Mathematical formula tackles complex moral decision-making in AI

October 10 2022, by Matt Shipman

Credit: Pixabay/CC0 Public Domain

An interdisciplinary team of researchers has developed a blueprint for creating algorithms that more effectively incorporate ethical guidelines into artificial intelligence (AI) decision-making programs. The project

was focused specifically on technologies in which humans interact with AI programs, such as virtual assistants or "carebots" used in healthcare settings.

"Technologies like carebots are supposed to help ensure the safety and comfort of hospital patients, older adults and other people who require health monitoring or physical assistance," says Veljko Dubljević, corresponding author of a paper on the work and an associate professor in the Science, Technology & Society program at North Carolina State University. "In practical terms, this means these technologies will be placed in situations where they need to make ethical judgments."

"For example, let's say that a carebot is in a setting where two people require medical assistance. One patient is unconscious but requires urgent care, while the second patient is in less urgent need but demands that the carebot treat him first. How does the carebot decide which patient is assisted first? Should the carebot even treat a patient who is unconscious and therefore unable to consent to receiving the treatment?"

"Previous efforts to incorporate ethical decision-making into AI programs have been limited in scope and focused on utilitarian reasoning, which neglects the complexity of human moral decision-making," Dubljević says. "Our work addresses this and, while I used carebots as an example, is applicable to a wide range of human-AI teaming technologies."

Utilitarian decision-making focuses on outcomes and consequences. But when humans make moral judgments they also consider two other factors.

The first factor is the intent of a given action and the character of the agent performing the action. In other words, who is performing a given action and what are they trying to accomplish? Is it benevolent or

malevolent? The second factor is the action itself. For example, people tend to view certain actions, such as lying, as inherently bad.

And all of these factors interact with each other. For example, we may agree that lying is bad, but if a nurse lies to a patient making obnoxious demands in order to prioritize treating a second patient in more urgent need, most people would view this as morally acceptable.

To address the complexity of moral decision-making, the researchers developed a mathematical formula and a related series of decision trees that can be incorporated into AI programs. These tools draw on something called the Agent, Deed, and Consequence (ADC) Model, which was developed by Dubljević and colleagues to reflect how people make complex ethical decisions in the real world.

"Our goal here was to translate the ADC Model into a format that makes it viable to incorporate into AI programming," Dubljević says. "We're not just saying that this ethical framework would work well for AI, we're presenting it in language that is accessible in a computer science context.

"With the rise of AI and robotics technologies, society needs such collaborative efforts between ethicists and engineers. Our future depends on it."

The paper is published open access in *AI and Ethics*.

More information: Michael Pflanzer et al, Ethics in human–AI teaming: principles and perspectives, *AI and Ethics* (2022). DOI: 10.1007/s43681-022-00214-z

Provided by North Carolina State University