Can an algorithm deliver justice?

Hannah Fry's new book Hello World is up for the Royal Society Science Book Prize - in this extract she explores how algorithms are helping judges decide sentences.

Article: https://www.sciencefocus.com/future-technology/can-an-algorithm-deliver-justice/

In 2017, a group of researchers set out to discover just how well a machine's predictions stacked up against the decisions of a bunch of human judges.

To help them in their mission, the team had access to the records of every person arrested in New York City over a five-year period between 2008 and 2013. During that time, three-quarters of a million people were subject to a bail hearing, which meant easily enough data to test an algorithm on a head-to-head basis with a human judge.

An algorithm hadn't been used by the New York judicial system during these cases, but looking retrospectively, the researchers got to work building lots of decision trees to see how well one could have predicted the defendants' risk of breaking bail conditions at the time. In went the data on an offender: their rap sheet, the crime they'd just committed, and so on. Out came a probability of whether or not that defendant would go on to violate the terms of their bail.

In the real data, 408,283 defendants were released before they faced trial. Anyone of those was free to flee or commit other crimes, which means we can use the benefit of hindsight to test how accurate the algorithm's predictions and the humans' decisions were. We know exactly who failed to appear in court later (15.2 percent) and who was re-arrested for another crime while on bail (25.8 percent).

Unfortunately for the science, any defendant deemed high risk by a judge would have been denied bail at the time – and hence, on those cases, there was no opportunity to prove the judges' assessment right or wrong. That makes things a little complicated. It means there's no way to come up with a cold, hard number that captures how accurate the judges were overall. And without a 'ground truth' for how those defendants would have behaved, you can't state an overall accuracy for the algorithm either. Instead, you have to make an educated guess on what the jailed defendants would have done if released and make your comparisons of human versus machine in a bit more of a roundabout way.

One thing is for sure, though: the judges and the machine didn't agree on their predictions. The researchers showed that many of the defendants flagged by the

algorithm as real bad guys were treated by the judges as though they were low risk. In fact, almost half of the defendants the algorithm flagged as the riskiest group were given bail by the judges.

But who was right? The data showed that the group the algorithm was worried about did indeed pose a risk. Just over 56 percent of them failed to show up for their court appearances, and 62.7 percent went on to commit new crimes while out on bail – including the worst crimes of all: rape and murder. The algorithm had seen it all coming.

The researchers argued that, whichever way you use it, their algorithm vastly outperforms the human judge. And the numbers back them up. If you're wanting to incarcerate fewer people awaiting trial, the algorithm could help by consigning 41.8 percent fewer defendants to jail while keeping the crime rate the same. Or, if you were happy with the current proportion of defendants given bail, then that's fine too: just by being more accurate at selecting which defendants to release, the algorithm could reduce the rate of skipping bail by 24.7 percent.

These benefits aren't just theoretical. Rhode Island, where the courts have been using these kinds of algorithms for the last eight years, has achieved a 17 percent reduction in prison populations and a 6 percent drop in recidivism rates. That's hundreds of low-risk offenders who aren't unnecessarily stuck in prison, hundreds of crimes that haven't been committed. Plus, given that it costs over £30,000 a year to incarcerate one prisoner in the UK – while in the United States spending a year in a high-security prison can cost about the same as going to Harvard – that's hundreds of thousands of taxpayers' money saved. It's a win-win for everyone.

Or is it?

Finding Darth Vader

Of course, no algorithm can perfectly predict what a person is going to do in the future. Individual humans are too messy, irrational, and impulsive for a forecast ever to be certain of what's going to happen next. They might give better predictions, but they will still make mistakes. The question is, what happens to all the people whose risk scores are wrong?

There are two kinds of mistakes that the algorithm can make. Richard Berk, a professor of criminology and statistics at the University of Pennsylvania and a pioneer in the field of predicting recidivism, has a noteworthy way of describing them.

‘There are good guys and bad guys,’ he told me. ‘Your algorithm is effectively asking: “Who are the Darth Vaders? And who are the Luke Skywalkers?”‘

Letting a Darth Vader go free is one kind of error, known as a false negative. It happens whenever you fail to identify the risk that an individual poses.

Incarcerating Luke Skywalker, on the other hand, would be a false positive. This is when the algorithm incorrectly identifies someone as a high-risk individual.

These two kinds of error, false positive and false negative, are not unique to recidivism. They’ll crop up repeatedly throughout this book. Any algorithm that aims to classify can be guilty of these mistakes.

Berk’s algorithms claim to be able to predict whether someone will go on to commit homicide with 75 percent accuracy, which makes them some of the most accurate around. When you consider how free we believe our will to be, that is a remarkably impressive level of accuracy. But even at 75 percent, that’s a lot of Luke Skywalkers who will be denied bail because they look like Darth Vaders from the outside.

The consequences of mislabelling a defendant become all the more serious when the algorithms are used in sentencing, rather than just decisions on bail or parole. This is a modern reality: recently, some US states have begun to allow judges to see a convicted offender’s calculated risk score while deciding on their jail term. It’s a development that has sparked a heated debate, and not without cause: it’s one thing calculating whether to let someone out early, quite another to calculate how long they should be locked away in the first place.

Part of the problem is that deciding the length of a sentence involves consideration of a lot more than just the risk of a criminal re-offending – which is all the algorithm can help with. A judge also has to take into account the risk the offender poses to others, the deterrent effect the sentencing decision will have on other criminals, the question of retribution for the victim, and the chance of rehabilitation for the defendant. It’s a lot to balance, so it’s little wonder that people raise objections to the algorithm being given too much weight in the decision. Little wonder that people find stories like that of Paul Zilly so deeply troubling.

Zilly was convicted of stealing a lawnmower. He stood in front of Judge Babler in Baron County, Wisconsin, in February 2013, knowing that his defense team had already agreed to a plea deal with the prosecution. Both sides had agreed that, in his

case, a long jail term wasn't the best course of action. He arrived expecting the judge to simply rubber-stamp the agreement.

Unfortunately for Zilly, Wisconsin judges were using a proprietary risk-assessment algorithm called COMPAS. As with the Idaho budget tool in the 'Power' chapter, the inner workings of COMPAS are considered a trade secret. Unlike the budget tool, however, the COMPAS code still isn't available to the public. What we do know is that the calculations are based on the answers a defendant gives to a questionnaire. This includes questions such as: 'A hungry person has a right to steal, agree or disagree?' and: 'If you lived with both your parents and they separated, how old were you at the time?' The algorithm was designed with the sole aim of predicting how likely a defendant would be to re-offend within two years, and in this task had achieved an accuracy rate of around 70 percent. That is, it would be wrong for roughly one in every three defendants. None the less, it was being used by judges during their sentencing decisions.

Zilly's score wasn't good. The algorithm had rated him as a high risk for future violent crime and a medium risk for general recidivism. 'When I look at the risk assessment,' Judge Babler said in court, 'it is about as bad as it could be.'

After seeing Zilly's score, the judge put more faith in the algorithm than in the agreement reached by the defense and the prosecution, rejected the plea bargain and doubled Zilly's sentence from one year in county jail to two years in state prison.

It's impossible to know for sure whether Zilly deserved his high-risk score, although a 70 percent accuracy rate seems a remarkably low threshold to justify using the algorithm to over-rule other factors.

Zilly's case was widely publicized, but it's not the only example. In 2003, Christopher Drew Brooks, a 19-year-old man, had consensual sex with a 14-year-old girl and was convicted of statutory rape by a court in Virginia. Initially, the sentencing guidelines suggested a jail term of 7 to 16 months. But, after the recommendation was adjusted to include his risk score (not built by COMPAS in this case), the upper limit was increased to 24 months. Taking this into account, the judge sentenced him to 18 months in prison.

Here's the problem. This particular algorithm used age as a factor in calculating its recidivism score. Being convicted of a sex offense at such a young age counted against Brooks, even though it meant he was closer in age to the victim. In fact, had Brooks been 36 years old (and hence 22 years older than the girl) the algorithm would have recommended that he not be sent to prison at all.

These are not the first examples of people trusting the output of a computer over their own judgment, and they won't be the last. The question is, what can you do about it? The Supreme Court of Wisconsin has its own suggestion. Speaking specifically about the danger of judges relying too heavily on the COMPAS algorithm, it stated: 'We expect that circuit courts will exercise discretion when assessing a COMPAS risk score with respect to each individual defendant.' But Richard Berk suggests that might be optimistic: 'The courts are concerned about not making mistakes – especially the judges who are appointed by the public. The algorithm provides them a way to do less work while not being accountable.'

There's another issue here. If an algorithm classifies someone as high risk and the judge denies them their freedom as a result, there is no way of knowing for sure whether the algorithm was seeing their future accurately. Take Zilly. Maybe he would have gone on to be violent. Maybe he wouldn't have. Maybe, being labeled as a high-risk convict and being sent to a state prison set him on a different path from the one he was on with the agreed plea deal. With no way to verify the algorithm's predictions, we have no way of knowing whether the judge was right to believe the risk score, no way of verifying whether Zilly was in fact a Vader or a Skywalker.

This is a problem without an easy solution. How do you persuade people to apply a healthy dose of common sense when it comes to using these algorithms? But even if you could, there's another problem with predicting recidivism. Arguably the most contentious of all.

Read More:

https://www.penguin.co.uk/books/1114076/hello-world/9781784163068.html