Weapons of Math Destruction and the Ethical Use of Data
by Diana Skrzydlo

Algorithms are amazing. They can drive your car, suggest what someone might want to watch or buy, track down fraud, keep communications secure, or even predict where the next goose will attack, all with remarkable efficiency.

But an algorithm is only as good as the data it's trained on. If reality is biased, an algorithm trained on that reality will reflect existing inequalities, and worse: perpetuate them with that same remarkable efficiency. The best algorithms get feedback to improve – if a prediction is found to be incorrect, tweaks can be made or additional data included. But many algorithms are black boxes, and end up creating loops that reinforce their own biases rather than correcting them.

Imagine the city of Waterloo is trying to optimize where it sends its police officers to minimize violent crime. One possibility would be to look at data from where violent crimes were committed in the past, and spend more time in those areas. But violent crime is relatively rare, and a model based on that little data might have poor predictive power. So what if they also include data from non-violent, or "nuisance" crimes, such as loitering or bylaw violations? That will give a huge dataset, and would probably reveal more such crimes reported in student areas. So the city would have its police officers spend more time in student areas, where they would probably catch more nuisance crimes, simply by spending more time there. These reports would go back into the algorithm, telling them to spend even more time there, etc. The algorithm would look like a success by catching more "criminals" and validating the high crime areas, but would the city accomplish its goal of reducing violent crime? Probably not! All it would do is catch a lot of minor crimes, while discriminating against students.

Something very similar to this hypothetical example actually happened in many major US cities. Predictive policing algorithms, when fed with "nuisance" crime data, had police spending more and more time in poorer neighbourhoods, which were also highly correlated with racialized neighbourhoods. One result was that while marijuana use is almost equal between white and black teenagers, black teenagers were more than four times as likely to be arrested for possession. The police probably even thought they were being unbiased by "blindly" following the algorithm's recommendations. Unfortunately people believe that algorithms are inherently fair, but they're only as good as their inputs.

Consider another example: the credit score. This started out as a good kind of algorithm - the definition is relatively transparent, the things that help you improve your score (pay bills on time, stop ordering new credit cards, etc.) are actually related to your creditworthiness, and you are entitled to know your information and correct errors if any are found. Sadly, over time it became used as a proxy for things other than the ability to pay back loans.

Insurance companies realized that people with higher credit scores also tended to have fewer claims, and, surmising that people who were conscientious with their finances would be conscientious drivers as well, gave them discounts on car insurance. Employers, looking for a shortcut to screening employees, began using credit score to influence hiring decisions. The problem is that credit score is also a proxy for socio-economic status. This resulted in people who were poorer being charged more for insurance and being less likely to get hired, which in turn reinforced the cycle of poverty. The issue here is that using credit score as a proxy for things other than your ability to pay bills is invalid. Of course the wealthy are more able to pay back loans, but that doesn’t actually make them better drivers or more valuable employees. At a certain point using credit score in this way is just a flimsy excuse to discriminate against poor people. In Canada it is illegal to use credit score as a rating factor in insurance or for hiring decisions, but in the United States it is still used. Punishing those who can least afford it shouldn’t be the goal of any financial system, but unless fairness is explicitly accounted for in the algorithms, bias creeps in. Our present society is biased and our systems should seek to correct this as opposed to reinforce it.

Transparency in algorithms is important, but sometimes it can be a problem as well, opening up the door for people and institutions to game the system. This is readily apparent in university rankings. There are many lists of rankings for universities & colleges, both national and international, and they usually have fairly clear criteria. The problem is, a huge amount of time and effort (and money) is spent improving the factors that affect the ranking, but which may or may not have any bearing on actual quality of education. This is the invalid proxy problem cropping up again.

So what can you do? You will be working for companies who make these kinds of decisions one day. It’s important to determine whether the data you’re using is actually measuring what you want to measure, or is a proxy for something else. Ensure that feedback loops are corrective in nature as opposed to reinforcing bias. Think about whether all the data is actually relevant: sometimes leaving out certain data is the right thing to do. Occasionally it comes down to a moral issue, requiring balance between efficient algorithms and getting non-harmful effects. A small sacrifice in efficiency that prevents a feedback loop of unfairness is a sacrifice you should make.

For more information and lots more examples, please see the book “Weapons of Math Destruction” by Cathy O’Neil. She gave a fascinating talk and workshop at UW in Winter 2018, and most of this article is based on her work.