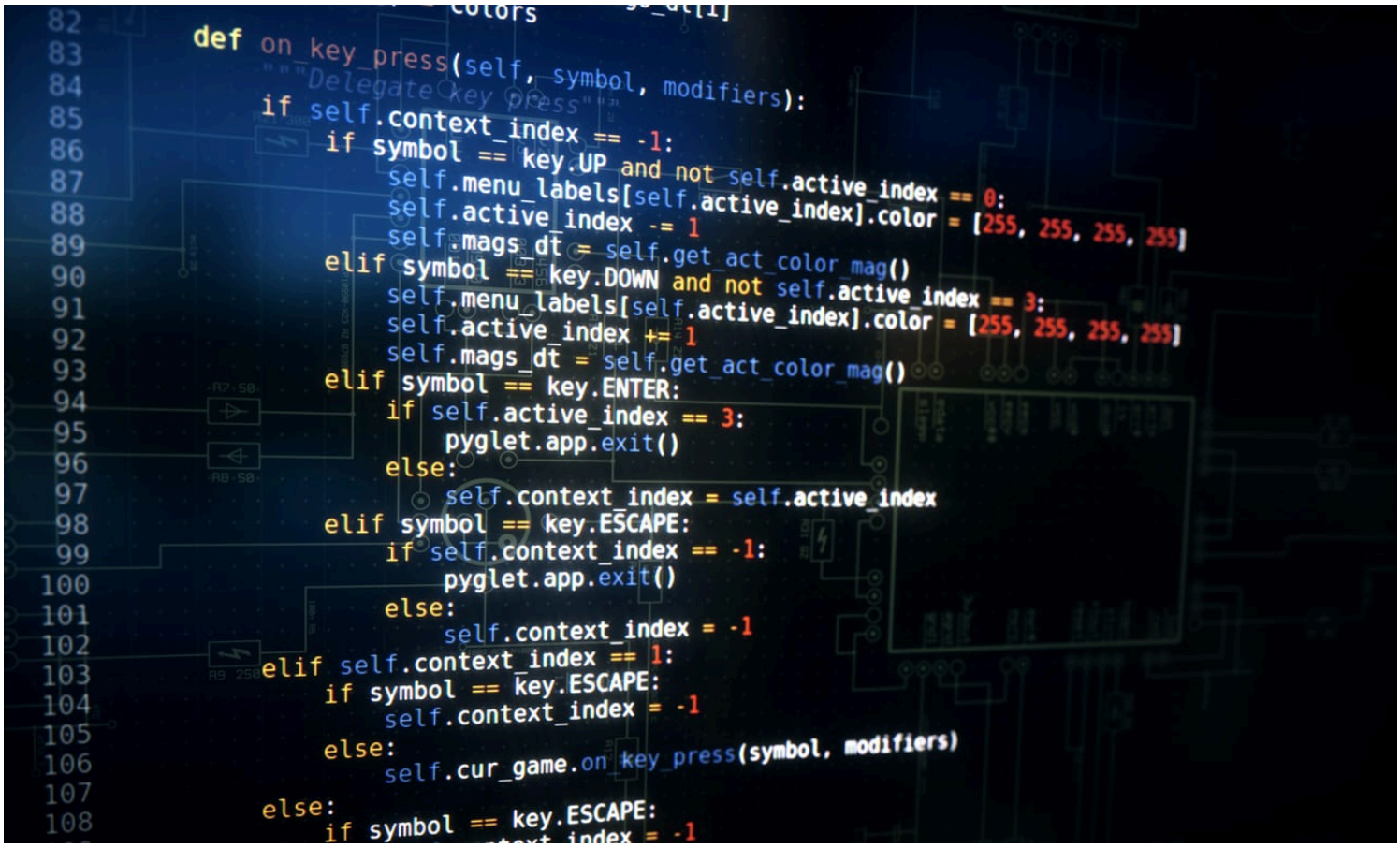
theguardian

How can we stop algorithms telling lies?

Algorithms can dictate whether you get a mortgage or how much you pay for insurance. But sometimes they're wrong – and sometimes they are designed to deceive



How might an algorithm sort your data? Photograph: MatejMo/Getty Images/iStockphoto

Lots of algorithms go bad unintentionally. Some of them, however, are made to be criminal. Algorithms are formal rules, usually written in computer code, that make predictions on future events based on historical patterns. To train an algorithm you need to provide historical data as well as a definition of success.

We've seen finance get taken over by algorithms in the past few decades. Trading algorithms use historical data to predict movements in the market. Success for that algorithm is a predictable market move, and the algorithm is vigilant for patterns that have historically happened just before that move. Financial risk models also use historical market changes to predict cataclysmic events in a more global sense, so not for an individual stock but rather for an entire market. The risk model for mortgage-backed securities was famously bad –

intentionally so - and the trust in those models can be blamed for much of the scale and subsequent damage wrought by the 2008 financial crisis.

Since 2008, we've heard less from algorithms in finance, and much more from big data algorithms. The target of this new generation of algorithms has been shifted from abstract markets to individuals. But the underlying functionality is the same: collect historical data about people, profiling their behaviour online, location, or answers to questionnaires, and use that massive dataset to predict their future purchases, voting behaviour, or work ethic.

Cathy O'Neill. Photograph: Adam Morganstern

The recent proliferation in big data models has gone largely unnoticed by the average person, but it's safe to say that most important moments where people interact with large bureaucratic systems now involve an algorithm in the form of a scoring system. Getting into college, getting a job, being assessed as a worker, getting a credit card or insurance, voting, and even policing are in many cases done algorithmically. Moreover, the technology introduced into these systematic decisions is largely opaque, even to their creators, and has so far largely escaped meaningful regulation, even when it fails. That makes the question of which of these algorithms are working on our behalf even more important and urgent.

I have a four-layer hierarchy when it comes to bad algorithms. At the top there are the unintentional problems that reflect cultural biases. For example, when Harvard professor Latanya Sweeney found that Google searches for names perceived to be black generated ads associated with criminal activity, we can assume that there was no Google engineer writing racist code. In fact, the ads were trained to be bad by previous users of Google search, who were more likely to click on a criminal records ad when they searched for a black sounding name. Another example: the Google image search result for "unprofessional hair", which returned almost exclusively black women, is similarly trained by the people posting or clicking on search results throughout time.

One layer down we come to algorithms that go bad through neglect. These would include scheduling programs that prevent people who work minimum wage jobs from leading decent lives. The algorithms treat them like cogs in a machine, sending them to work at different times of the day and on different days each week, preventing them from having regular childcare, a second job, or going to night school. They are brutally efficient, hugely scaled, and largely legal, collecting pennies on the backs of workers. Or consider Google's

system for automatically tagging photos. It had a consistent problem whereby black people were being labelled gorillas. This represents neglect of a different nature, namely quality assessment of the product itself: they didn't check that it worked on a wide variety of test cases before releasing the code.

Algorithms are used to approve applicants before their CVs are viewed by human eyes, which can lead to discrimination. Photograph: Danny Lawson/PA

The third layer consists of nasty but legal algorithms. For example, there were Facebook executives in Australia showing advertisers ways to find and target vulnerable teenagers. Awful but probably not explicitly illegal. Indeed online advertising in general can be seen as a spectrum, where on the one hand the wealthy are presented with luxury goods to buy but the poor and desperate are preyed upon by online payday lenders. Algorithms charge people more for car insurance if they don't seem likely to comparison shop and Uber just halted an algorithm it was using to predict how low an offer of pay could be, thereby reinforcing the gender pay gap.

Finally, there's the bottom layer, which consists of intentionally nefarious and sometimes outright illegal algorithms. There are hundreds of private companies, including dozens in the UK, that offer mass surveillance tools. They are marketed as a way of locating terrorists or criminals, but they can be used to target and root out citizen activists. And because they collect massive amounts of data, predictive algorithms and scoring systems are used to filter out the signal from the noise. The illegality of this industry is under debate, but a recent undercover operation by journalists at Al Jazeera has exposed the relative ease with which middlemen representing repressive regimes in Iran and South Sudan have been able to buy such systems. For that matter, observers have criticised China's social credit scoring system. Called "Sesame Credit," it's billed as mostly a credit score, but it may also function as a way of keeping tabs on an individual's political opinions, and for that matter as a way of nudging people towards compliance.

Closer to home, there's Uber's "Greyball," an algorithm invented specifically to avoid detection when the taxi service is functioning illegally in a city. It used data to predict which riders were violating the terms of service of Uber, or which riders were undercover government officials. Telltale signs that Greyball picked up included multiple use of the app in a single day and using a credit card tied to a police union.

The most famous malicious and illegal algorithm we've discovered so far is the one used by Volkswagen in 11 million vehicles worldwide to deceive the emissions tests, and in particular to hide the fact that the vehicles were emitting nitrogen oxide at up to 35 times the levels permitted by law. And although it seemed simply like a devious device, this qualifies as an algorithm as well. It was trained to identify and predict testing conditions versus road conditions, and to function differently depending on that result. And, like Greyball, it was designed to deceive.

In 2015, e-commerce business Poster Revolution was found guilty of using algorithms to collude with other poster sellers to set prices. Photograph: Bob Handelman/Getty Images

It's worth dwelling on the example of car manufacturers because the world of algorithms - a very young, highly risky new industry with no safety precautions in place - is rather like the early car industry. With its naive and exuberant faith in its own technology, the world of AI is selling the equivalent of cars without bumpers whose wheels might fall off at any moment. And I'm sure there were such cars made once upon a time, but over time, as we saw more damage being done by faulty design, we came up with more rules to protect passengers and pedestrians. So, what can we learn from the current, mature world of car makers in the context of illegal software?

First, similar types of software are being deployed by other car manufacturers that turn off emissions controls in certain settings. In other words, this was not a situation in which there was only one bad actor, but rather a standard operating procedure. Moreover, we can assume this doesn't represent collusion, but rather a simple case of extreme incentives combined with a calculated low probability of getting caught on the part of the car manufacturers. It's reasonable to expect, then, that there are plenty of other algorithms being used to skirt rules and regulations deemed too expensive, especially when the builders of the algorithms remain smug about their chances.

Next, the VW cheating started in 2009, which means it went undetected for five years. What else has been going on for five years? This line of thinking makes us start looking around, wondering which companies are currently hoodwinking regulators, evading privacy laws, or committing algorithmic fraud with impunity.

Indeed it might seem like a slam dunk business model, in terms of cost-benefit analysis: cheat until regulators catch up with us, if they ever do, and then pay a limited fine that

doesn't make much of a dent in our cumulative profit. That's how it worked in the aftermath of the financial crisis, after all. In the name of shareholder value, we might be obliged to do this.

Put it another way. We're all expecting cars to be self-driving in a few years or a couple of decades at most. When that happens, can we expect there to be international agreements on what the embedded self-driving car ethics will look like? Or will pedestrians be at the mercy of the car manufacturers to decide what happens in the case of an unexpected pothole? If we get rules, will the rules differ by country, or even by the country of the manufacturer?

If this sounds confusing for something as easy to observe as car crashes, imagine what's going on under the hood, in the relatively obscure world of complex "deep learning" models.

The tools are there already, to be sure. China has recently demonstrated how well facial recognition technology already works - enough to catch jaywalkers and toilet paper thieves. That means there are plenty of opportunities for companies to perform devious tricks on customers or potential hires. For that matter, the incentives are also in place. Just last month Google was fined €2.4bn for unfairly placing its own shopping search results in a more prominent place than its competitors. A similar complaint was levelled at Amazon by ProPublica last year with respect to its pricing algorithm, namely that it was privileging its own, in-house products - even when they weren't a better deal - over those outside its marketplace. If you think of the internet as a place where big data companies vie for your attention, then we can imagine more algorithms like this in our future.

There's a final parallel to draw with the VW scandal. Namely, the discrepancy in emissions was finally discovered in 2014 by a team of professors and students at West Virginia University, who applied and received a measly grant of $50,000 from the International Council on Clean Transportation, an independent nonprofit organisation paid for by US taxpayers. They spent their money driving cars around the country and capturing the emissions, a cheap and straightforward test.

In 2015, Volkswagen was found to have used a malicious algorithm to deceive the emissions test. Seven VW executives have been charged in the US. Photograph: Patrick T Fallon/Bloomberg/Getty

What organisation will put a stop to the oncoming crop of illegal algorithms? What is the analogue of the International Council on Clean Transportation? Does there yet exist an organisation that has the capacity, interest, and ability to put an end to illegal algorithms, and to prove that these algorithms are harmful? The answer is, so far, no. Instead, at least in the US, a disparate group of federal agencies is in charge of enforcing laws in their industry or domain, none of which is particularly on top of the complex world of big data algorithms. Elsewhere, the European commission seems to be looking into Google's antitrust activity, and Facebook's fake news problems, but that leaves multiple industries untouched by scrutiny.

Even more to the point, though, is the question of how involved the investigation of algorithms would have to be. The current nature of algorithms is secret, proprietary code, protected as the "secret sauce" of corporations. They're so secret that most online scoring systems aren't even apparent to the people targeted by them. That means those people also don't know the score they've been given, nor can they complain about or contest those scores. Most important, they typically won't know if something unfair has happened to them.

Given all of this, it's difficult to imagine oversight for algorithms, even when they've gone wrong and are actively harming people. For that matter, not all kinds of harm are distinctly measurable in the first place. One can make the argument that, what with all the fake news floating around, our democracy has been harmed. But how do you measure democracy?

That's not to say there is no hope. After all, by definition, an illegal algorithm is breaking an actual law that we can point to. There is, ultimately, someone that should be held accountable for this. The problem still remains, how will such laws be enforced?

Ben Shneiderman, a computer science professor at the University of Maryland, proposed the concept of a National Algorithms Safety Board, in a talk at the Alan Turing Institute. Modelled on the National Transportation Safety Board, which investigates ground and air traffic accidents, this body would similarly be charged with investigating harm, and specifically in deciding who should be held responsible for algorithmic harm.

Algorithms sift through historical data to value homes. In the US, one homeowner is suing Zoopla for knocking $100,000 from the value of her property by drawing on the wrong data.
Photograph: Yui Mok/PA

This is a good idea. We should investigate problems when we find them, and it's good to have a formal process to do so. If it has sufficient legal power, the board can perhaps get to the bottom of lots of commonsense issues. But it's not clear how comprehensive it could be.

Because here's where the analogy with car makers breaks down: there is no equivalent of a 30-car pile-up in the world of algorithms. Most of the harm happens to isolated individuals, separately and silently. A proliferation of silent and undetectable car crashes is harder to investigate than when it happens in plain sight.

I'd still maintain there's hope. One of the miracles of being a data sceptic in a land of data evangelists is that people are so impressed with their technology, even when it is unintentionally creating harm, they openly describe how amazing it is. And the fact that we've already come across quite a few examples of algorithmic harm means that, as secret and opaque as these algorithms are, they're eventually going to be discovered, albeit after they've caused a lot of trouble.

What does this mean for the future? First and foremost, we need to start keeping track. Each criminal algorithm we discover should be seen as a test case. Do the rule-breakers get into trouble? How much? Are the rules enforced, and what is the penalty? As we learned after the 2008 financial crisis, a rule is ignored if the penalty for breaking it is less than the profit pocketed. And that goes double for a broken rule that is only discovered half the time.

Even once we start building a track record of enforcement, we have ourselves an arms race. We can soon expect a fully fledged army of algorithms that skirt laws, that are sophisticated and silent, and that seek to get around rules and regulations. They will learn from how others were caught and do it better the next time. In other words, it will get progressively more difficult to catch them cheating. Our tactics have to get better over time too.

Predictive policing algorithms use historical data to forecast where crime will happen next. Civil rights groups argue that these systems exacerbate existing police prejudices. Photograph: Stuart Emmerson/Alamy

We can also expect to be told that the big companies are "dealing with it privately". This is already happening with respect to fighting terrorism. We should not trust them when they say this. We need to create a standard testing framework - a standard definition of harm -

and require that algorithms be submitted for testing. And we cannot do this only in "test lab conditions," either, or we will be reconstructing the VW emissions scandal.

One of the biggest obstacles to this is that Google, Facebook, or for that matter Amazon, don't allow testing of multiple personas - or online profiles - by outside researchers. Since those companies offer tailored and personalised service, the only way to see what that service looks like would be to take on the profile of multiple people, but that is not allowed. Think about that in the context of the VW testing: it would be like saying research teams could not have control of a car to test its emissions. We need to demand more access and ongoing monitoring, especially once we catch them in illegal acts. For that matter, entire industries, such as algorithms for insurance and hiring, should be subject to these monitors, not just individual culprits.

It's time to gird ourselves for a fight. It will eventually be a technological arms race, but it starts, now, as a political fight. We need to demand evidence that algorithms with the potential to harm us be shown to be acting fairly, legally, and consistently. When we find problems, we need to enforce our laws with sufficiently hefty fines that companies don't find it profitable to cheat in the first place. This is the time to start demanding that the machines work for us, and not the other way around.

Cathy O'Neil is the author of Weapons of Math Destruction (Allen Lane £9.99). To order a copy for £8.49, go to bookshop.theguardian.com or call 0330 333 6846

Since you're here ...

... we have a small favour to ask. More people are reading the Guardian than ever but advertising revenues across the media are falling fast. And unlike many news organisations,

we haven't put up a paywall - we want to keep our journalism as open as we can. So you can see why we need to ask for your help. The Guardian's independent, investigative journalism takes a lot of time, money and hard work to produce. But we do it because we believe our perspective matters - because it might well be your perspective, too.

I appreciate there not being a paywall: it is more democratic for the media to be available for all and not a commodity to be purchased by a few. I'm happy to make a contribution so others with less means still have access to information. *Thomasine F-R.*
If everyone who reads our reporting, who likes it, helps to support it, our future would be much more secure.

Become a supporter
Make a contribution
Topics

- Big data
- The Observer
- Data protection
- Privacy
- Internet
- features