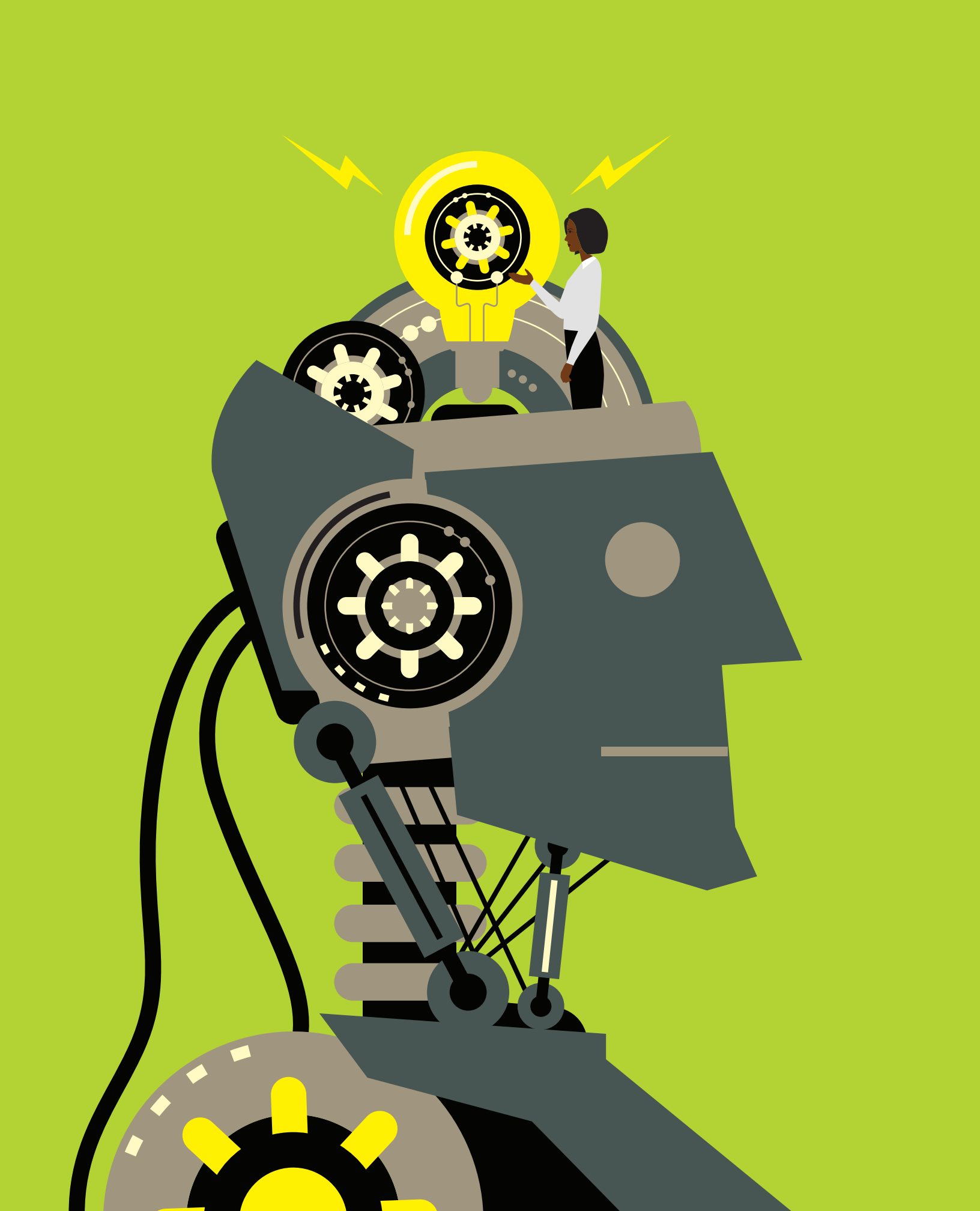
IS YOUR ORGANIZATION FUTURE-READY? PAGES 18, 42, 60, 78

Rotman MANAGEMENT

The Magazine of the Rotman School of Management
UNIVERSITY OF TORONTO
FALL 2022

The Future-Proof Issue

Rotman School of Management
UNIVERSITY OF TORONTO



POWER AND PREDICTION: THE ANTI-DISCRIMINATION OPPORTUNITY

AI system solutions have the potential to reduce discrimination across domains, from education to healthcare and banking.

by Ajay Agrawal, Joshua Gans and Avi Goldfarb

IN RETROSPECT, it was unlikely to turn out well. But in 2016, **Microsoft** researchers released an AI algorithm called Tay to learn how to interact on **Twitter**. Within hours, it learned alright, and began to spew out offensive tweets. Tay was not alone in becoming the worst of us. Stories like this abound and make many organizations reluctant to adopt AI. This is not because AI prediction performs worse than people. Instead, AI may be too good at behaving like them.

This shouldn't be a surprise. AI prediction requires data and, especially for data that involves predicting something about people, the training data comes from people. There can be merit in this, such as when training to play a game against people, but people are imperfect, and AI inherits those imperfections.

What many don't recognize is that this is a current problem because of how we have been thinking about AI solutions. When you are interested in, say, allowing your human resources department to screen hundreds of applicants, a potential use for AI is to use an algorithm rather than people for that job. It is, after all, a predictive task: What is the likelihood that *this* person with *these* credentials will succeed in *this* role? But this way of using AI is what is known as a 'point solution' — a tool that addresses a single-use case or challenge that exists within an organization. These can work, but a full system-level redesign is often warranted. This is why removing the adverse consequences of bias requires a *system mindset*.

The Opportunity Before Us

When viewed using a system mindset, the opportunities for AI with respect to bias are all upside. We believe they offer a solution to many aspects of discrimination. And it is precisely because they offer this that they face resistance. The uncomfortable truth about discrimination is that, as power shifts, winners and losers are generated. Thus, resistance to adopting AIs is likely to be higher precisely when AIs have the potential to engender new systems that eliminate many aspects of discrimination.

Consider a simple example. People of colour report much more knee pain than whites. There are two distinct explanations for this. First, people of colour might have more severe osteoarthritis within the knee. Alternatively, other factors external to the knee — such as life stress or social isolation — may lead to higher levels of knee pain. These explanations imply different treatments: If the issue is severe osteoarthritis, physiotherapy, medication and surgery can help; but if the issue is *external* to the knee, the most effective treatments might focus instead on

It is easier to catch a deliberately discriminatory AI than a deliberately discriminatory human being.

addressing mental health. Many doctors have suspected that factors external to the knee were more important in explaining the racial disparities. Studies have compared the pain reported by patients with radiologists' assessments of knee osteoarthritis based on medical imaging. The radiologists base their assessments on methods such as the **Kellgren-Lawrence** (KL) classification, through which doctors examine images of a patient's knee and assign a score based on the presence of bone spurs, deformities and other factors. Even after adjusting for these assessments, people of colour report higher levels of pain.

Computer scientist **Emma Pierson** and her co-authors suspected the issue might be in the classification system. The methods for measuring osteoarthritis, including KL, were developed decades ago in white populations. As such, they might miss the physical causes of pain in non-white populations. Radiologists may also be biased in their assessments of non-white patients, downplaying their pain in developing a diagnosis.

Pierson and her co-authors took thousands of images of knees. For each image, they had the patient's self-reported level of pain. When radiologists scored the images, only nine per cent of the racial disparities in pain appeared to be explained by factors internal to the knee. The authors then assessed whether an AI could use the images to predict the reported pain. The result: Their AI predicted 43 per cent of the racial disparities in pain, identifying factors within the knee that the humans missed — and these factors explained nearly five times as much of the difference in reported pain between people of colour and whites.

This matters for racial disparity in treatment because it suggests that many non-white patients would receive treatment external to the knee when there is clearly something going on *in* the knee. Here, AI helped identify systemic discrimination — and a path for fixing it. To address discrimination, both of these are necessary. You need to *detect* the discrimination and you need to *fix it*. This is true of both human and machine predictions. In other words, eliminating discrimination requires a system.

Detecting Discrimination

Detecting discrimination is hard. Despite plenty of legal claims decrying discrimination in technology and other industries, few are decided in favour of the plaintiffs. Many of these cases focus on whether the firms discriminated in terms of salary or promotions. Suppose a tech firm is accused of gender discrimination in promoting its leaders. There would be no question that the firm promoted numerous men instead of the female plaintiff, who has been at the firm longer; but the question at the heart of the litigation would be, *why*?

The plaintiff will claim that the firm intentionally discriminated against her. The firm will respond that the plaintiff "is less a victim of discrimination than a difficult employee who rejected advice to improve," as the *New York Times* described one defendant's approach. When asked if they discriminated in their recommendations for promotions, of course, managers will say no.

Even when there is discrimination, it is hard to prove. Managers consider a variety of factors when making promotion and hiring decisions, so without an explicit statement of intent to discriminate, it is difficult for a judge or jury to be confident that a human's decision was discriminatory. It is impossible to know what is truly in someone's mind, and no two people are exactly alike. Unless they are.

Sendhil Mullainathan is a world expert on detecting discrimination. In 2001, just three years out of his PhD, he and co-author **Marianne Bertrand** set out to measure discrimination in the U.S. labour market. They sent fictitious résumés to help-wanted ads in Boston and Chicago newspapers. For each ad, they sent four résumés: two were high quality and two were low quality. They randomly assigned one of the high-quality résumés an African-American name (Lakisha Washington or Jamal Jones) and the other a white-sounding name (Emily Walsh or Greg Baker.) Similarly, they randomly assigned one of the low-quality résumés an African-American name and the other a white-sounding name. Then they waited to see if their fictitious applicants would be called back for interviews.

The result: White names received 50 per cent more callbacks. The gap between high-quality résumés with white names and high-quality résumés with African American names was even larger. There was clearly discrimination in the labour market. Fifteen years later, Mullainathan did it again. Now a University of Chicago professor, he and his co-authors discovered that a widely used algorithm employed to identify patients with complex health needs was racially biased. At a given risk score, African-American patients were found to be considerably sicker than white patients. Remedying the disparity would nearly triple the fraction of African-American patients receiving additional resources to manage their care.

The bias arose because the machine was designed to predict healthcare costs as a proxy for illness, rather than the illness itself. Unequal access to care means that the U.S. healthcare system spends less money caring for African-American patients than for white patients. This means that a prediction machine that uses healthcare spending as a proxy for illness will underestimate the severity of illness in African Americans and other patient groups with limited access to care. In the aftermath of

With an AI designed to reduce bias, hiring through social connections will be harder.

this study, Mullainathan reflected on the two projects:

> *Both studies documented racial injustice, but they differed in one crucial respect. In the first, hiring managers made biased decisions. In the second, the culprit was a computer program. As a co-author of both studies, I see them as a lesson in contrasts. Side by side, they show the stark differences between two types of bias: human and algorithmic.*

The earlier study required an extraordinary amount of creativity and effort to detect discrimination and went on for months. In contrast, the later study was more straightforward:

> *This was a statistical exercise — the equivalent of asking the algorithm 'what would you do with this patient?' hundreds of thousands of times and mapping out the racial differences. The work was technical and rote, requiring neither stealth nor resourcefulness.*

Measuring discrimination in people is hard, requiring careful control over the context. But measuring discrimination by machines is more straightforward: Feed the machines the right data and see what comes out. The researcher can go to the AI and say, what if the person is like this? What if the person is like that? It is possible to try thousands of what-ifs. That is not possible with humans. "Humans are inscrutable in a way that algorithms are not," Mullainathan noted.

Fixing Discrimination

Of course, once discrimination is detected, we want to fix it. But humans are hard to fix. In the résumés study, even if you could get over the challenge of figuring out which companies were at fault, changing people's hearts and minds is no simple matter. The evidence on tools like implicit bias training is mixed. We don't know of a fix available that can reduce discrimination perpetuated by thousands or even millions of humans on a daily basis. Two decades after that initial study went into the field, Emily and Greg remain more employable than Lakisha and Jamal.

Contrast that with an AI. Even before the study on algorithmic discrimination was published, Mullainathan and his co-authors were already working with the organization to fix the problem. They started by contacting the company, which was able to replicate the study's result with its own simulations. As a first step, they showed that 'including health prediction with the existing cost prediction' would reduce bias by 84 per cent. The authors offered their services, for free, to a number of healthcare systems using these types of algorithms. Many took them up on the offer.

The research paper concludes: "Because labels are the key determinant of both predictive quality and predictive bias, careful choice can allow us to enjoy the benefits of algorithmic predictions while minimizing their risk." As Mullainathan put it, "software on computers can be updated; the 'wetware' in our brains has so far proven much less pliable."

We do not mean to leave the impression that fixing discrimination is easy. First, it requires humans who *want* to fix the bias. If the humans who manage the AI want to deploy a tool that discriminates, they will have little difficulty doing so. And because the AI is software, their discrimination can happen at scale. However, it is easier to catch a deliberately discriminatory AI than a deliberately discriminatory human, because the AI leaves an audit trail: A well-funded regulator with well-trained auditors who can access the AI can run simulations to look for discrimination, just like Mullainathan and his co-authors did. Unfortunately, our current legal and regulatory systems struggle with these challenges as they were designed for a world of human decision-makers.

Second, even when deployed by well-intentioned humans who want to reduce biases, *details matter* — and focusing on details is time-consuming and expensive. There are many ways bias can seep into an AI's predictions. Fixing bias requires understanding its *source*, which requires investments in storing data about past decisions. It also requires investments in simulating potential sources of bias to see how the AI holds up. And the first attempt might not work. New data might need to be collected and new processes required.

Third, an AI that reduces bias can change who holds decision-making power in an organization. Without AI, it might have been individual managers making decisions on who to hire. Even with the best intentions, these managers might hire through their social connections in a way that leads to unintended bias. With an AI designed to reduce bias, hiring through social connections will be harder. A more senior executive would set the threshold for which résumés should be considered. That executive might recognize that if all the company's managers were hired through their social connections, a diverse workforce would be impossible. The AI reduces discrimination, but it also reduces the discretion that individual managers have in hiring relative to the objectives set by the executive suite. As a result, those managers might resist a system-level change that would reduce their power.

In 2003, **Major League Baseball** used a new tool for identifying the location of pitches over the plate called the QuesTec

Umpire Information System. QuesTec was used to evaluate the balls and strikes called by umpires. Unsurprisingly, the umpires resisted the tool. So did some of the star players. **Sandy Alderson**, MLB's vice president of operations, described one motivation for the tool, claiming that some veteran players were getting the benefit of the doubt and having balls and strikes called in their favour. Many of the game's biggest stars complained, including award-winning pitcher **Tom Glavine** and multiple MVP winner **Barry Bonds**. An automated tool in which a computer predicted balls and strikes might have decreased bias — but those who were benefiting from the bias didn't like that.

It Takes a System

One out of every 153 workers in the U.S. is an **Amazon** employee. Thus, it should not surprise you that the company was very interested in developing an AI to assist with its recruiting. In 2014, they did just that; but a year later, the system was scrapped and never made it to the field. Why? Because it was found to not be evaluating candidates for technical jobs in a gender-neutral manner. The reason was a familiar one: Amazon's AI was trained on past data that was overwhelmingly male. When they looked under the hood, the AI was explicitly down-weighting references to women, including women's colleges. Simple tweaks could not restore neutrality.

You might read stories like this and think AI is hopelessly biased. But the other way you can read this is: The AI was biased and was judged to be such and so was not deployed. Could the same have been said for human recruiters? We actually know the answer: The AI was trained on those recruiters in the first place. At the same time, this experience has taught AI developers that training on past data is often not good enough. New sources of data are required, and this takes time to develop. But in the end, the resulting AI can be evaluated. What's more, it can be continually monitored for performance.

This is a potentially profound improvement over how we deal with discrimination today. Today's interventions to alleviate discrimination are primarily outcome-based: Is there a difference between outcomes for different groups? And the interventions are often direct rules to try and redress a balance and achieve outcome parity. The problem is that those interventions can be divisive.

By contrast, what people often want is to remove the source of the bias — in particular, the motivations of the people who are making decisions. They don't want to fix equal outcomes per se, they want equal treatment. However, when people are making the decisions and we can't see their motivations, how can we have confidence that there is ever equal treatment? If AI prediction can be placed at the heart of such decisions, an objective benchmark can be achieved: We can see how the AI treats people, and because we know it cannot have explicit motivations to treat people differently, we can work on ensuring that it actually doesn't.

Automated predictions make it easier to create standards. Just like all baseball players face the same strike zone, all drivers face the same traffic enforcement standards. There are well-documented biases in traffic enforcement. For example, Black drivers get pulled over more than whites. An easy point solution is to automate speeding tickets. We have the technology for this: Detect a car's speed, take pictures and then punish the drivers accordingly. An automated system is fairer and safer, reducing the chance of a violent encounter between police and the public.

But the benefits of automating this go well beyond the point solution. Having confidence that everyone is being treated equally changes how people interact with a system and how safe they feel behaving within it. It also removes the need for interventions that simply try to look good on the books — like having outcomes be part of fixed quotas.

Automated systems will not be welcomed by everyone. Just like the star baseball players we described, drivers that would have received the benefit of officer discretion might resent the cameras. Furthermore, an automated system cannot know how to be lenient if someone is speeding for a good reason — like a medical emergency. Still, having drivers stay below the speed limit unquestionably saves lives. If enforcement is often discriminatory, automated enforcement will catch more dangerous drivers *and* reduce discrimination.

In closing

We see the potential for AIs to reduce bias in all sorts of decisions. But this optimism disguises a broader pessimism about human decision-making. As MIT Computer Scientist **Marzyeh Ghassemi** put it after a lecture on biases in machine learning in healthcare, "Humans are awful." But AI bias *can* be detected and addressed.

The good news is that new AI system solutions across domains — from education to healthcare and from banking to policing — can be designed and implemented to reduce discrimination. And these systems can be continuously and retroactively monitored to ensure continued success at removing discrimination. If only it were that easy to fix humans. ***RM***



Ajay Agrawal, O.O.C., is the Geoffrey Taber Chair in Entrepreneurship and Innovation and Professor of Strategic Management at the Rotman School of Management, where he founded and leads the Creative Destruction Lab. **Joshua Gans** is a Professor of Strategic Management and holds the Jeffrey Skoll Chair in Technical Innovation and Entrepreneurship at the Rotman School. **Avi Goldfarb** is a Professor of Marketing and holds the Rotman Chair in Artificial Intelligence and Healthcare at the Rotman School. They are co-authors of *Power and Prediction: The Disruptive Economics of Artificial Intelligence* (Harvard Business Review Press, 2022) and *Prediction Machines: The Simple Economics of Artificial Intelligence* (HBR Press, 2018.)

Explore more rotmanmagazine.ca