How can we make sure that algorithms are fair?

16 December 2019, by Karthik Kannan

When algorithms make decisions with real-world consequences, they need to be fair. Credit: R-Type/Shutterstock.com

Using machines to augment human activity is nothing new. Egyptian hieroglyphs show the use of horse-drawn carriages even before 300 B.C. Ancient Indian literature such as "Silapadikaram" has described animals being used for farming. And one glance outside shows that today people use motorized vehicles to get around.

Where in the past human beings have augmented ourselves in physical ways, now the nature of augmentation also is more intelligent. Again, all one needs to do is look to cars—engineers are seemingly on the cusp of self-driving cars guided by artificial intelligence. Other devices are in various stages of becoming more intelligent. Along the way, interactions between people and machines are changing.

Machine and human intelligences bring different strengths to the table. Researchers like me are working to understand how algorithms can complement human skills while at the same time minimizing the liabilities of relying on machine intelligence. As a machine learning expert, I predict there will soon be a new balance between human and machine intelligence, a shift that humanity hasn't encountered before.

Such changes often elicit fear of the unknown, and in this case, one of the unknowns is how machines make decisions. This is especially so when it comes to fairness. Can machines be fair in a way that people understand?

When people are illogical

To humans, fairness is often at the heart of a good decision. Decision-making tends to rely on both the emotional and rational centers of our brains, what Nobel laureate Daniel Kahneman calls System 1 and System 2 thinking. Decision theorists believe that the emotional centers of the brain have been quite well developed across the ages, while brain areas involved in rational or logical thinking evolved more recently. The rational and logical part of the brain, what Kahneman calls System 2, has given humans an advantage over other species.

However, because System 2 was more recently developed, human decision-making is often buggy. This is why many decisions are illogical, inconsistent and suboptimal.

For example, preference reversal is a well-known yet illogical phenomenon that people exhibit: In it, a person who prefers choice A over B and B over C does not necessarily prefer A over C. Or consider that researchers have found that criminal court judges tend to be more lenient with parole decisions right after lunch breaks than at the close of the day.

Part of the problem is that our brains have trouble precisely computing probabilities without appropriate training. We often use irrelevant information or are influenced by extraneous factors. This is where machine intelligence can be helpful.

Machines are logical … to a fault

Well-designed machine intelligence can be consistent and useful in making optimal decisions. By their nature, they can be logical in the mathematical sense—they simply don't stray from the program's instruction. In a well-designed machine-learning algorithm, one would not encounter the illogical preference reversals that people frequently exhibit, for example. Within margins of statistical errors, the decisions from machine intelligence are consistent.

The problem is that machine intelligence is not always well designed.

As algorithms become more powerful and are incorporated into more parts of life, scientists like me expect this new world, one with a different balance between machine and human intelligence, to be the norm of the future.

In the criminal justice system, judges use algorithms during parole decisions to calculate recidivism risks. In theory, this practice could overcome any bias introduced by lunch breaks or exhaustion at the end of the day. Yet when journalists from ProPublica conducted an investigation, they found these algorithms were unfair: white men with prior armed robbery convictions were rated as lower risk than African American females who were convicted of misdemeanors.

There are many more such examples of machine learning algorithms later found to be unfair, including Amazon and its recruiting and Google's image labeling.

Researchers have been aware of these problems and have worked to impose restrictions that ensure fairness from the outset. For example, an algorithm called CB (color blind) imposes the restriction that any discriminating variables, such as race or gender, should not be used in predicting the outcomes. Another, called DP (demographic parity), ensures that groups are proportionally fair. In other words, the proportion of the group receiving a positive outcome is equal or fair across both the discriminating and nondiscriminating groups.

Researchers and policymakers are starting to take up the mantle. IBM has open-sourced many of their algorithms and released them under the "AI Fairness 360" banner. And the National Science Foundation recently accepted proposals from scientists who want to bolster the research foundation that underpins fairness in AI.

Improving the fairness of machines' decisions

I believe that existing fair machine algorithms are weak in many ways. This weakness often stems from the criteria used to ensure fairness. Most algorithms that impose "fairness restriction" such as demographic parity (DP) and color blindness (CB) are focused on ensuring fairness at the outcome level. If there are two people from different subpopulations, the imposed restrictions ensure that the outcome of their decisions is consistent across the groups.

While this is a good first step, researchers need to look beyond the outcomes alone and focus on the process as well. For instance, when an algorithm is used, the subpopulations that are affected will naturally change their efforts in response. Those changes need to be taken into account, too. Because they have not been taken into account, my colleagues and I focus on what we call "best response fairness."

If the subpopulations are inherently similar, their effort level to achieve the same outcome should also be the same even after the algorithm is implemented. This simple definition of best response fairness is not met by DP- and CB-based algorithms. For example, DP requires the positive rates to be equal even if one of the subpopulations does not put in effort. In other words, people in one subpopulation would have to work significantly harder to achieve the same outcome. While a DP-based algorithm would consider it fair—after all, both subpopulations achieved the same outcome—most humans would not.

There is another fairness restriction known as equalized odds (EO) which satisfies the notion of best response fairness—it ensures fairness even if

you take into account the response of the subpopulations. However, to impose the restriction, the algorithm needs to know the discriminating variables (say, black/white), and it will end up setting explicitly different thresholds for subpopulations—so, the thresholds will be explicitly different for white and black parole candidates.

While that would help increase fairness of outcomes, such a procedure may violate the notion of equal treatment required by the Civil Rights Act of 1964. For this reason, a California Law Review article has urged policymakers to amend the legislation so that fair algorithms that utilize this approach can be used without potential legal repercussion.

These constraints motivate my colleagues and me to develop an algorithm that is not only "best response fair" but also does not explicitly use discriminating variables. We demonstrate the performance of our algorithms theoretically using simulated data sets and real sample data sets from the web. When we tested our algorithms with the widely used sample data sets, we were surprised at how well they performed relative to open-source algorithms assembled by IBM.

Our work suggests that, despite the challenges, machines and algorithms will continue to be useful to humans—for physical jobs as well as knowledge jobs. We must remain vigilant that any decisions made by algorithms are fair, and it is imperative that everyone understands their limitations. If we can do that, then it's possible that human and machine intelligence will complement each other in valuable ways.

Provided by The Conversation