Biased algorithms: here's a more radical approach to creating fairness

January 21 2019, by Tom Douglas

Credit: AI-generated image (disclaimer)

Our lives are increasingly affected by algorithms. People may be denied loans, jobs, insurance policies, or even parole on the basis of risk scores that they produce.

Yet algorithms are notoriously prone to biases. For example, algorithms

used to assess the risk of criminal recidivism often have higher error rates in minority ethic groups. As ProPublica found, the COMPAS algorithm – widely used to predict re-offending in the US criminal justice system – had a higher false positive rate in black than in white people; black people were more likely to be wrongly predicted to re-offend.

Findings such as these have led some to claim that algorithms are unfair or discriminatory. In response, AI researchers have sought to produce algorithms that avoid, or at least minimise, unfairness, for example, by equalising false positive rates across racial groups. Recently, an MIT group reported that they had developed a new technique for taking bias out of algorithms without compromising accuracy. But is fixing algorithms the best way to combat unfairness?

It depends on what kind of fairness we're after. Moral and political philosophers often contrast two types of fairness: procedural and substantive. A policy, procedure, or course of action, is procedurally fair when it is fair independently of the outcomes it causes. A football referee's decision may be fair, regardless of how it affects the game's outcome, simply because the decision was made on the basis of an impartial application of the rules. Or a parent's treatment of his two children may be fair because it manifests no partiality or favouritism, even if it has the result that one child's life goes much better than the other's.

By contrast, something that is substantively fair produces fair outcomes. Suppose a football referee awards a soft penalty to a team that is 1-0 down because she thinks the other team's lead was the result of pure luck. As a result, the game finishes in a 1-1 draw. This decision seems procedurally unfair – the referee applies the rules less stringently to one team than the other. But if a draw reflects the relative performance of the two teams, it may be substantively fair.

Alternatively, imagine that a mother and father favour different children. Each parent treats the disfavoured child unfairly, in a procedural sense. But if the end result is that the two children receive equal love, then their actions may be substantively fair.

What's fair?

AI researchers concerned about fairness have, for the most part, been focused on developing algorithms that are procedurally fair – fair by virtue of the features of the algorithms themselves, not the effects of their deployment. But what if it's substantive fairness that really matters?

There is usually a tension between procedural fairness and accuracy – attempts to achieve the most commonly advocated forms of procedural fairness increase the algorithm's overall error rate. Take the COMPAS algorithm for example. If we equalised the false positive rates between black and white people by ignoring the predictors of recidivism that tended to be disproportionately possessed by black people, the likely result would be a loss in overall accuracy, with more people wrongly predicted to re-offend, or not re-offend.

We could avoid these difficulties if we focused on substantive rather than procedural fairness and simply designed algorithms to maximise accuracy, while simultaneously blocking or compensating for any substantively unfair effects that these algorithms might have. For example, instead of trying to ensure that crime prediction errors affect different racial groups equally – a goal that may in any case be unattainable – we could instead ensure that these algorithms are not used in ways that disadvantage those at high risk. We could offer people deemed "high risk" rehabilitative treatments rather than, say, subjecting them to further incarceration.

Alternatively, we could take steps to offset an algorithm's tendency to

assign higher risk to some groups than others – offering risk-lowering rehabilitation programmes preferentially to black people, for instance.

Aiming for substantive fairness outside of the algorithm's design would leave algorithm designers free to focus on maximising accuracy, with fairness left to state regulators, with expert and democratic input. This approach has been successful in other areas. In medicine, for instance, doctors focus on promoting the well-being of their patients while health funders and policymakers promote the fair allocation of healthcare resources across patients.

In substance or procedure

Of course, most of us would be reluctant to give up on procedural fairness entirely. If a referee penalises every minor infringement by one team, while letting another get away with major fouls, we'd think something had gone wrong – even if the right team wins. If a judge ignores everything a defendant says and listens attentively to the plaintiff, we'd think this was unfair, even if the defendant is a jet-setting billionaire who would, even if found guilty, be far better off than a more deserving plaintiff.

We do care about procedural fairness. Yet substantive fairness often matters more – at least, many of us have intuitions that seem to be consistent with this. Some of us think that presidents and monarchs should have the discretion to offer pardons to convicted offenders, even though this applies legal rules inconsistently – letting some, but not others, off the hook. Why think this is justified? Perhaps because pardons help to ensure substantive fairness where procedurally fair processes result in unfairly harsh consequences.

Many of us also think that affirmative action is justified, even when it looks, on the face of it, to be procedurally unfair, since it gives some

groups greater consideration than others. Perhaps we tolerate this unfairness because, through mitigating the effects of past oppression, affirmative action tends to promote substantive fairness.

If substantive fairness generally matters more than procedural fairness, countering biased algorithms through changes to algorithmic design may not be the best path to fairness after all.

Provided by The Conversation