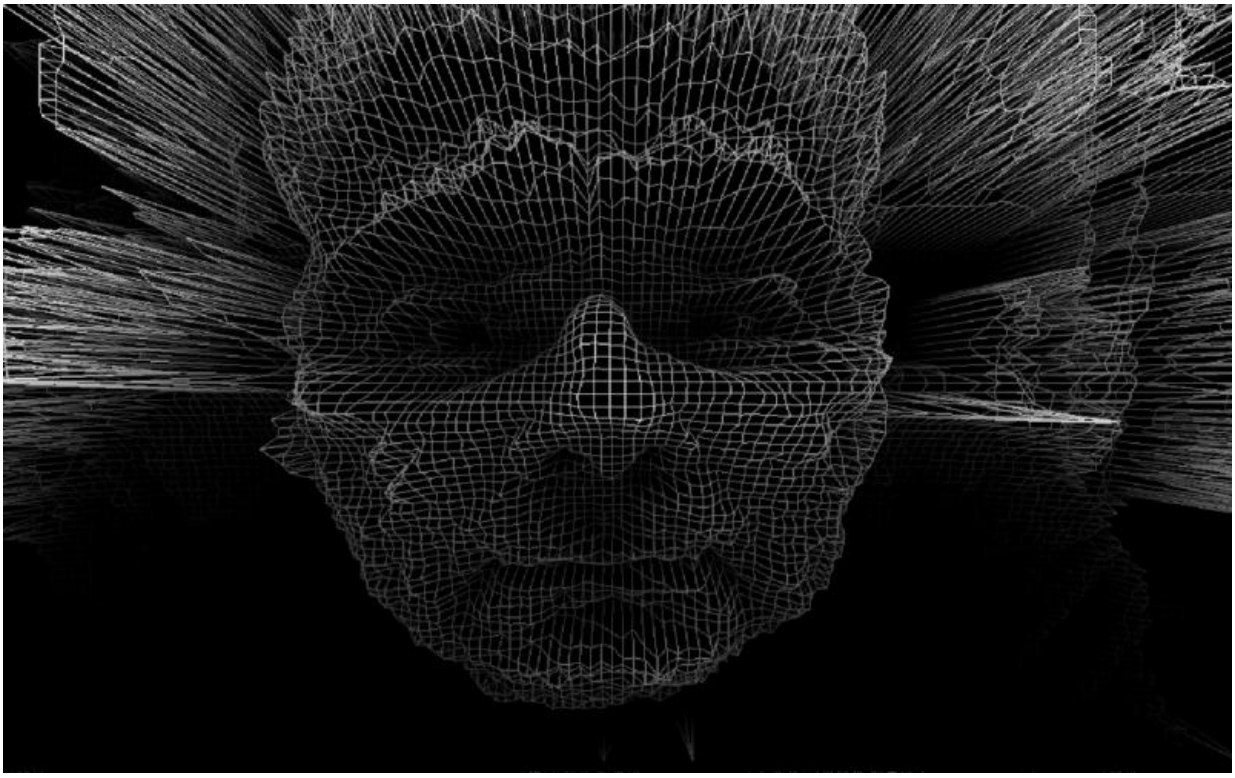
Enhancing the reliability of artificial intelligence

October 21 2016



Credit: Thierry Ehrmann

Computers that learn for themselves are with us now. As they become more common in 'high-stakes' applications like robotic surgery, terrorism detection and driverless cars, researchers ask what can be done to make sure we can trust them.

There would always be a first death in a driverless car and it happened in May 2016. Joshua Brown had engaged the autopilot system in his Tesla when a tractor-trailor drove across the road in front of him. It seems that neither he nor the sensors in the autopilot noticed the white-sided truck against a brightly lit sky, with tragic results.

Of course many people die in car crashes every day – in the USA there is one fatality every 94 million miles, and according to Tesla this was the first known fatality in over 130 million miles of driving with activated autopilot. In fact, given that most road fatalities are the result of human error, it has been said that autonomous cars should make travelling safer.

Even so, the tragedy raised a pertinent question: how much do we understand – and trust – the computers in an autonomous vehicle? Or, in fact, in any machine that has been taught to carry out an activity that a human would do?

We are now in the era of machine learning. Machines can be trained to recognise certain patterns in their environment and to respond appropriately. It happens every time your digital camera detects a face and throws a box around it to focus, or the personal assistant on your smartphone answers a question, or the adverts match your interests when you search online.

Machine learning is a way to program computers to learn from experience and improve their performance in a way that resembles how humans and animals learn tasks. As machine learning techniques become more common in everything from finance to healthcare, the issue of trust is becoming increasingly important, says Zoubin Ghahramani, Professor of Information Engineering in Cambridge's Department of Engineering.

Faced with a life or death decision, would a driverless car decide to hit

pedestrians, or avoid them and risk the lives of its occupants? Providing a medical diagnosis, could a machine be wildly inaccurate because it has based its opinion on a too-small sample size? In making financial transactions, should a computer explain how robust is its assessment of the volatility of the stock markets?

"Machines can now achieve near-human abilities at many cognitive tasks even if confronted with a situation they have never seen before, or an incomplete set of data," says Ghahramani. "But what is going on inside the 'black box'? If the processes by which decisions were being made were more transparent, then trust would be less of an issue."

His team builds the algorithms that lie at the heart of these technologies (the "invisible bit" as he refers to it). Trust and transparency are important themes in their work: "We really view the whole mathematics of machine learning as sitting inside a framework of understanding uncertainty. Before you see data – whether you are a baby learning a language or a scientist analysing some data – you start with a lot of uncertainty and then as you have more and more data you have more and more certainty.

"When machines make decisions, we want them to be clear on what stage they have reached in this process. And when they are unsure, we want them to tell us."

One method is to build in an internal self-evaluation or calibration stage so that the machine can test its own certainty, and report back.

Two years ago, Ghahramani's group launched the Automatic Statistician with funding from Google. The tool helps scientists analyse datasets for statistically significant patterns and, crucially, it also provides a report to explain how sure it is about its predictions.

"The difficulty with machine learning systems is you don't really know what's going on inside – and the answers they provide are not contextualised, like a human would do. The Automatic Statistician explains what it's doing, in a human-understandable form."

Where transparency becomes especially relevant is in applications like medical diagnoses, where understanding the provenance of how a decision is made is necessary to trust it.

Dr Adrian Weller, who works with Ghahramani, highlights the difficulty: "A particular issue with new artificial intelligence (AI) systems that learn or evolve is that their processes do not clearly map to rational decision-making pathways that are easy for humans to understand." His research aims both at making these pathways more transparent, sometimes through visualisation, and at looking at what happens when systems are used in real-world scenarios that extend beyond their training environments – an increasingly common occurrence.

"We would like AI systems to monitor their situation dynamically, detect whether there has been a change in their environment and – if they can no longer work reliably – then provide an alert and perhaps shift to a safety mode." A driverless car, for instance, might decide that a foggy night in heavy traffic requires a human driver to take control.

Weller's theme of trust and transparency forms just one of the projects at the newly launched £10 million Leverhulme Centre for the Future of Intelligence (CFI). Ghahramani, who is Deputy Director of the Centre, explains: "It's important to understand how developing technologies can help rather than replace humans. Over the coming years, philosophers, social scientists, cognitive scientists and computer scientists will help guide the future of the technology and study its implications – both the concerns and the benefits to society."

CFI brings together four of the world's leading universities (Cambridge, Oxford, Berkeley and Imperial College, London) to explore the implications of AI for human civilisation. Together, an interdisciplinary community of researchers will work closely with policy-makers and industry investigating topics such as the regulation of autonomous weaponry, and the implications of AI for democracy.

Ghahramani describes the excitement felt across the machine learning field: "It's exploding in importance. It used to be an area of research that was very academic – but in the past five years people have realised these methods are incredibly useful across a wide range of societally important areas.

"We are awash with data, we have increasing computing power and we will see more and more applications that make predictions in real time. And as we see an escalation in what machines can do, they will challenge our notions of intelligence and make it all the more important that we have the means to trust what they tell us."

Provided by University of Cambridge