The tiny changes that can cause AI to fail

Machines still have a long way to go before they learn like humans do – and that's a potential danger to privacy, safety, and more.

By Aviva Hope Rutkin
11 April 2017

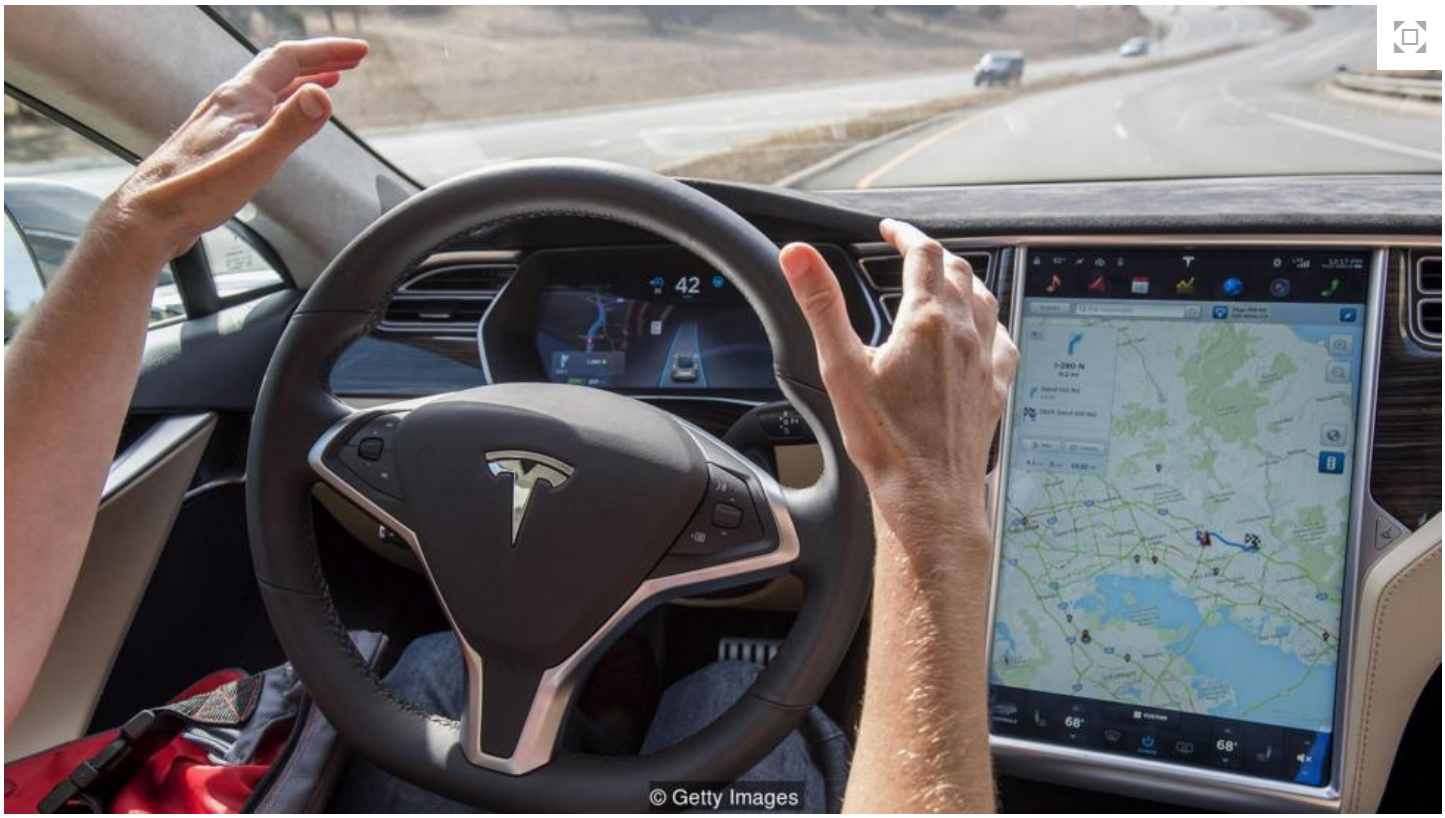
futurenow More articles >

The year is 2022. You're riding along in a self-driving car on a routine trip through the city. The car comes to a stop sign it's passed a hundred times before – but this time, it blows right through it.

To you, the stop sign looks exactly the same as any other. But to the car, it looks like something entirely different. Minutes earlier, unbeknownst to either you or the machine, a scam artist stuck a small sticker onto the sign: unnoticeable to the human eye, inescapable to the technology.

In other words? The tiny sticker smacked on the sign is enough for the car to “see” the stop sign as something completely different from a stop sign.

It may sound far-fetched. But a growing field of research proves that artificial intelligence can be fooled in more or less the same way, seeing one thing where humans would see something else entirely. As machine learning algorithms **increasingly find their way** into our roads, our finances, our healthcare system, computer scientists hope to learn more about how to defend them against these “adversarial” attacks – before someone tries to bamboozle them for real.



Artificial intelligence fuels our everyday lives in increasingly inextricable ways, from self-driving cars to household appliances that self-activate (Credit: Getty Images)

“It’s something that’s a growing concern in the machine learning and AI community, especially because these algorithms are being used more and more,” says Daniel Lowd, assistant professor of computer and information science at the University of Oregon. “If spam gets through or a few emails get blocked, it’s not the end of the word. On the other hand, if you’re relying on the vision system in a self-driving car to know where to go and not crash into anything, then the stakes are much higher.”

Whether or not a smart machine malfunctions, or is hacked, hinges on the very different way that machine learning algorithms 'see' the world. In this way, to a machine, a panda could look like a gibbon, or a school bus could read as an ostrich.

In one experiment, researchers from France and Switzerland showed how such perturbations could cause a computer to mistake a squirrel for an grey fox, or a coffee pot for a macaw.

How can this be? Think of a child learning to recognise numbers. As they look at each one in turn, she starts to pick up on certain common characteristics: ones are tall and slender, sixes and nines contain one big loop while eights have two, and so on. Once they've seen enough examples, they can quickly recognise new digits as fours or eights or threes – even if, thanks to the font or the handwriting, it doesn't look exactly like any other four or eight or three they've ever seen before.

> **In this way, to a machine, a panda could look like a gibbon, or a school bus could read as an ostrich**

Machine learning algorithms learn to read the world through a somewhat similar process. Scientists will feed a computer with hundreds or thousands of (usually labelled) examples of whatever it is they'd like the computer to detect. As the machine sifts through the data – this is a number, this is not, this is a number, this is not – it starts to pick up on features that give the answer away. Soon, it's able to look at a picture and declare, "This is a five!" with high accuracy.

In this way, both human children and computers alike can learn to recognise a huge array of objects, from numbers to cats to boats to individual human faces.

But, unlike a human child, the computer isn't paying attention to high-level details like a cat's furry ears or the number four's distinctive angular shape. It's not considering the whole picture.

Instead, it's likely looking at the individual pixels of the picture – and for the fastest way to tell objects apart. If the vast majority of number ones have a black pixel in one particular spot and a couple of white pixels in another particular spot, then the machine may make a call after only checking that handful of pixels.

Now, think back to the stop sign again. With an imperceptible tweak to the pixels of the image – or what experts call "perturbations" – the computer is fooled into thinking that the stop sign is something it isn't.

> **If these vulnerabilities exist, someone will figure out how to exploit them. Someone likely already has**

Similar research from the Evolving Artificial Intelligence Laboratory at the University of Wyoming and Cornell University has produced a **bounty of optical illusions** for artificial intelligence. These psychedelic images of abstract patterns and colours look like nothing much to humans, but are rapidly

recognised by the computer as snakes or rifles. These suggest how AI can look at something and be way off base as to what the object actually is or looks like.

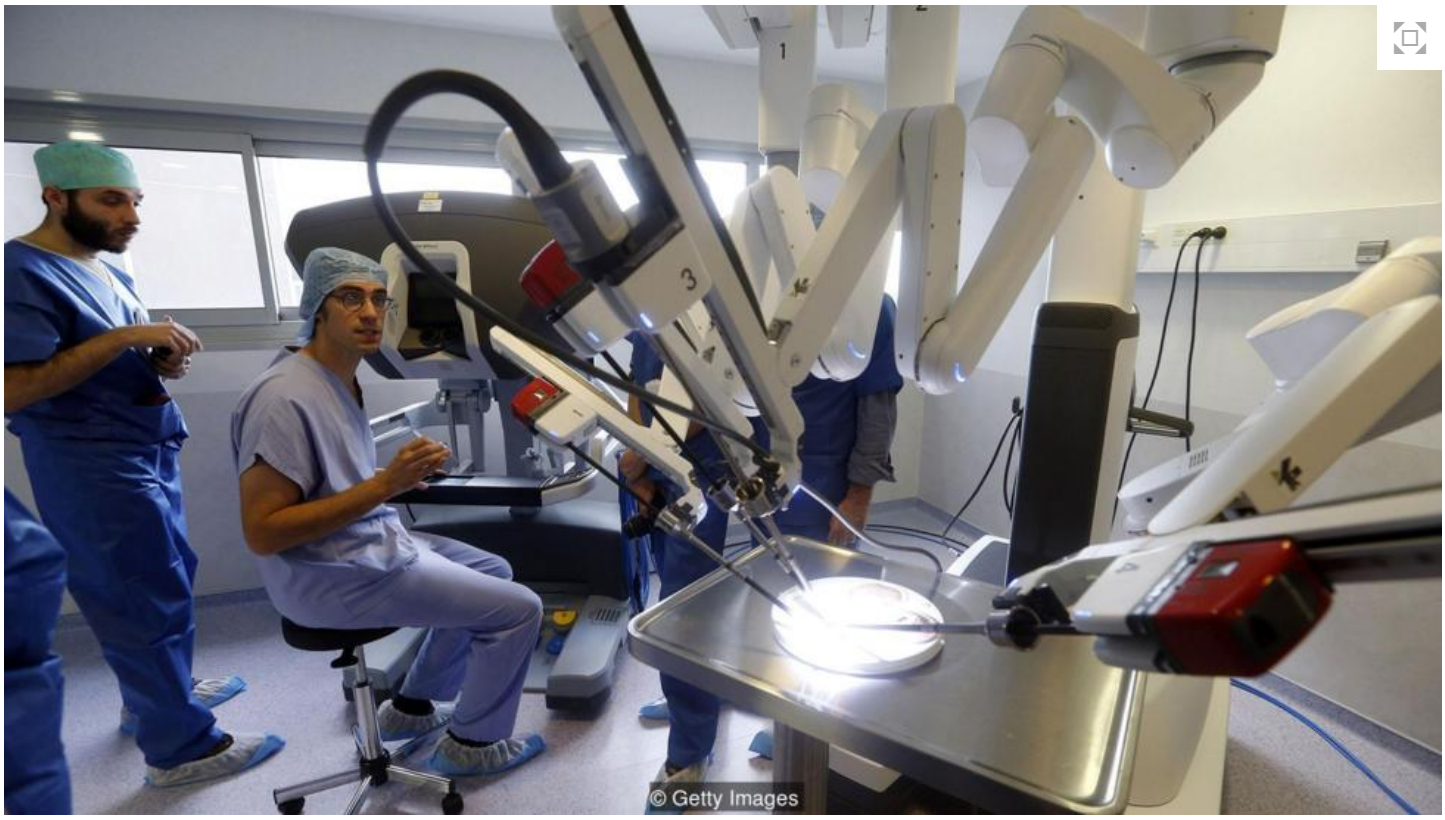
This weakness is common across all types of machine learning algorithms. “One would expect every algorithm has a chink the armour,” says Yevgeniy Vorobeychik, assistant professor of computer science and computer engineering at Vanderbilt University. “We live in a really complicated multi-dimensional world, and algorithms, by their nature, are only focused on a relatively small portion of it.”

Voyobeychik is “very confident” that, if these vulnerabilities exist, someone will figure out how to exploit them. Someone likely already has.

Consider spam filters, automated programmes that weed out any dodgy-looking emails. Spammers can try to scale over the wall by tweaking the spelling of words (Viagra to Vi@gra) or by appending a list of “good words” typically found in legitimate emails: words like, according to one algorithm, “glad”, “me” or “yup”. Meanwhile, spammers could try to drown out words that often pop up in illegitimate emails, like “claim” or “mobile” or “won”.

What might this allow scammers to one day pull off? That self-driving car hoodwinked by a stop sign sticker is a classic scenario that’s been floated by experts in the field. Adversarial data might help slip porn past safe-content filters. Others might try to boost the numbers on a cheque. Or hackers could tweak the code of malicious software just enough to slip undetected past digital security.

Troublemakers can figure out how to create adversarial data if they have a copy of the machine learning algorithm they want to fool. But that’s not necessary for sneaking through the algorithm’s doors. They can simply brute-force their attack, throwing slightly different versions of an email or image or whatever it is against the wall until one gets through. Over time, this could even be used to generate a new model entirely, one that learns what the good guys are looking for and how to produce data that fools them.



Autonomous vehicles and surgical robots put a lot on the line, so modern machines leave little room for error (Credit: Getty Images)

“People have been manipulating machine learning systems since they were first introduced,” says Patrick McDaniel, professor of computer science and engineering at Pennsylvania State University. “If people are using these techniques in the wild, we might not know it.”

Scammers might not be the only ones to make hay while the sun shines. Adversarial approaches could come in handy for people hoping to avoid the X-ray eyes of modern technology.

“If you’re some political dissident inside a repressive regime and you want to be able to conduct activities without being targeted, being able to avoid automated surveillance techniques based on machine learning would be a positive use,” says Lowd.

In **one project**, published in October, researchers at Carnegie Mellon University built a pair of glasses that can subtly mislead a facial recognition system – making the computer confuse actress Reese Witherspoon for Russell Crowe. It sounds playful, but such technology could be handy for someone desperate to avoid censorship by those in power.

> **McDaniel suggests we consider leaving humans in the loop when we can, providing some sort of external verification**

In the meantime, what's an algorithm to do? "The only way to completely avoid this is to have a perfect model that is right all the time," says Lowd. Even if we could build artificial intelligence that bested humans, the world would still contain ambiguous cases where the right answer wasn't readily apparent.

Machine learning algorithms are usually scored by their accuracy. A programme that recognises chairs 99% of the time is obviously better than one that only hits the mark six times out of 10. But some experts now argue that they should also measure how well the algorithm can handle an attack: the tougher, the better.

Another solution might be for experts to put the programmes through their paces. Create your own example attacks in the lab based on what you think perpetrators might do, then show them to the machine learning algorithm. This could help it become more resilient over time – provided, of course, that the test attacks match the type that will be tried in the real world.

McDaniel suggests we consider leaving humans in the loop when we can, providing some sort of external verification that the algorithms' guesses are correct. Some "intelligent assistants", like Facebook's M, have humans double-check and soup up their answers; others have suggested that human checks could be useful in sensitive applications such as court judgments.

"Machine learning systems are a tool to do reasoning. We need to be smart and rational about what we give them and what they tell us," he says. "We shouldn't treat them as perfect oracles of truth."