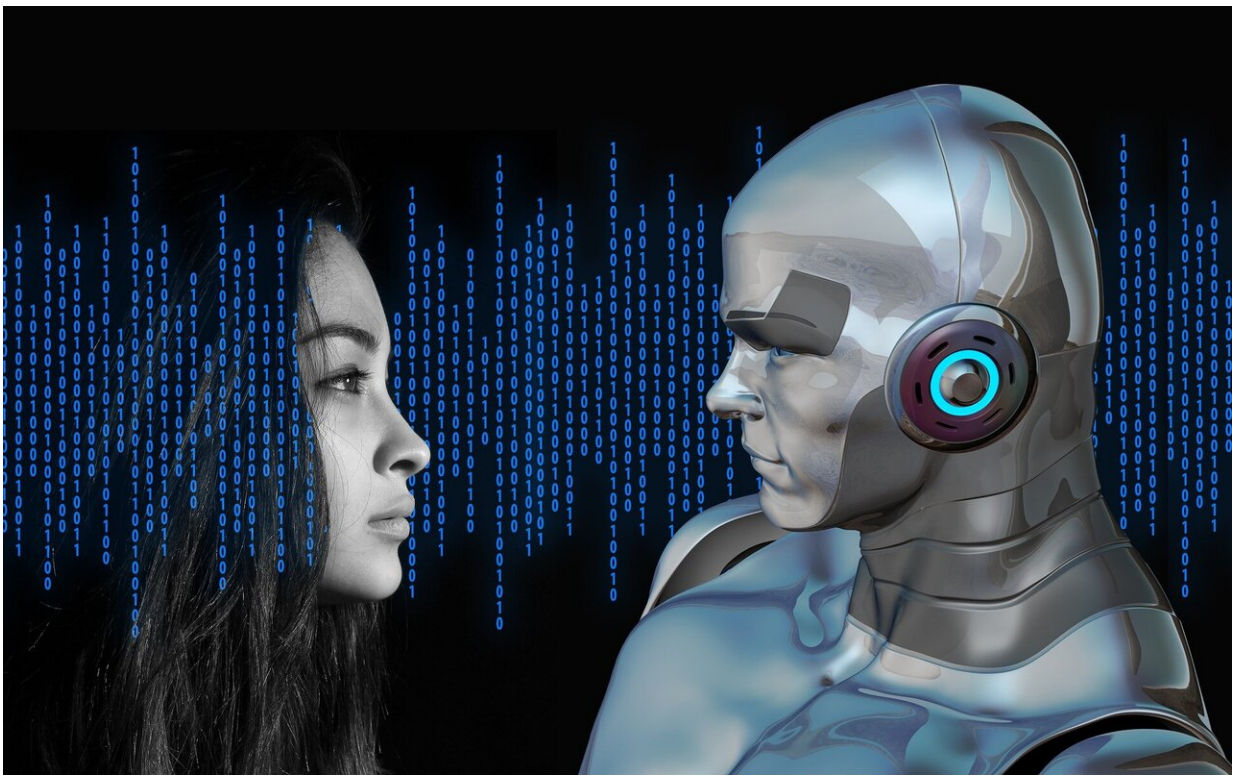
New mathematical idea reins in AI bias towards making unethical and costly commercial choices

June 30 2020



Credit: CC0 Public Domain

Researchers from the University of Warwick, Imperial College London, EPFL (Lausanne) and Sciteb Ltd have found a mathematical means of helping regulators and business manage and police Artificial Intelligence

systems' biases towards making unethical, and potentially very costly and damaging commercial choices—an ethical eye on AI.

Artificial intelligence (AI) is increasingly deployed in commercial situations. Consider for example using AI to set prices of insurance products to be sold to a particular customer. There are legitimate reasons for setting different prices for different people, but it may also be profitable to 'game' their psychology or willingness to shop around.

The AI has a vast number of potential strategies to choose from, but some are unethical and will incur not just moral cost but a significant potential economic penalty as stakeholders will apply some penalty if they find that such a strategy has been used—regulators may levy significant fines of billions of Dollars, Pounds or Euros and customers may boycott you—or both.

So in an environment in which decisions are increasingly made without human intervention, there is therefore a very strong incentive to know under what circumstances AI systems might adopt an unethical strategy and reduce that risk or eliminate entirely if possible.

Mathematicians and statisticians from University of Warwick, Imperial, EPFL and Sciteb Ltd have come together to help business and regulators creating a new "Unethical Optimization Principle" and provide a simple formula to estimate its impact. They have laid out the full details in a paper bearing the name "An unethical optimization principle", published in *Royal Society Open Science* on Wednesday 1st July 2020.

The four authors of the paper are Nicholas Beale of Sciteb Ltd; Heather Battey of the Department of Mathematics, Imperial College London; Anthony C. Davison of the Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne; and Professor Robert MacKay of the Mathematics Institute of the University of Warwick.

Professor Robert MacKay of the Mathematics Institute of the University of Warwick said:

"Our suggested 'Unethical Optimization Principle' can be used to help regulators, compliance staff and others to find problematic strategies that might be hidden in a large strategy space. Optimisation can be expected to choose disproportionately many unethical strategies, inspection of which should show where problems are likely to arise and thus suggest how the AI search algorithm should be modified to avoid them in future.

"The Principle also suggests that it may be necessary to re-think the way AI operates in very large strategy spaces, so that unethical outcomes are explicitly rejected in the optimization/learning process."

More information: An Unethical Optimization Principle, *Royal Society Open Science* (2020). URL after publication: royalsocietypublishing.org/doi/10.1098/rsos.200462

Provided by University of Warwick