‘Algorithmic death spiral’: The failing mental health of our machines

Is my car hallucinating? Is the algorithm that runs the police surveillance system in my city paranoid? Marvin the android in Douglas Adams’s *Hitchhikers Guide to the Galaxy* had a pain in all the diodes down his left-hand side. Is that how my toaster feels?

This all sounds ludicrous until we realise that our algorithms are increasingly being made in our own image. As we’ve learned more about our own brains, we’ve enlisted that knowledge to create algorithmic versions of ourselves. These algorithms control the speeds of driverless cars, identify targets for autonomous military drones, compute our susceptibility to commercial and political advertising, find our soulmates in online dating services, and evaluate our insurance and credit risks. Algorithms are becoming the near-sentient backdrop of our lives.

The most popular algorithms currently being put into the workforce are deep learning algorithms. These algorithms mirror the architecture of human brains by building complex representations of information. They learn to understand environments by experiencing them, identify what seems to matter, and figure out what predicts what. Being like our brains, these algorithms are increasingly at risk of mental-health problems.

Deep Blue, the algorithm that beat the world chess champion Garry Kasparov in 1997, did so through brute force, examining millions of positions a second, up to 20 moves in the future. Anyone could understand how it worked even if they couldn’t do it themselves. AlphaGo, the deep learning algorithm that beat Lee Sedol at the game of Go in 2016, is fundamentally different. Using deep neural networks, it created its own understanding of the game, considered to be the most complex of board games. AlphaGo learned by watching others and by playing itself. Computer scientists and Go players alike are befuddled by AlphaGo’s unorthodox play. Its strategy seems at first to be awkward. Only in retrospect do we understand what AlphaGo was thinking, and even then it’s not all that clear.

Garry Kasparov playing against Deep Blue, the chess-playing computer built by IBM.
Image credit: Adam Nadel, AP Images

To give you a better understanding of what I mean by thinking, consider this. Programs such as Deep Blue can have a bug in their programming. They can crash from memory overload. They can enter a state of paralysis due to a neverending loop or simply spit out the wrong answer on a lookup table. But all of these problems are solvable by a programmer with access to the source code, the code in which the algorithm was written.

Algorithms such as AlphaGo are entirely different. Their problems are not apparent by looking at their source code. They are embedded in the way that they represent information. That representation is an ever-changing high-dimensional space, much like walking around in a dream. Solving problems there requires nothing less than a psychotherapist for algorithms.

Take the case of driverless cars. A driverless car that sees its first stop sign in the real world will have already seen millions of stop signs during training, when it built up its mental representation of what a stop sign is. Under various light conditions, in good weather and bad, with and without bullet holes, the stop signs it was exposed to contain a bewildering variety of information. Under most normal conditions, the driverless car will recognise a stop sign for what it is. But not all conditions are normal. Some recent demonstrations have shown that a few black stickers on a stop sign can fool the algorithm into thinking that the stop sign is a 60 mph sign. Subjected to something frighteningly similar to the high-contrast shade of a tree, the algorithm hallucinates.

How many different ways can the algorithm hallucinate? To find out, we would have to provide the algorithm with all possible combinations of input stimuli. This means that there are potentially infinite ways in which it can go wrong. Crackerjack programmers already know this, and take advantage of it by creating what are called adversarial examples. The AI research group LabSix at the Massachusetts Institute of Technology has shown that, by presenting images to Google's image-classifying algorithm and using the data it sends back, they can identify the algorithm's weak spots. They can then do things similar to fooling Google's image-recognition software into believing that an X-rated image is just a couple of puppies playing in the grass.

Algorithms also make mistakes because they pick up on features of the environment that are correlated with outcomes, even when there is no causal relationship between them. In the algorithmic world, this is called overfitting. When this happens in a brain, we call it superstition.

The biggest algorithmic failure due to superstition that we know of so far is called the parable of Google Flu. Google Flu used what people type into Google to predict the location and intensity of influenza outbreaks. Google Flu's predictions worked fine at first, but they grew worse over time, until eventually it was predicting twice the number of cases as were submitted to the US Centers for Disease Control. Like an algorithmic witchdoctor, Google Flu was simply paying attention to the wrong things.

Algorithmic pathologies might be fixable. But in practice, algorithms are often proprietary black boxes whose updating is commercially protected. Cathy O'Neil's *Weapons of Math Destruction* (2016) describes a veritable freakshow of commercial algorithms whose insidious pathologies play out collectively to ruin peoples' lives. The algorithmic faultline that separates the wealthy from the poor is particularly compelling. Poorer people are more likely to have bad credit, to live in high-crime areas, and to be surrounded by other poor people with similar problems. Because of this, algorithms target these individuals for misleading ads that prey on their desperation, offer them subprime loans, and send more police to their neighbourhoods, increasing the likelihood that they will be stopped by police for crimes committed at similar rates in wealthier neighbourhoods. Algorithms used by the judicial system give these individuals longer prison sentences, reduce their chances for parole, block them from jobs, increase their mortgage rates, demand higher premiums for insurance, and so on.

deep 3 29 18 3

Image not found or type unknown

Self-driving car. Image credit: Uber

This algorithmic death spiral is hidden in nesting dolls of black boxes: black-box algorithms that hide their processing in high-dimensional thoughts that we can't access are further hidden in black boxes of proprietary ownership. This has prompted some places, such as New York City, to propose laws enforcing the monitoring of fairness in algorithms used by municipal services. But if we can't detect bias in ourselves, why would we expect to detect it in our algorithms?

By training algorithms on human data, they learn our biases. One recent study led by Aylin Caliskan at Princeton University found that algorithms trained on the news learned racial and gender biases essentially overnight. As Caliskan noted: 'Many people think machines are not biased. But machines are trained on human data. And humans are biased.'

Social media is a writhing nest of human bias and hatred. Algorithms that spend time on social media sites rapidly become bigots. These algorithms are biased against male nurses and female engineers. They will view issues such as immigration and minority rights in ways that don't stand up to investigation. Given half a chance, we should expect algorithms to treat people as unfairly as people treat each other. But algorithms are by construction overconfident, with no sense of their own infallibility. Unless they are trained to do so, they have no reason to question their incompetence (much like people).

For the algorithms I've described above, their mental-health problems come from the quality of the data they are trained on. But algorithms can also have mental-health problems based on the way they are built. They can forget older things when they learn new information. Imagine learning a new co-worker's name and suddenly forgetting where you live. In the extreme, algorithms can suffer from what is called catastrophic forgetting, where the entire algorithm can no longer learn or remember anything. A theory of human age-related cognitive decline is based on a similar idea: when memory becomes overpopulated, brains and desktop computers alike require more time to find what they know.

When things become pathological is often a matter of opinion. As a result, mental anomalies in humans routinely go undetected. Synaesthetes such as my daughter, who perceives written letters as colours, often don't realise that they have a perceptual gift until they're in their teens. Evidence based on Ronald Reagan's speech patterns now suggests that he probably had dementia while in office as US president. And *The Guardian* reports that the mass shootings that have occurred every nine out of 10 days for roughly the past five years in the US are often perpetrated by so-called 'normal' people who happen to break under feelings of persecution and depression.

In many cases, it takes repeated malfunctioning to detect a problem. Diagnosis of schizophrenia requires at least one month of fairly debilitating symptoms. Antisocial personality disorder, the modern term for psychopathy and sociopathy, cannot be diagnosed in individuals until they are 18, and then only if there is a history of conduct disorders before the age of 15.

There are no biomarkers for most mental-health disorders, just like there are no bugs in the code for AlphaGo. The problem is not visible in our hardware. It's in our software. The many ways our minds go wrong make each mental-health problem unique unto itself. We sort them into broad categories such as schizophrenia and Asperger's syndrome, but most are spectrum disorders that cover symptoms we all share to different degrees. In 2006, the psychologists Matthew Keller and Geoffrey Miller argued that this is an inevitable property of the way that brains are built.

There is a lot that can go wrong in minds such as ours. Carl Jung once suggested that in every sane man hides a lunatic. As our algorithms become more like ourselves, it is getting easier to hide.

Thomas T. Hills is professor of psychology at the University of Warwick in Coventry, UK. Follow him on Twitter @thomhills

This article was originally published at Aeon as "Does my algorithm have a mental-health problem?" and has been republished here with permission.