The bias detectives

As machine learning infiltrates society, scientists grapple with how to make algorithms fair.

BY RACHEL COURTLAND

In 2015, a worried father asked Rhema Vaithianathan a question that still weighs on her mind. A small crowd had gathered in a basement room in Pittsburgh, Pennsylvania, to hear her explain how software might tackle child abuse. Each day, the area's hotline receives dozens of calls from people who suspect that a child is in danger; some of these are then flagged by call-centre staff for investigation. But the system does not catch all cases of abuse. Vaithianathan and her colleagues had just won a half-million-dollar contract to build an algorithm to help.

Vaithianathan, a health economist who co-directs the Centre for Social Data Analytics at the Auckland University of Technology in New Zealand, told the crowd how the algorithm might work. For example, a tool trained on reams of data — including family backgrounds and criminal records — could generate risk scores when calls come in. That could help call screeners to flag which families to investigate.

After Vaithianathan invited questions from her audience, the father stood up to speak. He had struggled with drug addiction, he said, and social workers had removed a child from his home in the past. But he had been clean for some time. With a computer assessing his records, would the effort he'd made to turn his life around count for nothing? In other words: would algorithms judge him unfairly?

Vaithianathan assured him that a human would always be in the loop, so his efforts would not be overlooked. But now that the automated tool has been deployed, she still thinks about his question. Computer calculations are increasingly being used to steer potentially life-changing decisions, including which people to detain after they have been charged with a crime; which families to investigate for potential child abuse,

ILLUSTRATION BY MARIO WAGNER

and — in a trend called 'predictive policing' — which neighbourhoods police should focus on. These tools promise to make decisions more consistent, accurate and rigorous. But oversight is limited: no one knows how many are in use. And their potential for unfairness is raising alarm. In 2016, for instance, US journalists argued that a system used to assess the risk of future criminal activity discriminates against black defendants.

"What concerns me most is the idea that we're coming up with systems that are supposed to ameliorate problems [but] that might end up exacerbating them," says Kate Crawford, co-founder of the AI Now Institute, a research centre at New York University that studies the social implications of artificial intelligence.

With Crawford and others waving red flags, governments are trying to make software more accountable. Last December, the New York City Council passed a bill to set up a task force that will recommend how to publicly share information about algorithms and investigate them for bias. This year, France's president, Emmanuel Macron, has said that the country will make all algorithms used by its government open. And in guidance issued this month, the UK government called for those working with data in the public sector to be transparent and accountable. Europe's General Data Protection Regulation (GDPR), which came into force at the end of May, is also expected to promote algorithmic accountability.

Police in Camden, New Jersey, use automated tools to help determine which areas need patrolling.

TIMOTHY CLARY/AFP/GETTY

In the midst of such activity, scientists are confronting complex questions about what it means to make an algorithm fair. Researchers such as Vaithianathan, who work with public agencies to try to build responsible and effective software, must grapple with how automated tools might introduce bias or entrench existing inequity — especially if they are being inserted into an already discriminatory social system.

The questions that automated decision-making tools raise are not entirely new, notes Suresh Venkatasubramanian, a theoretical computer scientist at the University of Utah in Salt Lake City. Actuarial tools for assessing criminality or credit risk have been around for decades. But as large data sets and more-complex models become widespread, it is becoming harder to ignore their ethical implications, he says. "Computer scientists have no choice but to be engaged now. We can no longer just throw the algorithms over the fence and see what happens."

FAIRNESS TRADE-OFFS

When officials at the Department of Human Services in Allegheny County, where Pittsburgh is located, called in 2014 for proposals for an automated tool, they hadn't yet decided how to use it. But they knew they wanted to be open about the new system. "I'm very against using government money for black-box solutions where I can't tell my community what we're doing," says Erin Dalton, deputy director of the department's Office of Data Analysis, Research and Evaluation. The department has a centralized data warehouse, built in 1999, that contains a wealth of information about individuals — including on housing, mental health and criminal records. Vaithianathan's team put in an impressive bid to focus on child welfare, Dalton says.

The Allegheny Family Screening Tool (AFST) launched in August 2016. For each phone call to the hotline, call-centre employees see a score between 1 and 20 that is generated by the automated risk-assessment system, with 20 corresponding to a case designated as highest risk. These are families for which the AFST predicts that children are most likely to be removed from their homes within two years, or to be referred to the county again because a caller has suspected abuse (the county is in the process of dropping this second metric, which does not seem to closely reflect the cases that require further investigation).

An independent researcher, Jeremy Goldhaber-Fiebert at Stanford University in California, is still assessing the tool. But Dalton says preliminary results suggest that it is helping. The cases that call-centre staff refer to investigators seem to include more instances of legitimate concern, she says. Call screeners also seem to be making more consistent decisions about cases that have similar profiles. Still, their decisions don't necessarily agree with the algorithm's risk scores; the county is hoping to bring the two into closer alignment.

As the AFST was being deployed, Dalton wanted more help working out whether it might be biased. In 2016, she enlisted Alexandra Chouldechova, a statistician at Carnegie Mellon University in Pittsburgh, to analyse whether the software was discriminating against particular groups. Chouldechova had already been thinking about bias in algorithms — and was about to weigh in on a case that has triggered substantial debate over the issue. In May that year, journalists at the news website ProPublica reported on commercial software used by judges in Broward County, Florida, that helps to decide whether a person charged with a crime should be released from jail before their trial. The journalists said

"If you want to be fair in one way, you might necessarily be unfair in another."

that the software was biased against black defendants. The tool, called COMPAS, generated scores designed to gauge the chance of a person committing another crime within two years if released.

The ProPublica team investigated COMPAS scores for thousands of defendants, which it had obtained through public-records requests. Comparing black and white defendants, the journalists found that a disproportionate number of black defendants were 'false positives': they were classified by COMPAS as high risk but subsequently not charged with another crime.

The developer of the algorithm, a Michigan-based company called Northpointe (now Equivant, of Canton, Ohio), argued that the tool was not biased. It said that COMPAS was equally good at predicting whether a white or black defendant classified as high risk would reoffend (an example of a concept called 'predictive parity'). Chouldechova soon showed that there was tension between Northpointe's and ProPublica's measures of fairness[1]. Predictive parity, equal false-positive error rates, and equal false-negative error rates are all ways of being 'fair', but are

statistically impossible to reconcile if there are differences across two groups — such as the rates at which white and black people are being rearrested (see 'How to define 'fair'"). "You can't have it all. If you want to be fair in one way, you might necessarily be unfair in another definition that also sounds reasonable," says Michael Veale, a researcher in responsible machine learning at University College London.

In fact, there are even more ways of defining fairness, mathematically speaking: at a conference this February, computer scientist Arvind Narayanan gave a talk entitled '21 fairness definitions and their politics' — and he noted that there were still others. Some researchers who have examined the ProPublica case, including Chouldechova, note that it's not clear that unequal error rates are indicative of bias. They instead reflect the fact that one group is more difficult to make predictions about than another, says Sharad Goel, a computer scientist at Stanford. "It turns out that that's more or less a statistical artefact."

For some, the ProPublica case highlights the fact that many agencies lack resources to ask for and properly assess algorithmic tools. "If anything, what it's showing us is that the government agency who hired Northpointe did not give them a well-defined definition to work with," says Rayid Ghani, who directs the Center for Data Science and Public Policy at the University of Chicago, Illinois. "I think that governments need to learn and get trained in how to ask for these systems, how to define the metrics they should be measuring and to make sure that the systems they are being given by vendors, consultants and researchers are actually fair."

Allegheny County's experience shows how difficult it is to navigate these questions. When Chouldechova, as requested, began digging through the Allegheny data in early 2017, she found that its tool also suffered similar statistical imbalances. The model had some "pretty undesirable properties", she says. The difference in error rates was much higher than expected across race and ethnicity groups. And, for reasons that are still not clear, white children that the algorithm scored as at highest risk of maltreatment were less likely to be removed from their homes than were black children given the highest risk scores[2]. Allegheny and Vaithianathan's team are currently considering switching to a different model. That could help to reduce inequities, says Chouldechova.

Although statistical imbalances are a problem, a deeper dimension of unfairness lurks within algorithms — that they might reinforce societal injustices. For example, an algorithm such as COMPAS might purport to predict the chance of future criminal activity, but it can only rely on measurable proxies, such as being arrested. And variations in policing practices could mean that some communities are disproportionately targeted, with people being arrested for crimes that might be ignored in other communities. "Even if we are accurately predicting something, the thing we are accurately predicting might be the imposition of injustice," says David Robinson, a managing director at Upturn, a non-profit social-justice organization in Washington DC. Much would depend on the extent to which judges rely on such algorithms to make their decisions — about which little is known.

Allegheny's tool has come under criticism along similar lines. Writer and political scientist Virginia Eubanks has argued that, irrespective of whether the algorithm is accurate, it is acting on biased inputs, because black and biracial families are more likely to be reported to hotlines. Furthermore, because the model relies on public-services information in the Allegheny system — and because the families who used such services are generally poor — the algorithm unfairly penalizes poorer families by subjecting them to more scrutiny. Dalton acknowledges that the available data are a limitation, but she thinks the tool is needed. "The unfortunate societal issue of poverty does not negate our responsibility to improve our decision-making capacity for those children coming to our attention," the county said in a response to Eubanks, posted on the AFST website earlier this year.

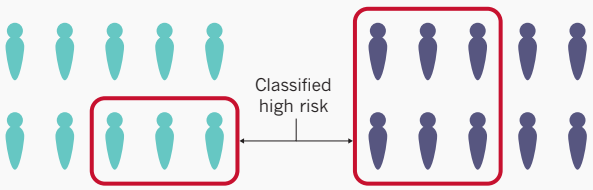
TRANSPARENCY AND ITS LIMITS

Although some agencies build their own tools or use commercial software, academics are finding themselves in demand for work on public-sector algorithms. At the University of Chicago, Ghani has been working with a range of agencies, including the public-health department

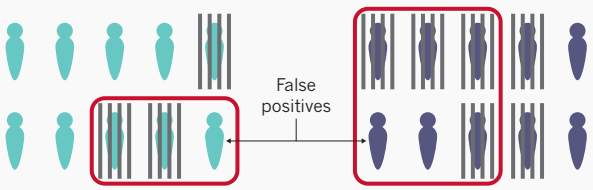
How to define 'fair'

Researchers studying bias in algorithms say there are many ways of defining fairness, which are sometimes contradictory.

Imagine that an algorithm for use in the criminal-justice system assigns scores to two groups (blue and purple) for their risk of being rearrested. Historical data indicate that the purple group has a higher rate of arrest, so the model would classify more people in the purple group as high risk (see figure, top). This could occur even if the model's developers try to avoid bias by not directly telling their model whether a person is blue or purple. That is because other data used as training inputs might correlate with being blue or purple.



A high-risk status cannot perfectly predict rearrest, but the algorithm's developers try to make the prediction equitable: for both groups, 'high risk' corresponds to a two-thirds chance of being rearrested within two years. (This kind of fairness is termed predictive parity.) Rates of future arrests might not follow past patterns. But in this simple example, assume that they do: as predicted, 3 out of 10 in the blue group and 6 out of 10 in the purple group (and two-thirds of those labelled high risk in each group) are indeed rearrested (indicated in grey bars in figure, bottom).



This algorithm has predictive parity. But there's a problem. In the blue group, 1 person out of 7 (14%) was misidentified as high risk; in the purple group, it was 2 people out of 4 (50%). So purple individuals are more likely to be 'false positives': misidentified as high risk.

As long as blue and purple group members are rearrested at different rates, then it will be difficult to achieve predictive parity and equal false-positive rates. And it is mathematically impossible to achieve this while also satisfying a third measure of fairness: equal false-negative rates (individuals who are identified as low risk but subsequently rearrested; in the example above, this happens to be equal, at 33%, for both purple and blue groups).

Some would see the higher false-positive rates for the purple group as discrimination. But other researchers argue that this is not necessarily clear evidence of bias in the algorithm. And there could be a deeper source for the imbalance: the purple group might have been unfairly targeted for arrest in the first place. In accurately predicting from past data that more people in the purple group will be rearrested, the algorithm could be recapitulating — and perhaps entrenching — a pre-existing societal bias. **R.C.**

SOURCE: A. CHOULDECHOVA

of Chicago on a tool to predict which homes might harbour hazardous lead. In the United Kingdom, researchers at the University of Cambridge have worked with police in County Durham on a model that helps to identify who to refer to intervention programmes, as an alternative to prosecution. And Goel and his colleagues this year launched the Stanford Computational Policy Lab, which is conducting collaborations with government agencies, including the San Francisco District Attorney's office. Partnerships with outside researchers are crucial, says Maria McKee, an analyst at the district attorney's office. "We all have a sense of what is right and what is fair," she says. "But we often don't have the tools or the research to tell us exactly, mechanically, how to get there."

There is a large appetite for more transparency, along the lines adopted by Allegheny, which has engaged with stakeholders and opened its doors to journalists. Algorithms generally exacerbate problems when they are "closed loops that are not open for algorithmic auditing, for review, or for public debate", says Crawford at the AI Now Institute. But it is not clear how best to make algorithms more open. Simply releasing all the parameters of a model won't provide much insight into how it works, says Ghani. Transparency can also conflict with efforts to protect privacy. And in some cases, disclosing too much information about how an algorithm works might allow people to game the system.

One big obstacle to accountability is that agencies often do not collect data on how the tools are used or their performance, says Goel. "A lot of times there's no transparency because there's nothing to share." The California legislature, for instance, has a draft bill that calls for risk-assessment tools to help reduce how often defendants must pay bail — a practice that has been criticized for penalizing lower-income defendants. Goel wants the bill to mandate that data are collected on instances when judges disagree with the tool and on specific details, including outcomes, of every case. "The goal is fundamentally to decrease incarceration while maintaining public safety," he says, "so we have to know — is that working?"

Rhema Vaithianathan builds algorithms to help flag potential cases of child abuse.

Crawford says that a range of 'due process' infrastructure will be needed to ensure that algorithms are made accountable. In April, the AI Now Institute outlined a framework[3] for public agencies interested in responsible adoption of algorithmic decision-making tools; among other things, it called for soliciting community input and giving people the ability to appeal decisions made about them.

Many are hoping that laws could enforce such goals. There is some precedent, says Solon Barocas, a researcher who studies ethics and policy issues around artificial intelligence at Cornell University in Ithaca, New York. In the United States, some consumer-protection rules grant citizens an explanation when an unfavourable decision is made about their credit[4]. And in France, legislation that gives a right to explanation and the ability to dispute automated decisions can be found as early as the 1970s, says Veale.

The big test will be Europe's GDPR, which entered into force on 25 May. Some provisions — such as a right to meaningful information about the logic involved in cases of automated decision-making — seem to promote algorithmic accountability. But Brent Mittelstadt, a data ethicist at the Oxford Internet Institute, UK, says the GDPR might actually hamper it by creating a "legal minefield" for those who want to assess fairness. The best way to test whether an algorithm is biased along certain lines — for example, whether it favours one ethnicity over another — requires knowing the relevant attributes about the people who go into the system. But the GDPR's restrictions on the use of such sensitive data are so severe and the penalties so high, Mittelstadt says, that companies in a position to evaluate algorithms might have little incentive to handle the information. "It seems like that will be a limitation on our ability to assess fairness," he says. The scope of GDPR provisions that might give the public insight into algorithms and the ability to appeal is also in question. As written, some GDPR rules apply only to systems that are fully automated, which could exclude situations in which an algorithm affects a decision but a human is supposed to make the final call. The details, Mittelstadt says, should eventually be clarified in the courts.

AUDITING ALGORITHMS

Meanwhile, researchers are pushing ahead on strategies for detecting bias in algorithms that haven't been opened up for public scrutiny. Firms might be unwilling to discuss how they are working to address fairness, says Barocas, because it would mean admitting that there was a problem in the first place. Even if they do, their actions might ameliorate bias but not eliminate it, he says. "So any public statement about this will also inevitably be an acknowledgment that the problem persists." But in recent months, Microsoft and Facebook have both announced the development of tools to detect bias.

Some researchers, such as Christo Wilson, a computer scientist at Northeastern University in Boston, try to uncover bias in commercial algorithms from the outside. Wilson has created mock passengers who purport to be in search of Uber taxi rides, for example, and has uploaded dummy CVs to a jobs website to test for gender bias. Others are building software that they hope could be of general use in self-assessments. In May, Ghani and his colleagues released open-source software called Aequitas to help engineers, policymakers and analysts to audit machine-learning models for bias. And mathematician Cathy O'Neil, who has been vocal about the dangers of algorithmic decision-making, has launched a firm that is working privately with companies to audit their algorithms.

Some researchers are already calling for a step back, in criminal-justice applications and other areas, from a narrow focus on building algorithms that make forecasts. A tool might be good at predicting who will fail to appear in court, for example. But it might be better to ask why people don't appear and, perhaps, to devise interventions, such as text reminders or transportation assistance, that might improve appearance rates. "What these tools often do is help us tinker around the edges, but what we need is wholesale change," says Vincent Southerland, a civil-rights lawyer and racial-justice advocate at New York University's law school. That said, the robust debate around algorithms, he says, "forces us all to ask and answer these really tough fundamental questions about the systems that we're working with and the ways in which they operate".

Vaithianathan, who is now in the process of extending her child-abuse prediction model to Douglas and Larimer counties in Colorado, sees value in building better algorithms, even if the overarching system they are embedded in is flawed. That said, "algorithms can't be helicopter-dropped into these complex systems", she says: they must be implemented with the help of people who understand the wider context. But even the best efforts will face challenges, so in the absence of straight answers and perfect solutions, she says, transparency is the best policy. "I always say: if you can't be right, be honest." ■

Rachel Courtland *is a science journalist based in New York City.*

1. Chouldechova, A. Preprint at https://arxiv.org/abs/1703.00056 (2017).
2. Chouldechova, A., Putnam-Hornstein, E., Benavides-Prado, D., Fialko, O. & Vaithianathan, R. *Proc. Machine Learn. Res.* **81**, 134–148 (2018).
3. Reisman, D., Schultz, J., Crawford, K. & Whittaker, M. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability* (AI Now, 2018).
4. Wachter, S., Mittelstadt, B. & Floridi, L. *Sci. Robotics* **2**, eaan6080 (2017).