MEREDITH BROUSSARD

How to Investigate an Algorithm

Algorithmic auditing has the potential to decrease bias and prevent or fix the harms caused by artificial intelligence.

At the end of 2021, Sujin Kim was a senior at the University of Michigan. Eager to follow up her undergraduate political science studies with a PhD, Kim was applying to 15 schools on the East and West Coasts. Her applications were in, she was on top of the deadlines, and the only thing that remained was to take the GRE, the standardized test required by many graduate schools. She scheduled it carefully—the scores were supposed to be returned in 10 to 15 days, so she picked a test date in early November, well before her December 1 application deadlines. The GRE would be administered remotely because the COVID-19 pandemic was still raging in Michigan. She reserved a study room on campus so she could have a quiet spot to take the test without distractions. She knew that the test would be administered using proctoring software called ProctorU, and she knew the software would be finicky.

ProctorU, like ExamSoft or Proctorio, is one of the many AI-based remote proctoring software systems on the market. But unlike its competitors, ProctorU has backed off from trying to use AI without human intervention. "We believe that only a human can best determine whether test-taker behavior is suspicious or violates test rules," said Scott McFarland, CEO of ProctorU, in a May 2021 press release. "Depending exclusively on AI and outside review can lead to mistakes or incorrect conclusions as well as create other problems."

When the day came to take the test, everything seemed ordinary at first. Kim checked in using ProctorU and held up her ID to the camera so the software could take a picture of her face and her ID. The software took control of her desktop, and she made sure the camera could see the door behind her, as instructed, so the remote proctor could see if anyone came into the room.

Problems arose quickly. "I got through the first section and the connection dropped," Kim said. She restarted the program and checked in again with the photo verification and began to write the required essay portion of the test. The connection dropped again. She tried to log back in, this time without success. Eventually, she connected with a human proctor who seemed flustered. She seemed to feel bad that Kim couldn't take her exam. The proctor took over Kim's desktop and checked her in again, and this time it worked. Kim finished the exam and the program issued her a set of preliminary scores, which would be verified by the Educational Testing Service (ETS), the organization that administers the GRE, and she would get her final results in 10 to 15 days.

Ten days later, the scores had not arrived. Friends who had taken the test at the same time had received their scores in a week. On the fifteenth day, nothing arrived. "I stayed up till midnight to see if they would come out and they didn't," Kim recalled. "I didn't get an email. When I went into my portal, it just said, 'scores unavailable.' I emailed

ETS. I didn't get anything. Then, I spent four days on the phone—or trying to get through to them on the phone, on hold for literally hours—trying to figure out what was happening because no one would tell me."

Eventually, she connected with an agent who said there was a security hold on her scores. "She was working at home; I could hear her kid crying in the background," Kim said. "I asked her when I could expect the scores by. She said it depends on what the hold is for. I said, 'OK, can I have documentation or something to send to the schools so I can tell them I'm not lying about my scores being delayed?' She said, 'It looks like there was a problem with your pictures, your ID photo, so that's usually two to eight weeks.' She said not to worry, the schools would give me extensions. She told me to just call the schools I was applying to and ask. Then she hung up."

Kim blamed herself. Maybe she had closed her eyes or been out of the frame in one of the verification photos? Could that be the reason for the security hold? She went on Reddit and the ProctorU website to look up security holds. "I assumed it was a person doing the verification," she said. "It's actually facial recognition. They use a biometric facial ID, and if the score is above a certain threshold match you can proceed. That's how I found out it was a tech problem."

"A web of serious problems"

Kim's story could have ended there: unable to get final scores, contacting 15 different graduate programs to try to get extensions on her application deadlines. But she was not easily cowed. And she had paid attention in school. Kim worked as a research assistant for Shobita Parthasarathy and Molly Kleinman at the University of Michigan's Ford Science, Technology, and Public Policy program. Parthasarathy, the program director, is a professor of public policy; Kleinman is the program's managing director. The lab had published a study the year before called "Cameras in the Classroom" which criticized facial recognition in education and called for a ban on its use. "Not only is the [facial recognition] technology not suited to security purposes, but it also creates a web of serious problems beyond racial discrimination, including normalizing surveillance and eroding privacy, institutionalizing inaccuracy and creating false data on school life, commodifying data and marginalizing nonconforming students," Parthasarathy wrote.

Kim told Kleinman about the situation, who tweeted about it. The thread went viral in ethical AI circles. People called the situation Kafkaesque, Orwellian. The professoriate was outraged. "This student understood exactly what was happening and why it was wrong because she happens to work with a research team that literally wrote the report on why facial recognition in education should be banned," wrote Kleinman. "Imagine all the students who don't have that knowledge or access."

Just before the December 1 deadline (and after Kleinman's thread), ETS resolved its security hold and released Kim's final scores.

When I spoke with Kim about the facial recognition fail that could have wrecked her chances of getting into graduate school, she was keenly aware of how institutional privilege and educational capital had helped her overcome this case of algorithmic bias. "I find it hard to believe it's just a coincidence. I'm sure the scores coming out had a lot to do with Molly advocating for me," Kim told me.

We debated what could have happened inside the software. She thought the problem was a check-in photo that had her face out of frame. I thought the problem was racism in the facial recognition software. In a National Institute of Standards and Technology study, facial recognition technology was 10 to 100 times more likely to inaccurately identify the face of a Black or East Asian person compared to that of a white person. There is also a case in New Zealand where a man of Asian descent was unable to get his passport photo automatically approved because an AI program repeatedly registered his eyes as being closed. Both explanations were plausible.

Sujin Kim's experience offers a starting point for sorting through the many ways that ideas about race, gender, and ability are embedded in today's technology. Digital technology is wonderful and world-changing; it is also racist, sexist, and ableist. For many years, developers and marketers have focused on the positives about technology, pretending that the problems are only glitches. Calling something a glitch means it's a temporary blip, something unexpected but inconsequential. A glitch can be fixed. The biases embedded in technology are more than mere glitches; they're baked in from the beginning. They are structural biases, and they can't be addressed with a quick code update. It's time to address this issue head-on, unflinchingly, taking advantage of everything we know about culture and how the biases of the real world take shape inside our computational systems. Only then can the slow, painstaking process of accountability and remediation begin.

Algorithmic auditing

Sujin Kim didn't have all the information she would have needed to understand the complete picture of what happened with her GRE test—but she knew enough to recognize that something wasn't right. Kim's self-advocacy gives me hope. The repercussions of algorithmic failures fall hardest on already marginalized communities; hence the urgency to address these issues. But people who experience these failures having the understanding and agency to call them out is an important step in the right direction.

Algorithmic auditing is a developing area of public interest technology, which aims to get more talented people working on projects to serve the public good. It shows great promise for decreasing bias and fixing or preventing algorithmic harms such as the kind experienced by Sujin Kim. Algorithmic auditing is the process of examining an algorithm for bias or unfairness, then evaluating and revising it to make it better. Rarely are the problems solved in one shot. But auditing is the best tool we have right now.

Ideally, algorithmic auditing will be integrated into the compliance process for a range of industries. It hasn't been adopted widely in the United States yet, but the European Union's regulatory progress like General Data Protection Regulation (GDPR) and the EU's proposed AI legislation suggest that compliance for AI is coming soon. As far as the question of why audit at all, I think it is best articulated by auditing expert Inioluwa Deborah Raji, who tweeted: "We can't keep regulating AI as if it works. Most policy interventions start with the assumption that the technology lives up to its claims of performance, but policymakers & critical scholars need to stop falling for the corporate hype and should scrutinize these claims more."

Until recently, software developers have not paid enough attention to ensuring their algorithms operate within existing laws. Auditing is a way to make sure that the public interest is being preserved in and around algorithms. Generally, there are two ways of auditing: bespoke and automated. In bespoke auditing, the audit is done by hand: auditors break down the process, read code, run statistical tests, look at training data, write documents, and have meetings. In automated auditing, they do the same thing, plus use additional technical components to analyze the performance of a system on the level of code, using a platform or repeated tests. There are more thresholds in the automated method.

One of the people leading the field in algorithmic auditing is Cathy O'Neil, the author of *Weapons of Math Destruction*. Her book is one of the catalysts for the entire movement for algorithmic accountability. O'Neil's consulting company, O'Neil Risk Consulting & Algorithmic Auditing (ORCAA), does bespoke auditing to help companies and organizations manage and audit their algorithmic risks. I have had the good fortune to consult with ORCAA. When ORCAA's auditors consider an algorithm, they start by asking two questions: What does it mean for this algorithm to work? And how could this algorithm fail, and for whom?

One thing ORCAA does is what's called an internal audit, which means auditors ask these questions directly of companies and other organizations, focusing on algorithms as they are used in specific contexts. They have also asked these starting questions of regulators and lawmakers in the course of developing standards for algorithmic auditing. ORCAA's approach is inclusive: the company aims to incorporate and address concerns from all the stakeholders in an algorithm, not just those who built or deployed it. It is essential to include members of an affected community in an audit in order to evaluate whether harms have occurred.

ORCAA has worked with computer scientist Joy Buolamwini's organization, the Algorithmic Justice League, to perform audits with an intersectional focus. In Buolamwini's paper "Gender Shades," she and computer scientist Timnit Gebru propose an intersectional framework for analyzing an algorithm. This means evaluating the algorithm's performance for different subgroups. Not just men and women, but perhaps also nonbinary and trans folks, and for darker-skinned women and lighter-skinned men. Intersectionality looks at the intersection of different groups that an individual belongs to, like race and gender, and proposes that the intersection gives rise to different experiences and different forms of oppression or discrimination. This reality, often referred to as the matrix of domination, is different for a Latinx trans woman or a Black man or an Afro-Caribbean woman or a Pacific Islander domestic worker or a disabled Native American CEO or any other combination of identities. Thinking about race and gender and ability explicitly, and writing down an intersectional matrix of people for whom the algorithm might fail, makes it easier to spot problems.

> The biases embedded in technology are more than mere glitches; they're baked in from the beginning. They are structural biases, and they can't be addressed with a quick code update.

Auditing is fascinating because it requires digging into the history of an algorithm and weighing competing corporate and mathematical imperatives. One of the things we do is translate extremely complicated mathematical concepts for different corporate audiences. In math, you look to prove theorems that hold true everywhere, across time and space, in the same way. This is what physicists do too, but for the natural world. People trained in math and physics (which includes many data scientists and computer scientists) often make predictable mistakes when writing code for social contexts, because they are looking for one method that explains everything. In auditing, there's less of a focus on one single explanation, and we consider both quantitative and qualitative factors.

Auditing involves a lot of creativity, looking for the edge cases and figuring out what could go wrong. We look at the code or the design pattern that went into making an algorithmic system and do what in other contexts might be called threat modeling. Sociologist Ruha Benjamin's idea of tech discrimination is my own animating principle. Benjamin offers the frame that technology discriminates by default, not that discrimination is a glitch. Adopting this point of view makes it easier to see where technology might be disadvantaging certain groups or violating people's civil rights. My other default assumption is that AI doesn't work as well as people imagine. This perspective also makes it easier to spot algorithmic problems.

Discovering discriminatory patterns

In addition to internal audits, there are external audits, which (as the name suggests) are performed outside the company, without access to code or trade secrets. Usually external audits are initiated by journalists, lawyers, or watchdog groups. ORCAA, for example, has helped attorneys general to identify and prosecute cases where algorithms are used to break the law. An attorney general has the power to demand (via subpoena) documentation, system data, or code from the target company. External audits are sometimes quite creative. For example, a watchdog project called Exposing.ai lets you find out if your Flickr photos were used to train facial recognition systems. This is more common than most people expect. Rarely are they excited to find that their photos have been collected and used to train AI models. When the news broke that Clearview AI had scraped millions of Flickr photos and used them to create a facial recognition database for policing, there was a massive outcry. Clearview AI argued that its use was within the labeled use of the images, but many people do not feel it was an ethical use of their images. ChatGPT and other generative AI systems also use data scraped from the open web. A project by the *Washington Post* allows users to find out if their websites have been used to train generative AI. That investigation found 200 million examples of the copyright symbol, which indicates a work is proprietary.

Another external audit, by academic researchers at the University of California, Berkeley, revealed that Black and Latinx people pay more for mortgages and are denied at higher rates. "Latinos and African-Americans paid almost one tenth of a percentage point more for mortgages between 2008 and 2015, the study found—a disparity that sucked hundreds of millions of dollars from minority homeowners every year," wrote CBS reporter Kristopher J. Brooks about the study. Black and Latinx borrowers end up paying an additional $765 million per year in additional mortgage costs—a disparity that contributes to the racial wealth gap. Over time, if algorithmic lenders are allowed to continue this discriminatory pattern, the racial wealth gap could become insurmountable.

Auditing, especially automated auditing, is important because models decay. Many people imagine that they will be able to create or implement a computational system and then "set it and forget it." But nothing could be further from the truth. Every computational system needs to be updated, staffed, and tended. Computer systems need to change as the world changes.

Figuring out which fairness metrics to use is one of the biggest auditing challenges. Currently, there are about 21 different mathematical definitions of fairness. Interestingly, these definitions are mutually exclusive. It is mathematically unlikely that any solution can satisfy one kind of fairness, and also satisfy a second criteria for fairness. So, in order to consider an algorithm fair, a choice must be made as to which kind of fairness is the standard for a particular type of algorithm. From a policy perspective, this means that all similar algorithms would need to be evaluated according to the same fairness metric.

Auditors need to examine algorithmic systems for search, e-commerce, online advertising, advertising tech, maps, ridesharing, online reviews and ratings, natural language processing, education tech, recommendation systems, facial recognition inside and outside policing, predictive policing, criminal justice, housing, credit, background checking, financial services, insurance, child protective services, and more. These systems all operate in different contexts, and the same test won't necessarily suit every industry. Auditors need to choose which single fairness metric works best for each algorithm in each specific context. They can then choose among multiple software packages to run the chosen fairness test. One popular opensource package is called AI Fairness 360. There are also platforms for auditing, such as Aequitas. I helped ORCAA build a system called Pilot, a platform for automated, continuous algorithmic auditing.

Auditing is not a silver bullet—it is a tool in an imperfect system. It does not always work. One effort gone awry happened in New York City, when the city made a task force devoted to cataloguing and overseeing all the city's algorithms. The task force, which was made up of multiple people with world-class reputations in algorithmic fairness, disbanded after only a year. It didn't have enough funding or resources, and the city didn't have the capacity to do what it said it was going to do with the task force. "The task force was given no details into how even the simplest of automated decision systems worked," wrote task force member Albert Fox Cahn, founder and executive director of the Surveillance Technology Oversight Project, about the fiasco. "By January 2019," he recounted, "there was growing anger about the city's unwillingness to provide information on what automated decision systems it already used. This undercut the value of the task force, which aimed to escape the theories and generalizations of the ivory tower

Until recently, software developers have not paid enough attention to ensuring their algorithms operate within existing laws. Auditing is a way to make sure that the public interest is being preserved in and around algorithms.

to examine how these tools were operating in the real world, using the country's largest city as our test case. Only we never got the data."

The effort needed to be better resourced, have more power to compel disclosure, and needed far more time than it was given. Auditing is not inexpensive and needs wide-ranging institutional support. The NYC algorithm task force disaster highlights that it's important to have the ability to say "no" to the tech if it is not working well or as expected. Few people are prepared to let software projects go, especially after investing thousands or millions of dollars in their development. Everyone making an algorithmic system needs to be prepared to confront the shortcomings of the computational system and of the larger sociocultural context.

Another instructive audit situation comes from STOP LAPD Spying Coalition, a grassroots organization that led demands to audit PredPol, a predictive policing system used by the LA Police Department. "We released a report, 'Before the Bullet Hits the Body,' in May 2018 on predictive policing in Los Angeles, which led to the city of Los Angeles holding public hearings on data-driven policing, which were the first of their kind in the country," said the organization's founder, Hamid Khan. "We demanded a forensic audit of PredPol by the inspector general. In March 2019, the inspector general released the audit and it said that we cannot even audit PredPol because it's just not possible. It's so, so complicated."

This was interesting to me because it gets at some of the essential problems of auditing software. A forensic audit is different from an algorithmic audit, and not everyone in the legal or forensic world knows that algorithmic auditing exists. It's likely that the inspector general's office didn't understand how an algorithmic audit would work, and thus claimed it was impossible. An audit is possible, yes—but it requires the inspector general's office and the auditors and the audit report-readers and everyone else in the institutional context to have a high level of mathematical and computational literacy in order to understand and communicate the results. Few people know what algorithmic test results mean, and there isn't yet a standard report. An algorithm is itself a kind of amorphous thing in most contexts. When you look at an algorithmic system, or a machine learning model, it looks like gobbledygook unless you can read code and understand data and know the right kind of reports to request from exactly the right person. It's confusing.

Opening the black box

People like to have a "good reason" for a decision, and algorithms rarely give one. There is rarely a reason for algorithmic decisions that will make sense to a human being. Unless a data scientist is willing to sit down and explain every feature of an algorithmic system—and provide the code and the training data and the social context in which the system was developed and deployed—it's hard and deeply unsatisfying to explain what happened. When you examine or audit a system, you get familiar enough with its inputs and outputs that it feels like you understand its reasoning. Without that understanding of the code and the training data and the social context, it feels opaque—like a black box. It takes an investment of time in order to understand what's going on, and the person asking for the reason also needs to invest in understanding all the dimensions of the system.

Auditing doesn't have as much marketing hype behind it as innovation does. In part, this is because there is lots of funding (venture capital and otherwise) for building new things, but very little funding for fixing and improving the things that already exist. Public interest technology pushes back against this, with an awareness that society needs to fund infrastructure as well as innovation. If developers are building AI systems that intervene in people's lives, society needs to maintain and inspect and replace the systems the same way it maintains and inspects and replaces bridges and roads. Another thing that will help is decoupling innovation from social progress. Innovation and social progress are *not* the same thing. Using more technology does not bring about social progress if the technology causes algorithmic harms or (as is often the case) reverses hard-won civil rights advances. Finally, diversifying the landscape of technology creators will help, so that there are more people in the room who can bring more viewpoints and can raise awareness of potential issues that will need to be audited.

Meredith Broussard *is associate professor at the Arthur L. Carter Journalism Institute of New York University and research director at the NYU Alliance for Public Interest Technology. Her most recent book is* More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech *(The MIT Press, 2023), from which this essay is adapted.* *Reprinted with permission from The MIT Press. Copyright 2023.*