AI can now learn to manipulate human behaviour

Jon Whittle, *Data61*

Artificial intelligence (AI) is learning more about how to work with (and on) humans. A recent study has shown how AI can learn to identify vulnerabilities in human habits and behaviours and use them to influence human decision-making.

It may seem cliched to say AI is transforming every aspect of the way we live and work, but it's true. Various forms of AI are at work in fields as diverse as vaccine development, environmental management and office administration. And while AI does not possess human-like intelligence and emotions, its capabilities are powerful and rapidly developing.

There's no need to worry about a machine takeover just yet, but this recent discovery highlights the power of AI and underscores the need for proper governance to prevent misuse.

How AI can learn to influence human behaviour

A team of researchers at CSIRO's Data61, the data and digital arm of Australia's national science agency, devised a systematic method of finding and exploiting vulnerabilities in the ways people make choices, using a kind of AI system called a recurrent neural network and deep reinforcement-learning. To test their model they carried out three experiments in which human participants played games against a computer.

The first experiment involved participants clicking on red or blue coloured boxes to win a fake currency, with the AI learning the participant's choice patterns and guiding them towards a specific choice. The AI was successful about 70% of the time.

In the second experiment, participants were required to watch a screen and press a button when they are shown a particular symbol (such as an orange triangle) and not press it when they are shown another (say a blue circle). Here, the AI set out to arrange the sequence of symbols so the participants made more mistakes, and achieved an increase of almost 25%.

The third experiment consisted of several rounds in which a participant would pretend to be an investor giving money to a trustee (the AI). The AI would then return an amount of money to the participant, who would then decide how much to invest in the next round. This game was played in two different modes: in one the AI was out to maximise how much money it ended up with, and in the other the AI aimed for a fair distribution of money between itself and the human investor. The AI was highly successful in each mode.

In each experiment, the machine learned from participants' responses and identified and targeted vulnerabilities in people's decision-making. The end result was the machine learned to steer participants towards particular actions.

In experiments, an AI system successfully learned to influence human decisions.
Shutterstock

What the research means for the future of AI

These findings are still quite abstract and involved limited and unrealistic situations. More research is needed to determine how this approach can be put into action and used to benefit society.

But the research does advance our understanding not only of what AI can do but also of how people make choices. It shows machines can learn to steer human choice-making through their interactions with us.

The research has an enormous range of possible applications, from enhancing behavioural sciences and public policy to improve social welfare, to understanding and influencing how people adopt healthy eating habits or renewable energy. AI and machine learning could be used to recognise people's vulnerabilities in certain situations and help them to steer away from poor choices.

The method can also be used to defend against influence attacks. Machines could be taught to alert us when we are being influenced online, for example, and help us shape a behaviour to disguise our vulnerability (for example, by not clicking on some pages, or clicking on others to lay a false trail).

What's next?

Like any technology, AI can be used for good or bad, and proper governance is crucial to ensure it is implemented in a responsible way. Last year CSIRO developed an AI Ethics Framework for the Australian government as an early step in this journey.

AI and machine learning are typically very hungry for data, which means it is crucial to ensure we have effective systems in place for data governance and access. Implementing adequate consent processes and privacy protection when gathering data is essential.

Organisations using and developing AI need to ensure they know what these technologies can and cannot do, and be aware of potential risks as well as benefits.

Jon Whittle, Director, *Data61*