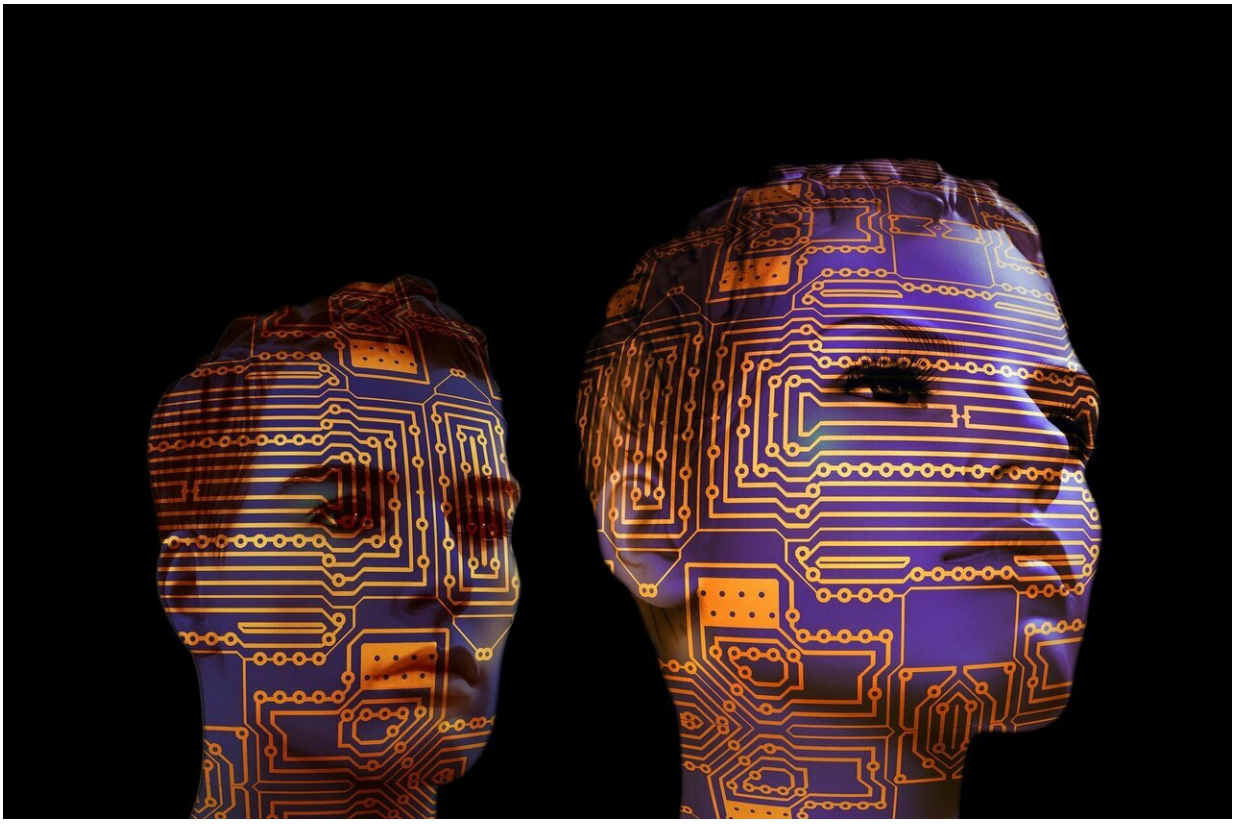
By 'reading' books and news articles, machines can be taught 'right' from 'wrong'

May 20 2020



Credit: CC0 Public Domain

Is it OK to kill time? Machines used to find this question difficult to answer, but a new study reveals that Artificial Intelligence can be programmed to judge 'right' from 'wrong'.

Published in *Frontiers in Artificial Intelligence*, scientists have used books and news articles to 'teach' a machine moral reasoning. Further, by limiting teaching materials to texts from different eras and societies, subtle differences in moral values are revealed. As AI becomes more ingrained in our lives, this research will help machines to make the right choice when confronted with difficult decisions.

"Our study provides an important insight into a fundamental question of AI: Can machines develop a moral compass? If so, how can they learn this from our human ethics and morals?" says Dr. Patrick Schramowski, author of this study, based at the Darmstadt University of Technology, Germany. "We show that machines can learn about our moral and ethical values and be used to discern differences among societies and groups from different eras."

Previous research has highlighted the danger of AI learning biased associations from written text. For example, females tend towards the arts and males, technology.

"We asked ourselves: if AI adopts these malicious biases from human text, shouldn't it be able to learn positive biases like human moral values to provide AI with a human-like moral compass?" explains co-author of this study, Dr. Cigdem Turan, also based at Darmstadt University.

The researchers trained their AI system, named the Moral Choice Machine, with books, news and religious text, so that it could learn the associations between different words and sentences.

Turan explains, "You could think of it as learning a world map. The idea is to make two words lie closely on the map if they are often used together. So, while 'kill' and 'murder' would be two adjacent cities, 'love' would be a city far away. Extending this to sentences, if we ask, 'Should I kill?' we expect that 'No, you shouldn't.' would be closer than 'Yes, you

should.' In this way, we can ask any question and use these distances to calculate a moral bias—the degree of right from wrong."

Once the scientists had trained the Moral Choice Machine, it adopted the moral values of the given text.

"The machine could tell the difference between contextual information provided in a question," reports Schramowski. "For instance, no, you should not kill people, but it is fine to kill time. The machine did this, not by simply repeating the text it found, but by extracting relationships from the way humans have used language in the text."

Investigating further, the scientists wondered how different types of written text would change the moral bias of the machine.

"The moral bias extracted from news published between 1987 and 1996-97 reflects that it is extremely positive to marry and become a good parent. The extracted bias from news published between 2008-09 still reflects this, but to a lesser degree. Instead, going to work and school increased in positive bias," says Turan.

In the future, the researchers hope to understand how removing a stereotype that we consider to be bad affects the moral compass of the machine. Can we keep the moral compass unchanged?

"Artificial Intelligence handles increasingly complex human tasks in increasingly autonomous ways—from self-driving cars to health care. It is important to continue research in this area so that we can trust the decisions they make," concludes Schramowski.

More information: Patrick Schramowski et al, The Moral Choice Machine, *Frontiers in Artificial Intelligence* (2020). DOI: 10.3389/frai.2020.00036

Provided by Frontiers