KEMENTERIAN SAINS, TEKNOLOGI DAN INOVASI
MINISTRY OF SCIENCE,TECHNOLOGY AND INNOVATION

NUKLEAR MALAYSIA

TARIKH	ISNIN, 9 SEPTEMBER 2024
AKHBAR	THE STAR
TAJUK ARTIKEL	WHO'S TRAINING WHOM?
M/S	5
BIDANG	SCIENCE AND TECHNOLOGY
KATA KUNCI	SCIENCE AND TECHNOLOGY

Who's training whom?

We train AI, but AI might be training us, too, US researchers find.

By SERINA DESALVIO

MANY artificial intelligence (AI) systems are "trained" by interacting with human behaviour and human-created information.

Now, Washington University researchers have found that humans change their own behaviour when they know their actions are being used to train AI.

And not only do people change, but those changes can last into the future – creating new habits in the human trainers.

And tendencies or biases in a person's behaviour, including those the person isn't even aware of, can change, too.

So, who is training whom, here?

"This study clearly shows us that we need to understand these behaviours of people interacting with AI, specifically when they are helping train these tools, so that we can measure that bias and mitigate it," said Dr Philip R.O. Payne, director of the WashU Institute for Informatics and a professor of medicine.

But there could be a downside to people improving their behaviour if their actions are being used to train AI.

Lead researcher Lauren Treiman, a WashU graduate student, said that if people were helping train an AI for self-driving cars, for example, they might drive especially carefully. This might make the AI a perfect driver.

However, in a place like St Louis where people often run short yellow lights, it might be more dangerous for a self-driving car to try to be the perfect driver.

"AI might need to learn to run yellows, if that's what people tend to do," she said.

Algorithms all around

Treiman said she first started thinking about the effects of how AI is trained and the algorithms that determine what we see online, while scrolling through her social media feed.

"When you get 'recommended' videos, the algorithm can be super sensitive," she said.

"Sit on something for a few seconds, and you see content just like that over and over again."

She said she'll intentionally swipe past something quickly, or not click on it at all, so that social media algorithms learn to show her less of that kind of content.

The experience, changing her own behaviour in response to AI, inspired experimentation.

The researchers used the *Ultimatum Game* – a "classical academic approach", according to Wouter Kool, an assistant professor in psychological and brain sciences at Washington University.

The *Ultimatum Game* has two players: One decides how to split US$10 (RM43) between the two of them and makes an offer the other can accept or reject.

If the offer is accepted, each player gets the money they have agreed to split. But if it's rejected, neither player receives any money.

In the study, some players were told that the way they play this game would be used to teach AI to play. They were reminded of this with a small webcam icon at the corner of their screen and the words "Offer used to train AI".

Researchers found that people who were playing to train AI tended to reject more unfair offers.

Rejecting an offer lowers players' financial gain at the end of the game, yet those who were playing to train AI did it much more often than people who weren't.

It appeared that people wanted to train AI to play fairly and changed their behaviour to meet that desire.

However, participants had been told they would later play the *Ultimatum Game* against the AI – so were they training it fairly for the sake of fairness, or so they'd have better luck in the next round?

To answer this question, the researchers set up the same experiment, but this time told participants they would train an AI that someone else would face in the next round.

The results were the same: People trained AI to play fairly, even if it cost them some money – and even if they wouldn't play against that AI in the future.

Finally, researchers checked if the feeling of being observed prompted participants to play fairly.

"This is an age-old problem in research," Payne said. "People know they are being observed, and that influences what (researchers) are observing."

So the researchers deliberately removed the webcam icon from the screen of the AI-training group to help people feel a little less like they were being watched.

Still, "people were willing to sacrifice reward to make the AI more fair", said Chien-Ju Ho, a researcher and assistant professor of computer science and engineering at WashU.

'Patterns of persistence'

All these experiments make a robust case for humans changing their behaviour when they train AI, but the three researchers agreed the most interesting thing is how long that behaviour change lasted.

"Kool said the initial experiment only took around five short minutes. Participants came back two or three days later and played the same way they had in the first session, even after being explicitly told they were no longer training an AI.

"Looking at the behaviour in the second session, we saw these really beautiful, clear patterns of persistence of behaviour," Kool said.

Treiman said the findings have potential to shape how AI is trained – and raise other questions.

"In the *Ultimatum Game*, there's a clear definition of fairness. But, in a lot of other circumstances, there's not a clear definition," Treiman said.

"You always want to get the most fair, honest AI you can.

"But in other situations, what is fair? Especially since people will instil their preferences, biases or beliefs into the AI systems they are training." – St. Louis Post-Dispatch/Tribune News Service

Photo:123rf

DISEDIAKAN OLEH	**NOR SURIANI BINTI MOHD ZIN (S44), BPM** UNIT PERPUSTAKAAN, BPM