Could AI robots develop prejudice on their own?

September 6 2018

Credit: CC0 Public Domain

Showing prejudice towards others does not require a high level of cognitive ability and could easily be exhibited by artificially intelligent machines, new research has suggested.

Computer science and psychology experts from Cardiff University and MIT have shown that groups of autonomous machines could demonstrate prejudice by simply identifying, copying and learning this behaviour from one another.

It may seem that prejudice is a human-specific phenomenon that requires human cognition to form an opinion of, or to stereotype, a certain person or group.

Though some types of computer algorithms have already exhibited prejudice, such as racism and sexism, based on learning from public records and other data generated by humans, this new work demonstrates the possibility of AI evolving prejudicial groups on their own.

The new findings, which have been published in the journal *Scientific Reports*, are based on computer simulations of how similarly prejudiced individuals, or virtual agents, can form a group and interact with each other.

In a game of give and take, each individual makes a decision as to whether they donate to somebody inside of their own group or in a different group, based on an individual's reputation as well as their own donating strategy, which includes their levels of prejudice towards outsiders.

As the game unfolds and a supercomputer racks up thousands of simulations, each individual begins to learn new strategies by copying others either within their own group or the entire population.

Co-author of the study Professor Roger Whitaker, from Cardiff University's Crime and Security Research Institute and the School of Computer Science and Informatics, said: "By running these simulations thousands and thousands of times over, we begin to get an understanding

of how prejudice evolves and the conditions that promote or impede it.

"Our simulations show that prejudice is a powerful force of nature and through evolution, it can easily become incentivised in virtual populations, to the detriment of wider connectivity with others. Protection from prejudicial groups can inadvertently lead to individuals forming further prejudicial groups, resulting in a fractured population. Such widespread prejudice is hard to reverse."

The findings involve individuals updating their prejudice levels by preferentially copying those that gain a higher short term payoff, meaning that these decisions do not necessarily require advanced cognitive abilities.

"It is feasible that autonomous machines with the ability to identify with discrimination and copy others could in future be susceptible to prejudicial phenomena that we see in the human population," Professor Whitaker continued.

"Many of the AI developments that we are seeing involve autonomy and self-control, meaning that the behaviour of devices is also influenced by others around them. Vehicles and the Internet of Things are two recent examples. Our study gives a theoretical insight where simulated agents periodically call upon others for some kind of resource."

A further interesting finding from the study was that under particular conditions, which include more distinct subpopulations being present within a population, it was more difficult for prejudice to take hold.

"With a greater number of subpopulations, alliances of non-prejudicial groups can cooperate without being exploited. This also diminishes their status as a minority, reducing the susceptibility to prejudice taking hold. However, this also requires circumstances where agents have a higher

disposition towards interacting outside of their group," Professor Whitaker concluded.

More information: Roger M. Whitaker et al. Indirect Reciprocity and the Evolution of Prejudicial Groups, *Scientific Reports* (2018). DOI: 10.1038/s41598-018-31363-z

Provided by Cardiff University