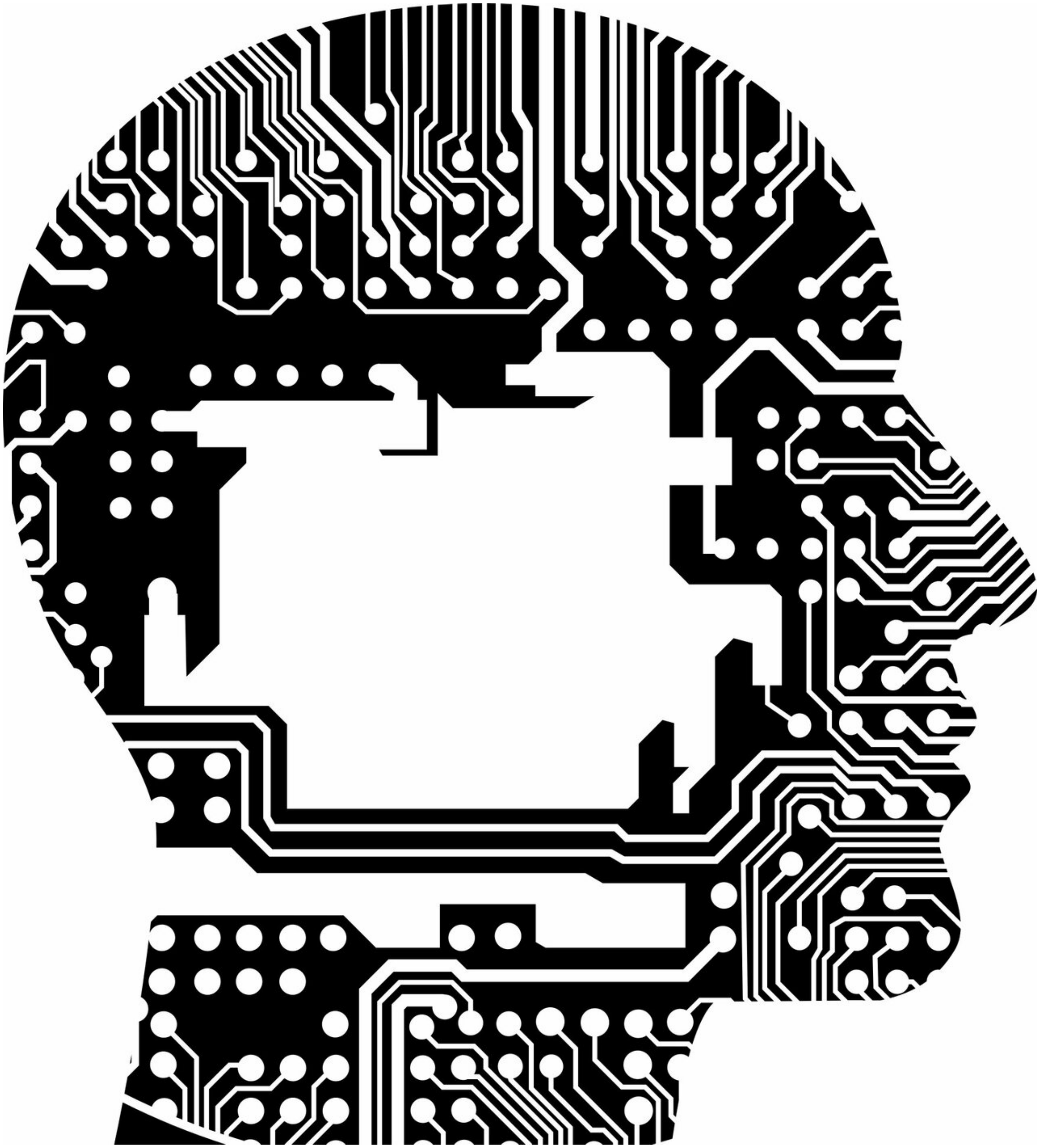
Defining blameworthiness to help make AI moral

March 28 2019, by Melanie Lefkowitz



Credit: CC0 Public Domain

Say 100 people live near a lake. If at least 10 of them overfish this year, the entire fish population will die out. Each assumes at least 10 others will overfish, and there won't be anything left to fish in the coming years.

Since the fish will be gone anyway, they all decide they might as well overfish. All the fish die. Do all of the people deserve blame?

It depends whether they could have coordinated with each other to change the outcome and the cost of doing so, according to new research by Joseph Halpern, the Joseph C. Ford Professor of Engineering, and Meir Friedenberg, a doctoral student in computer science. Building on Halpern's foundational work on causality, they developed a mathematical model to calculate blameworthiness on a scale from zero to one.

The research – lying at the intersection of computer science, philosophy and cognitive psychology – potentially could be used to guide the behavior of artificially intelligent agents, such as driverless vehicles, to help them behave in a "moral" way.

"One of the things we really wanted to do is give a framework that allows us to apply these kinds of legal and philosophical notions to autonomous systems," said Friedenberg, first author of "Blameworthiness in Multi-Agent Settings," which was presented at the 2019 AAAI Conference on Artificial Intelligence in February. "We think that's going to be important if we're going to effectively integrate autonomous systems into society."

In previous work, Halpern and colleagues defined individuals' blameworthiness roughly as the extent to which they believe their actions could change an event's outcome. For example, if you voted against a candidate who you believed would lose by a single vote, your blameworthiness would be one, the maximum; but if you believed the candidate would lose by thousands of votes, your blameworthiness would be far lower.

In the recent paper, Friedenberg and Halpern first gave a definition of a group's blameworthiness – essentially, a measure of the degree to which the group could have coordinated to bring about a different result. They then created a model to apportion the blameworthiness of the group to individual members.

"If you look at the group of fishermen, as a group they're responsible – obviously, if they didn't all fish there would be plenty for the next year," Halpern said. "The extent to which the fishermen are responsible is the extent to which they could coordinate to bring about a different outcome."

The researchers captured this by measuring the group members' ability to work together to change the outcome, taking into account the cost of doing so. Cost is a critical factor in blameworthiness: Someone who knocks over an expensive vase while running from a lion is less blameworthy than someone who is just not paying attention. If your vote swings an election, you're less blameworthy if someone is threatening to kill you unless you vote a certain way.

In future work, Halpern said he hopes to test the model by asking people, via crowdsourcing, to ascribe blameworthiness in various scenarios, and comparing their opinions with the numerical results.

When it comes to autonomous cars, developers or policymakers could consider their own definitions of cost when creating their algorithms, Halpern said. For example, if a government decides that no degree of risk is acceptable, a car would be designed to never pass another car, since that can increase the chance of an accident.

Although it can be difficult to determine how machine learning algorithms make decisions, it may be possible to develop more transparent algorithms that would allow for an easier assessment of blameworthiness.

"The advantage of our framework is that it gives you a formal way to think about these things and model them, and it forces you to be explicit about your assumptions and how you're defining the costs," Halpern said. "Our definition is trying to be quantitative, because like it or lump it you have to make tradeoffs, and this definition is forcing you to think about that. It's a tool to help people think about the tradeoffs without telling them what the tradeoffs should be."

More information: Meir Friedenberg and Joseph Y. Halpern. Blameworthiness in Multi-Agent Settings, arXiv:1903.04102 [cs.CY] arxiv.org/abs/1903.04102

Provided by Cornell University

Citation: Defining blameworthiness to help make AI moral (2019, March 28) retrieved 9 April 2024 from https://techxplore.com/news/2019-03-blameworthiness-ai-moral.html