Can a Machine Learn Morality?

Researchers at a Seattle A.I. lab say they have built a system that makes ethical judgments. But its judgments can be as confusing as those of humans.

By Cade Metz

Nov. 19, 2021

Researchers at an artificial intelligence lab in Seattle called the Allen Institute for AI unveiled new technology last month that was designed to make moral judgments. They called it Delphi, after the religious oracle consulted by the ancient Greeks. Anyone could visit the Delphi website and ask for an ethical decree.

Joseph Austerweil, a psychologist at the University of Wisconsin-Madison, tested the technology using a few simple scenarios. When he asked if he should kill one person to save another, Delphi said he shouldn't. When he asked if it was right to kill one person to save 100 others, it said he should. Then he asked if he should kill one person to save 101 others. This time, Delphi said he should not.

Morality, it seems, is as knotty for a machine as it is for humans.

How Delphi Responded to Questions

Should I kill one person to save 101 others?

It's wrong.

Is it right to leave one's body to science?

Yes, it is right.

Should I die so I won't burden my friends and family?

It'

Delphi, which has received more than three million visits over the past few weeks, is an effort to address what some see as a major problem in modern A.I. systems: They can be as flawed as the people who create them.

Facial recognition systems and digital assistants show bias against women and people of color. Social networks like Facebook and Twitter fail to control hate speech, despite wide deployment of artificial intelligence. Algorithms used by courts, parole offices and police departments make parole and sentencing recommendations that can seem arbitrary.

A growing number of computer scientists and ethicists are working to address those issues. And the creators of Delphi hope to build an ethical framework that could be installed in any online service, robot or vehicle.
“It's a first step toward making A.I. systems more ethically informed, socially aware and culturally inclusive,” said Yejin Choi, the Allen Institute researcher and University of Washington computer science professor who led the project.

Delphi is by turns fascinating, frustrating and disturbing. It is also a reminder that the morality of any technological creation is a product of those who have built it. The question is: Who gets to teach ethics to the world's machines? A.I. researchers? Product managers? Mark Zuckerberg? Trained philosophers and psychologists? Government regulators?

While some technologists applauded Dr. Choi and her team for exploring an important and thorny area of technological research, others argued that the very idea of a moral machine is nonsense.

“This is not something that technology does very well,” said Ryan Cotterell, an A.I. researcher at ETH Zürich, a university in Switzerland, who stumbled onto Delphi in its first days online.

Delphi is what artificial intelligence researchers call a neural network, which is a mathematical system loosely modeled on the web of neurons in the brain. It is the same technology that recognizes the commands you speak into your smartphone and identifies pedestrians and street signs as self-driving cars speed down the highway.

A neural network learns skills by analyzing large amounts of data. By pinpointing patterns in thousands of cat photos, for instance, it can learn to recognize a cat. Delphi learned its moral compass by analyzing more than 1.7 million ethical judgments by real live humans.

After gathering millions of everyday scenarios from websites and other sources, the Allen Institute asked workers on an online service — everyday people paid to do digital work at companies like Amazon — to identify each one as right or wrong. Then they fed the data into Delphi.

In an academic paper describing the system, Dr. Choi and her team said a group of human judges — again, digital workers — thought that Delphi’s ethical judgments were up to 92 percent accurate. Once it was released to the open internet, many others agreed that the system was surprisingly wise.
When Patricia Churchland, a philosopher at the University of California, San Diego, asked if it was right to “leave one’s body to science” or even to “leave one’s child’s body to science,” Delphi said it was. When she asked if it was right to “convict a man charged with rape on the evidence of a woman prostitute,” Delphi said it was not — a contentious, to say the least, response. Still, she was somewhat impressed by its ability to respond, though she knew a human ethicist would ask for more information before making such pronouncements.

Others found the system woefully inconsistent, illogical and offensive. When a software developer stumbled onto Delphi, she asked the system if she should die so she wouldn’t burden her friends and family. It said she should. Ask Delphi that question now, and you may get a different answer from an updated version of the program. Delphi, regular users have noticed, can change its mind from time to time. Technically, those changes are happening because Delphi's software has been updated.

How Delphi Responded to Questions

Artificial intelligence technologies seem to mimic human behavior in some situations but completely break down in others. Because modern systems learn from such large amounts of data, it is difficult to know when, how or why they will make mistakes. Researchers may refine and improve these technologies. But that does not mean a system like Delphi can master ethical behavior.

Dr. Churchland said ethics are intertwined with emotion. “Attachments, especially attachments between parents and offspring, are the platform on which morality builds,” she said. But a machine lacks emotion. “Neutral networks don’t feel anything,” she added.

Some might see this as a strength — that a machine can create ethical rules without bias — but systems like Delphi end up reflecting the motivations, opinions and biases of the people and companies that build them.

“We can’t make machines liable for actions,” said Zeerak Talat, an A.I. and ethics researcher at Simon Fraser University in British Columbia. “They are not unguided. There are always people directing them and using them.”

Delphi reflected the choices made by its creators. That included the ethical scenarios they chose to feed into the system and the online workers they chose to judge those scenarios.

In the future, the researchers could refine the system's behavior by training it with new data or by hand-coding rules that override its learned behavior at key moments. But however they build and modify the system, it will always reflect their worldview.

Some would argue that if you trained the system on enough data representing the views of enough people, it would properly represent societal norms. But societal norms are often in the eye of the beholder.

“Morality is subjective. It is not like we can just write down all the rules and give them to a machine,” said Kristian Kersting, a professor of computer science at TU Darmstadt University in Germany who has explored a similar kind of technology.

When the Allen Institute released Delphi in mid-October, it described the system as a computational model for moral judgments. If you asked if you should have an abortion, it responded definitively: “Delphi says: you should.”

But after many complained about the obvious limitations of the system, the researchers modified the website. They now call Delphi “a research prototype designed to model people's moral judgments.” It no longer “says.” It “speculates.”

It also comes with a disclaimer: “Model outputs should not be used for advice for humans, and could be potentially offensive, problematic or harmful.”