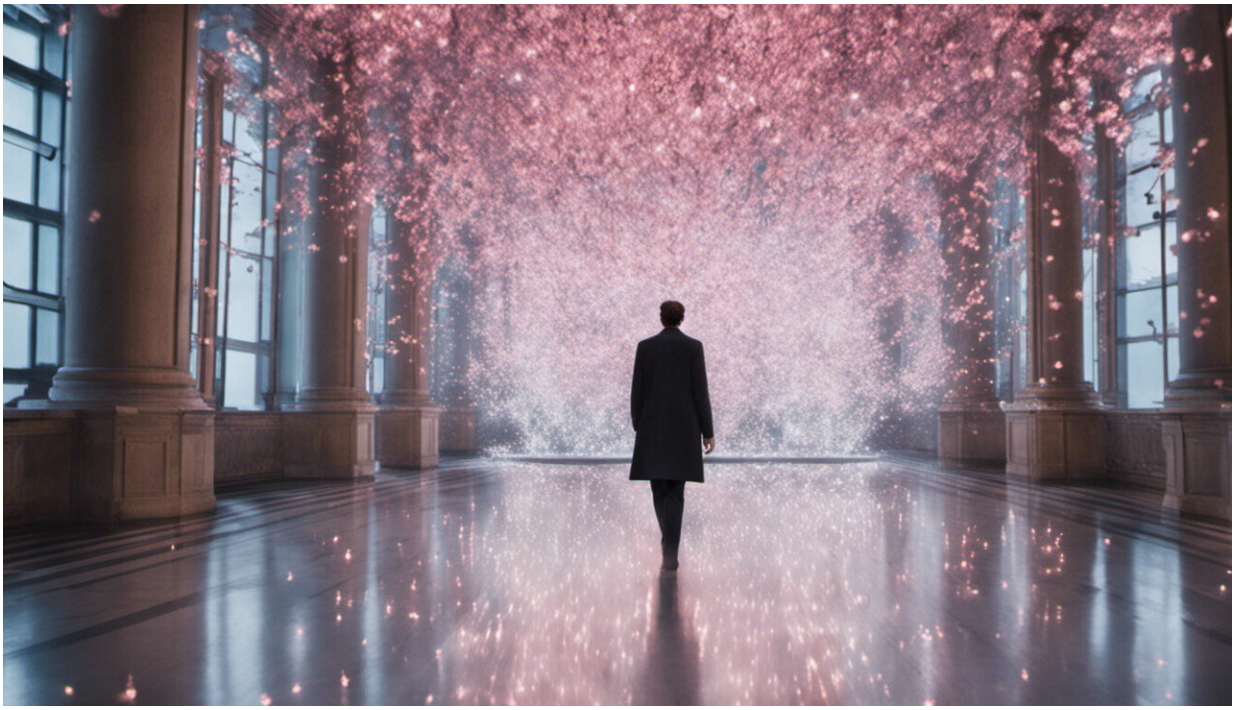
People don't trust AI—here's how we can change that

January 10 2018, by Vyacheslav Polonski



Credit: AI-generated image (disclaimer)

Artificial intelligence can already predict the future. Police forces are using it to map when and where crime is likely to occur. Doctors can use it to predict when a patient is most likely to have a heart attack or stroke. Researchers are even trying to give AI imagination so it can plan for unexpected consequences.

Many decisions in our lives require a good forecast, and AI agents are almost always better at forecasting than their human counterparts. Yet for all these technological advances, we still seem to deeply lack confidence in AI predictions. Recent cases show that people don't like relying on AI and prefer to trust human experts, even if these experts are wrong.

If we want AI to really benefit people, we need to find a way to get people to trust it. To do that, we need to understand why people are so reluctant to trust AI in the first place.

Should you trust Dr. Robot?

IBM's attempt to promote its supercomputer programme to cancer doctors (Watson for Onology) was a PR disaster. The AI promised to deliver top-quality recommendations on the treatment of 12 cancers that accounted for 80% of the world's cases. As of today, over 14,000 patients worldwide have received advice based on its calculations.

But when doctors first interacted with Watson they found themselves in a rather difficult situation. On the one hand, if Watson provided guidance about a treatment that coincided with their own opinions, physicians did not see much value in Watson's recommendations. The supercomputer was simply telling them what they already know, and these recommendations did not change the actual treatment. This may have given doctors some peace of mind, providing them with more confidence in their own decisions. But IBM has yet to provide evidence that Watson actually improves cancer survival rates.

On the other hand, if Watson generated a recommendation that contradicted the experts' opinion, doctors would typically conclude that Watson wasn't competent. And the machine wouldn't be able to explain why its treatment was plausible because its machine learning algorithms

were simply too complex to be fully understood by humans. Consequently, this has caused even more mistrust and disbelief, leading many doctors to ignore the seemingly outlandish AI recommendations and stick to their own expertise.

As a result, IBM Watson's premier medical partner, the MD Anderson Cancer Center, recently announced it was dropping the programme. Similarly, a Danish hospital reportedly abandoned the AI programme after discovering that its cancer doctors disagreed with Watson in over two thirds of cases.

The problem with Watson for Oncology was that doctors simply didn't trust it. Human trust is often based on our understanding of how other people think and having experience of their reliability. This helps create a psychological feeling of safety. AI, on the other hand, is still fairly new and unfamiliar to most people. It makes decisions using a complex system of analysis to identify potentially hidden patterns and weak signals from large amounts of data.

Even if it can be technically explained (and that's not always the case), AI's decision-making process is usually too difficult for most people to understand. And interacting with something we don't understand can cause anxiety and make us feel like we're losing control. Many people are also simply not familiar with many instances of AI actually working, because it often happens in the background.

Instead, they are acutely aware of instances where AI goes wrong: a Google algorithm that classifies people of colour as gorillas; a Microsoft chatbot that decides to become a white supremacist in less than a day; a Tesla car operating in autopilot mode that resulted in a fatal accident. These unfortunate examples have received a disproportionate amount of media attention, emphasising the message that we cannot rely on technology. Machine learning is not foolproof, in part because the

humans who design it aren't.

A new AI divide in society?

Feelings about AI also run deep. My colleagues and I recently ran an experiment where we asked people from a range of backgrounds to watch various sci-fi films about AI and then asked them questions about automation in everyday life. We found that, regardless of whether the film they watched depicted AI in a positive or negative light, simply watching a cinematic vision of our technological future polarised the participants' attitudes. Optimists became more extreme in their enthusiasm for AI and sceptics became even more guarded.

This suggests people use relevant evidence about AI in a biased manner to support their existing attitudes, a deep-rooted human tendency known as confirmation bias. As AI is reported and represented more and more in the media, it could contribute to a deeply divided society, split between those who benefit from AI and those who reject it. More pertinently, refusing to accept the advantages offered by AI could place a large group of people at a serious disadvantage.

Three ways out of the AI trust crisis

Fortunately we already have some ideas about how to improve trust in AI. Simply having previous experience with AI can significantly improve people's attitudes towards the technology, as we found in our study. Similar evidence also suggests the more you use other technologies such as the internet, the more you trust them.

Another solution may be to open the "black-box" of machine learning algorithms and be more transparent about how they work. Companies such as Google, Airbnb and Twitter already release transparency reports

about government requests and surveillance disclosures. A similar practice for AI systems could help people have a better understanding of algorithmic decisions are made.

Research suggests involving people more in the AI decision-making process could also improve trust and allow the AI to learn from human experience. For example,one study showed people were given the freedom to slightly modify an algorithm felt more satisfied with its decisions, more likely to believe it was superior and more likely to use it in the future.

We don't need to understand the intricate inner workings of AI systems, but if people are given at least a bit of information about and control over how they are implemented, they will be more open to accepting AI into their lives.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation