.

**marrying “structure” and “judgement”**

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-judgement” 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 summarizes the findings of Dirsmith and Haskins, who use the metaphors of “mechanism” and “organism” to describe the debate: “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.

“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.
and weaknesses. On the one hand, the technological complexity of machine learning applications and the time-intensive task of reviewing data call for some degree of process standardization. On the other hand, the context-dependence of algorithms necessitates engagement with subject matter experts and flexible judgement calls. Only when the strategies are applied in tandem can an audit reflect both the technical precision and social awareness that is needed to root out systematic unfairness.

**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 esteemed statistician Leo Breiman once wrote, “Predictive accuracy on test sets is the criterion for how good the model is.” While this focus on reducing test error rates 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.” 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 race, and notices that the model is more likely to predict a positive outcome for white individuals than other racial groups. (Note: this could happen even if race is not included as a direct input into the model.) 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 white people. To improve the error rate for the non-white subgroup, the developer might mechanically adjust some of the model’s parameters so that it performs better across all racial 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, meaning 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.

**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

---

VI The test error rate of a model represents its performance on data that was not included in the training set.
“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.” 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.

**iii. the practice**

deriving a framework for ML audits

To summarize how the abovementioned 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 judgement 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.” 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.**

Chapter one of this paper introduced the six technical steps involved in creating a machine learning application (e.g., training data identification, pre-processing, model selection and training, accuracy evaluation, insights presentation, and feedback incorporation), but for the purposes of conducting an audit, the non-technical components must also be taken into account (e.g., model purpose and constituent remediation). The discussion below outlines how these eight aspects of a model might be evaluated.
**Step 1 - Model Purpose:** What is the organization seeking to approximate with an algorithm, and why?

Prior to even reviewing the substance of a model, it is important for an auditor to obtain an understanding of why an organization opted to create it in the first place. This understanding might be developed through reviewing a project statement, as well as by interviewing people involved in the process. Whatever the goal of an application is, part of an auditor’s job is to ensure that it is clearly defined. In addition to examining how an algorithm is described in internal documents, an auditor should pay close attention to any discrepancies in how it is verbally described across an organization. For example, if management seems to have a different understanding of the model’s purpose than the technical team that built it, there may be cause for concern regarding deployment.

**Step 2 - Training Data Identification:** Does the available training data map to the target population?

An auditor must work to develop a clear sense of the relationship between the observations being used to train a model and the ultimate population of interest. While examining the technical nuances of algorithms is important, sampling problems can make even the most sophisticated of models ineffective. The auditor will want to review how various demographic groups are represented in the training data. Two types of sampling bias are particularly worth keeping in mind – first, a data collection process that fails to capture certain observations (e.g., a survey given over the Internet missed rural and elderly populations) and second, a precedent that excluded certain groups (e.g., a firm using historical resume data to inform hiring decisions has previously employed very few women). If such gaps in the training data do exist, the auditor must ensure that they are accounted for in the model design, perhaps through weighting, ensemble methods, or synthetic data generation. Determining the noteworthy demographic segments will also allow the auditor to compare the model’s respective performance across them.

Additionally, with an understanding of a model’s purpose in mind, an auditor must determine whether the available data logically map to the organization’s goals. As discussed in chapter one, data are “proxies” for genuine concepts of interest, and the appropriateness of a proxy depends on how the information it carries will be used. Ideally, the data inputs for a model are documented in a so-called “data dictionary,” which defines the metrics, their relevance to the population in question, and the data’s source. As a starting point, an auditor must look for two things in terms of data – variables that are poor proxies (e.g., using SAT scores as a measure of intelligence) and concepts (omitted variables) that are inevitably important but missing from the model (e.g., failing to include gender in a model that predicts a person’s height).

**Step 3 - Pre-Processing:** What assumptions are made in the processing of the data?

Beyond examining the contents of the “data dictionary,” an auditor will want to understand which variables were used to inform the cleaning and structuring of the training data. For example, if a developer opted to omit observations that have “N/A” listed for a particular metric, the model may perform differently on that subgroup. While such decisions are often essential steps in the machine learning development process, they must be clearly documented to appreciate the potential for bias.

---

**VII** See page 5, footnote III.
**Step 4 - Model Selection & Training:** How does the chosen model make sense of the information it receives to generate a prediction?

An auditor should seek to understand the basic strategies a model uses to generate predictions, including how input variables are narrowed down and how outliers are normalized. For example, in terms of feature selection, while some ML algorithms rely on humans to identify the relevant variables for analysis, other algorithms automate this process. It may be the case that an automated feature selection process performs well on the training data, but contradicts a subject matter expert’s (SME’s) opinions on the relevant variables for the target population. An auditor must work to verify that an algorithm is actually incorporating information that has been deemed mission critical.

**Step 5 - Accuracy Evaluation:** How are the tradeoffs around optimizing for predictive power handled?

Regardless of the exact algorithm a machine learning model uses, an audit should unpack how definitions of “accuracy” informed that decision. Namely, while the overall test error rate of a model is important, an organization that is responsibly deploying a machine learning application must consider the relative risks associated with false negative and false positive rates. As such, an auditor must consider the possible harms that could be inflicted on different constituencies if the predictions are wrong, and validate that the model is optimized to mitigate the worst of these harms.

**Step 6 - Insights Presentation:** What information is the model’s end-user ultimately given?

After reviewing the technical design of a model, an auditor must also examine how the predictions it generates are ultimately presented. This would involve reviewing whatever information the “user” (e.g., the decision-maker who will act on the model’s findings) receives, including digital interfaces and written instructions on how to interpret results. An auditor should be particularly aware of any language that falsely implies a causative link between the input variables and the prediction, since this could lead a user to be overly confident in the findings.

**Step 7 - Feedback Incorporation:** Is the model implemented in a manner that incorporates feedback?

To ensure that flawed models do not become entrenched in an institution, it is crucial that an auditor examines the process by which a model can be updated to reflect new information. In the absence of such measures, an algorithm could perpetuate cycles of bias through feedback loops – for example, an algorithm for creditworthiness might reject applicants from underrepresented backgrounds, thereby creating a bottleneck where no data exists to support granting loans to diverse applicants. Auditors must consider the potential ways such reinforcing cycles might emerge, and explore what measures exist to correct them.

**Step 8 - Constituent Remediation:** Do procedures to promote transparency and/or remediation exist?

Though outside the scope of the technical model, an auditor must assess the nature in which people working with the tool are using the insights it generates. Because correlation does not imply causation, it is crucial to ensure that algorithms are not being used to hand down irreversible decisions that could alter the course of people’s lives. Auditors must ensure that organization using machine learning maintain transparency with their constituents, including a process for contesting the results of a model. An organization should have these procedures explicitly documented and publicly available.
It is exceedingly important to highlight that the above steps (or any set of steps) are not exhaustive of all of the possible ways in which an auditor might evaluate bias. Still, building off the notion that bias may potentially enter the ML pipeline at any phase of development, it logically follows that an auditor should examine what efforts were taken to minimize this threat throughout the process. As the practice of auditing becomes more common, it may become possible to collect data on common pitfalls across specific industries and contexts, allowing further development of best practices for risk mitigation.
Since part of an auditor’s job is to reconstruct how an organization handles the issue of bias across a model, the following case helps demonstrate what an examination of “responsible” ML development might look like.

As a state with relatively high rates of recidivism, the Iowa Department of Corrections collects extensive data on whether or not an offender is re-incarcerated within three years of being released. According to the Department’s website, recidivism is used “as an indicator on whether strategies are reducing offenders relapse into criminal behavior.” With information now available on offenders released in the years 2010 through 2014, a machine learning application is introduced to help Iowa policymakers predict which individuals are most likely to recidivate following their release. An auditor is introduced to evaluate the extent to which bias considerations were addressed throughout the process.

1. **Model Purpose**: The Iowa Department of Correction wants to know if a person they release from prison will commit another crime within three years, but this answer is not sufficient to evaluate the ethics of an algorithm. Rather, it is crucial to consider the behavior that will be informed by the algorithm’s outputs. For example, if the Department plans to use this list for punitive measures, such as increasing surveillance on the predicted repeat offenders, then serious questions of fairness emerge. On the other hand, if the Department plans to use the list to increase the amount of support these individuals are given in the form of measures like job training and placement, then an algorithm’s outputs are less problematic.

The perhaps most problematic mentality for an organization to have around questions of an algorithm’s purpose is the belief that it can be used to guide many different policies simultaneously. Put another way – is the burden of proof required for profiling and surveilling someone the same as that required for helping them find a job? If not, the same algorithm should not inform both of these actions.

2. **Training Data Identification**: Iowa’s open data portal reports the information about whether or not individuals recidivated in a data file with 19 other variables included. It is important to keep in mind that the existence of these 19 variables on the same spreadsheet as the outcome of interest does not imply causation, or even correlation, and that significant information is not captured in the document. At a high level, several of the included variables seem likely to have an impact on whether or not an individual recidivates, such as their age, gender, and initial criminal conviction.

It is also extremely important to look beyond the immediate data for factors that could better inform recidivism predictions. For example, the specific experiences an inmate has while incarcerated may prove useful, such as if they are released early for good behavior or involved in any episodes of violence. Corrections departments likely collect this data in a systematic manner, making it worth exploring as an input to an algorithm.
3. **pre-processing:** The Department of Corrections data inevitably incorporates some judgment calls about how to categorize people. For example, rather than providing the precise age at which an individual was incarcerated, age is treated as a categorical variable. While a developer might not see any problem with simply grouping individuals in 10 year increments, this practice might obfuscate important information if 36- to 40-year-olds are actually significantly less likely to recidivate than 30- to 35-year-olds.

Additionally, the output variable for the Department of Corrections training data is a binary of “recidivated within three years” versus “did not recidivate within three years.” As such, any model trained on this data can only make similarly simple predictions. Iowa’s Department of Corrections has recently engaged in diverse programs to reduce rates of reoffending for individuals who cannot find a job after their release. With this context in mind, a more informative output variable in the training data might have three options, such as “recidivated while unemployed,” “recidivated while employed,” and “did not recidivate.” Conversations with SME’s can shed light on whether additional layers of complexity should be accounted for in the initial design of the data to make the model’s predictions more actionable.

4. **model selection & training:** Like any model, the Iowa Corrections model is a simplified representation of reality, meaning it cannot capture everything about why an individual recidivates. However, in this particular context, the Department has concluded that certain mental health conditions significantly increase the odds that a person is re-incarcerated. While some data on inmates’ mental health diagnoses exists, this variable is not consistently available for all prisoners in the training data, making it statistically more likely to be excluded from the analysis. To avoid such communication lapses, developers should share clear documentation with SME’s, explaining precisely which inputs are ultimately contributing to a model’s estimates.

5. **accuracy evaluation:** Because a model cannot be optimized for every possible task, the process of building and testing various algorithms needs to be informed by the goal at hand. Conventional machine learning wisdom focuses on minimizing the “test error” rate for the entire population, but this summary metric can detract from specific conversations about what the model is actually getting right and wrong.

As a simple example, using the Department of Corrections’ data for prisoners released in 2011, 2012, and 2013 for training, two hypothetical models produced the following results when tested on 2014 data (see Figure 9 on page 36). While the overall accuracy rates of these models are very similar, it is worth noting that the sensitivity of the first model is 10 percentage points lower than the sensitivity of the second model. At the same time, the second model has a lower specificity rate than the first.
Put another way, if every individual released from jail in 2014 had received a label of “will recidivate” or “will not recidivate,” the first model would have “caught” 70% of repeat offenders, while the second model would have “caught” 80%. If the Department was hoping to maximize the number of at-risk individuals who receive extra help finding a job, the second model would have been preferable. On the other hand, if resource constraints were the primary issue, the model could have been optimized to help the Department predict which prisoners they could release without actively addressing recidivism concerns. In this case, the first model would have been right about which prisoners were not going to end up back in jail 80% of the time, while the second model would have been correct 72% of the time. Regardless of what the Department’s priorities for a model are, it is crucial to establish them in order to inform frank conversations about the tradeoffs involved in predicting particular outcomes.

### Figure 9:
comparing performance tradeoffs of two models

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Actual</th>
<th>Recidivated</th>
<th>Did Not Recidivate</th>
<th>Positive Pred. Value</th>
<th>Negative Pred. Value</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro: higher specificity (fewer false positives, so it is more likely that a person labelled “likely to recidivate” actually recidivates)</td>
<td></td>
<td>1131</td>
<td>624</td>
<td>0.64</td>
<td>0.83</td>
<td>0.70</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>Con: lower sensitivity (more false negatives)</td>
<td></td>
<td>474</td>
<td>2458</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 2</th>
<th>Actual</th>
<th>Recidivated</th>
<th>Did Not Recidivate</th>
<th>Positive Pred. Value</th>
<th>Negative Pred. Value</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro: higher sensitivity (fewer false negatives, so there is less of a chance that a person who is likely to recidivate falls through the cracks)</td>
<td></td>
<td>1289</td>
<td>851</td>
<td>0.80</td>
<td>0.60</td>
<td>0.80</td>
<td>0.72</td>
<td>0.77</td>
</tr>
<tr>
<td>Con: lower specificity (more false positives)</td>
<td></td>
<td>316</td>
<td>2231</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

thought exercise for the auditor
What is the most important thing for the model to get “right?”

What would be the worst things for the model to get “wrong?”
6. **Insights presentation:** Even if the Department of Corrections model addresses many of the technical sources of bias, problems will arise if the estimates are not presented judiciously. For example, an interface could be designed in one of two ways – an ominous mugshot and bold label of “LIKELY TO RECIDIVATE” with no further information, or an imageless display that details some of the individual’s risk factors for re-incarceration. The precise information that a developer should opt to provide alongside a model’s predictions should obviously be informed by the audience’s role.

7. **Feedback incorporation:** If the Department of Corrections deploys an algorithm that fails to identify the individuals who are most at-risk for recidivism, it would be very easy for a problematic feedback loop to emerge. Consider what happens if a group of former offenders are wrongly identified as future “recidivaters,” leading them to receive extra support to help them return to society. Three years later, these individuals do not recidivate, leading agency officials to conclude that the intervention worked and that the same policy should be used moving forward. Very quickly, the former offenders who are most in-need of support are cut out of the system’s resources.

Avoiding feedback loops in institutional settings can be very difficult, but extra care needs to be taken to mitigate this harm, particularly when the stakes are as high as helping individuals avoid future imprisonment. A randomized control trial is one possible strategy for achieving this goal. For example, the Department of Corrections might intentionally select some individuals who the algorithm labels “not likely to recidivate” to include in its anti-recidivism programming. As such, three years into the future, the agency can evaluate the program’s relative effectiveness on individuals who were and were not identified as high-risk for re-imprisonment and use this information to make ongoing improvements to its predictive capacity.

8. **Constituent remediation:** In light of the negative attention garnered by algorithms used by justice departments and law enforcement agencies across the U.S., the Department of Corrections wants to be very forthcoming with the public about the nature of this model. For example, it is particularly important to explain that the model is optimized to provide resources to at-risk individuals, rather than to be used as a predictive policing tool. Additionally, in cases where an inmate who receives a prediction of “not likely to recidivate” would still like access to any extra support services granted to the “likely to recidivate” population, a formal process for attorneys to apply on behalf of their client should exist.
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.

**clear articulation of goals**

As the media continues to popularize sensationalized anecdotes about the negative impacts of algorithms, stakeholders are increasingly calling for researchers to shift to “explainable” types of AI. However, in the words of technologist David Weinberger, “Keeping AI simple enough to be explicable can forestall garnering the full value possible from unhobbled AI.” To avoid making this sacrifice, the Senior Researcher at the Berkman Klein Center for Internet and Society proposes an alternative: “Accept that we’re not always going to be able to understand our machine’s ‘thinking.’ Instead, use our existing policy-making processes...to decide what we want these systems optimized for. Measure the results. Fix the systems when they don’t hit their marks. Celebrate and improve them when they do.” Of course, such an approach is only possible if organizations deploying ML tools exercise a great deal of intentionality around their technology’s goals.

In virtually every industry, project teams notoriously underestimate the importance of a specific and consistent objective. Zig Zaglar, a twentieth-century icon in business acumen and motivational speaking, built his brand on quotes like “a goal properly set is halfway reached” and “lack of direction, not lack of time, is the problem.” Half a century later, such lessons still resonate with technologists like Andrew Ng, an esteemed Stanford Professor and co-founder of Coursera, who teaches the importance of picking a measurable metric at the outset of any data science project to inform all decisions. As AI engineer and author, Tirthajyoti Sarkar notes, “Advanced techniques of artificial intelligence or machine learning may be able to guide businesses toward a better optimal solution at a faster clip, but they must confront and solve the same (or more complex) optimization problems as before.”

Interestingly, local government teams that have been successful in developing ML tools often call out this process of goal definition as the driving force behind their positive outcomes. In the words of Joy Bunaguro, the former chief data officer for San Francisco, “The key barrier to data science is good questions.” Through several interviews with public servants in roles similar to Bunaguro’s, writer Ben Green summarizes their attitudes: “Improving operations with data often hinges not on developing a fancy algorithm but on thoughtfully implementing an algorithm to serve the precise needs of municipal staff.” When an organization can clearly articulate the intentions behind a machine learning application, the process of auditing can be better appreciated as another way of optimizing toward this goal.
interdisciplinary inputs

While goal clarity is crucial for getting an organization and an auditor “on the same team,” not all stated missions naturally facilitate collaboration. As Philip Virgo, the former vice president of the Council of the British Computer Society, notes, “Failure to consult those in the front line of delivery, or in receipt of services, is endemic among those planning new policy initiatives or changes to existing systems.” Former chief analytics officer for New York, Amen Ra Mashariki, specifically highlights the importance of incorporating expertise from non-technical civil servants for data science applications. “You come in with your fancy machine learning algorithm in your pocket,” he says. “But what’s always going to be your ace in the hole is the knowledge of the people who actually do the real work.” If internal disagreements about the utility of an ML model exist, it will be much more difficult for an auditor to objectively assess the development pipeline, thereby creating more room for suspicion and distrust. On the other hand, if an auditor comes into an organization where a model’s development has been informed by diverse inputs and a sense of teamwork, this spirit of collaboration is likely to continue during the evaluation process.

The notion that subject matter experts need to be robustly incorporated in ML tools is a common refrain of technologists, but the actual process of collaborating can be challenging. Bradford Cross, who works on AI startups, strongly emphasizes the importance of balancing different types of expertise throughout the development lifecycle. “Teams that manage to combine the subject matter and technical expertise are able to model the domain richly and drive innovation that comes from thinking outside the box by understanding what the box is,” he writes. However, “teams that come with a domain-first approach tend to get stuck inside the box, and teams that come with a tech-first tend to get stuck out in left field…if you’re unable to set the joint domain-tech DNA early, then one side dominates.” One report from McKinsey describes a multidisciplinary lens to ML as a means of building tools so that the relevant players will adopt them. Strategies might include “bringing target users into a model’s development process from the beginning,” “soliciting frequent reviews and input along the way,” or “designing a straightforward way to deliver and consume the model’s insights.” For an auditor, evidence of these engagements also enrich the evaluation process.

socially-responsible mindset

A final way in which an organization can foster productive attitudes toward algorithmic auditing relates to broader conceptions of the risks presented by modern innovations. In chapter two of this paper, machine learning models are compared to other technologies that involve wide-ranging consequences, such as nuclear energy and nanotechnology. A defining feature of these inventions is the fact that their risks can never be fully controlled, regardless of who is designing or implementing them. In the realm of artificial intelligence, the uncertainty around costs has largely led many innovators to argue that scientific progress should be pursued in the immediate term, with hazards being handled later. Such a mindset is notably at odds with the cautiousness of auditing, which seeks to prevent harm from unfolding wherever it can be predicted. Therefore, an organization can make itself more “audit-ready” by adopting a different attitude, instead viewing technological advancement and social risk mitigation as inextricably linked.

Thus far in history, scholars have largely been divided into two camps regarding how society might handle an amorphous problem like algorithmic bias – namely, either continue ahead with AI innovations and hope for the best, or halt progress until better tools for controlling dangers are available. The public sector’s use of machine learning tools presents an opportunity for government to serve as an example of a third approach
- continue exploring the benefits of machine learning where possible, but do so in a manner that thinks through risks. Algorithmic auditing is symbolic of such a strategy, marrying innovation with responsibility and future promises with present-day pragmatism. If an organization can acknowledge auditing as the inevitable social duty that comes with deploying complex technologies, the process will prove more worthwhile for everyone.

v. summary

- Because a model can only be evaluated in relation to the reality it claims to represent, algorithmic bias can be best identified using context-specific evaluations
- Auditing is a discipline that combines structured frameworks with the flexibility for professionals to exercise skepticism and ask additional questions as needed
- To carry out an algorithmic audit, an auditor must systematically review how bias considerations were taken into account in every step of the model development process
- Rather than viewing auditing as punitive, public sector organizations should seek to engage with the process as a means of protecting society against the risks of machine learning, while also exploring the technology’s benefits
- Clear articulation of ML goals and diverse inputs from relevant experts are key for allowing the auditor and the organization to collaborate on a shared mission
chapter four
recommendations & conclusion

“We can only see a short distance ahead, but we can see plenty there that needs to be done.” - Alan Turing (1912-1954)

In light of the speed at which new methods for data analysis are being developed, the notion of keeping up with “best practices” can be overwhelming for even the most technologically sophisticated organizations. The goal of this paper, however, has been to demonstrate that algorithmic auditing is more than a “check the box” compliance measure. Rather, the process codifies fundamental questions about Big Data usage, ensuring that organizations engage comprehensively with issues of bias.

Chapter 3 of this paper introduced a technical framework for an algorithmic audit, but a variety of broader efforts are required for an organization to implement the practice. Algorithmic audits require an institutional awareness that the benefits of Big Data analysis cannot be achieved without proper attention to the risks. The following recommendations are aimed at a public sector organization seeking to cultivate this mentality – and ultimately mitigate bias through robust audits.

recommendation one: signaling the utility of algorithmic auditing

It is important for an organization’s leaders to socialize the concept of algorithmic auditing as a pivot toward more socially responsible data analysis, rather than introducing it as a means of identifying inappropriate behaviors. Noteworthy cases of problematic government algorithms, like those mentioned at the end of Chapter 1, provide significant justification for why public sector entities need to diligently investigate issues of bias. Additionally, to encourage transparent communication throughout the process, leadership should emphasize that bias does not emerge from a bad decision by a single developer, but through institutional failures that must be collectively addressed.

viii Though the process of identifying individuals to conduct algorithmic audits is somewhat beyond the scope of this paper, it is worth noting that organizations may rely on a combination of internal and external parties to evaluate models. The primary goal is to approach the assessment with relatively “fresh eyes,” which can often be done by internal teams who were not involved in the original development process. In cases where a model is automating a key organizational function, it may be desirable to introduce an external auditor, though the professional industry is in its very early stages.
**Recommendation Two: Prioritizing High-Impact Algorithms**

While automated data tools may be widespread across a government institution, public sector organizations should prioritize auditing any applications that are currently used to make decisions with direct impacts on people’s lives. For example, is an algorithm being used to explicitly determine if someone is eligible for social services? An initial exercise might involve the individuals who developed this algorithm reviewing the development pipeline outlined in Chapter 3, noting their thoughts on where the process was most susceptible to bias. At the same time, other internal developers who did not work on the algorithm can “red team” this exercise, providing a more detached perspective. Depending on the findings of these exercises and the significance of the model’s impact on human lives, it might then be appropriate to engage with an external auditor.

**Recommendation Three: Translating Audit Results to Action Steps**

Though beyond the scope of the technical auditing framework introduced in this paper, organizations that audit their algorithms must then work to make the findings actionable. Particularly in cases where a model’s outputs are highly consequential, this process may involve answering challenging questions about steps to take in the short versus long term. For example, does the algorithm need to be taken offline until bias concerns are adjudicated? Do better remediation procedures need to be introduced? After addressing these initial decisions, the audit results might imply a need to address procedural factors, like integrating SME approvals to the development pipeline or refining data processing methods.

**Recommendation Four: Building Future Algorithms with an Auditor’s Perspective in Mind**

Chapter 3 of this paper outlined eight questions that an auditor might pose across the machine learning pipeline, but these questions can easily be used to inform a conscientious development process. For example, if a team working on a model recognizes that an auditor is likely to ask, “What is the organization seeking to approximate with an algorithm, and why?”, they will be encouraged to precisely define the tool’s goals at the outset. Developers should also think through the “thought exercises for the auditor” included in the discussion of a fictional Department of Corrections model in Chapter 3, applying them to their particular context. As an ancillary benefit, the creativity required for answering these questions might lead to innovative thinking around factors like viable data sources.

**Recommendation Five: Establishing a Schedule for Recurring Audits**

Because data analysis in the public sector is dynamic, with new observations and trends constantly emerging, algorithmic auditing must be conducted on an ongoing basis. Organization should define a cadence for auditing, including an evaluation that takes place fairly soon after an application goes live. The frequency of audits should be informed by the relative significance of algorithms in shaping decision-making, as well as the resource constraints of the institution housing them. However, in cases where an organization is unsure that it will be able to dedicate sufficient time to ongoing maintenance of a model, it is worth considering if automated algorithms are appropriate.
Because AI tools process data so quickly, left unmonitored, they can very quickly produce significant consequences.

**Recommendation six: Synthesizing findings and looking for trends**

Chapter 2 of this paper emphasized that the ambiguity of algorithmic bias makes it very difficult to manage through traditional “top-down” regulatory frameworks. However, as an organization audits its models and collects information on trends, it might be empowered to develop an institution-specific framework to inform all future model development. For example, it might be the case that multiple audits of a particular organization reveal consistent failures to validate feature selection with SME’s, since these experts sit in a different part of the government. In this situation, an internal procedure for facilitating (and even forcing) these conversations could be developed.

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. While this paper has highlighted the potential for these institutions to create socially responsible systems around automated tools, it also demonstrated that irresponsible behaviors are hidden in models with alarming ease. Big Data applications have unequivocally emerged as forces influencing almost every aspect of life, and if the public sector does not work to shape them to mitigate systemic harms, it is unclear who will.
references

chapter one


Ibid.


Ibid.


Ibid.


Ibid.


**Chapter Two**


48 Ibid.


Ibid.


chapter three


82 3-Year Recidivism for Offenders Released from Prison in Iowa. (2019, January 23). Retrieved from https://data.iowa.gov/Public-Safety/3-Year-Recidivism-for-Offenders-Released-from-Prison/mw8r-vqy4


85 Ibid.


88 Ibid.


