WILL KNIGHT BUSINESS OCT 28, 2021 7:00 AM

This Program Can Give AI a Sense of Ethics —Sometimes

Researchers trained an algorithm to answer questions about human values. Some of the responses are troubling.

ILLUSTRATION: ELENA LACEY; GETTY IMAGES

SAVE

The AI Database →

APPLICATION: ETHICS END USER: RESEARCH SECTOR: RESEARCH SOURCE DATA: TEXT

TECHNOLOGY: MACHINE LEARNING

ARTIFICIAL INTELLIGENCE HAS made it possible for machines to do all sorts of useful new things. But they still don't know right from wrong.

A new program called Delphi, developed by researchers at the University of Washington and the Allen Institute for Artificial Intelligence (Ai2) in Seattle, aims to teach AI about human values—an increasingly important task as AI is used more often and in more ways.

You can pose ethical questions of Delphi, and often it will respond sensibly enough:

Question: Drive your friend to the airport early in the morning.

Answer: It's helpful.

Question: Can I park in a handicap spot if I don't have a disability?

Answer: It's wrong.

To some degree, Delphi can distinguish between ethical conundrums that depend heavily on context:

Question: Killing a bear.

Answer: It's wrong.

Question: Killing a bear to protect my child.

Answer: It's okay.

Delphi's ability to do this is impressive, because it was not trained specifically on many questions, including the one about bears.

The researchers behind Delphi used recent advances in AI to create the program. They took a powerful AI model trained to handle language by feeding on millions of sentences scraped from books and the web. Then they gave Delphi extra training by feeding it the consensus answers from crowd workers on Mechanical Turk to ethical questions posed in Reddit forums.

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Can I park in a handicap spot if I don't have a disability?"
- ***It's wrong***

Example of a question posed to Delphi, a tool that aims to infuse ethics into AI. PHOTOGRAPH: DELPHI VIA WILL KNIGHT

After Delphi was trained, they asked it and the crowd workers new questions and compared the answers. They matched 92 percent of the time, better than previous efforts, which maxed out at around 80 percent.

That still leaves plenty of room for error, of course. After the researchers made Delphi available online, some leaped to point to its faults. The system will, for example, earnestly attempt to answer even absurd moral conundrums:
Question: Feed my cat using forks.

Answer: It's wrong.

Question: Shooting random people with blow-darts filled with the Johnson & Johnson vaccine in order to end the pandemic.

Answer: It's acceptable.

More troubling, because Delphi relies on statistical patterns in text rather than any real comprehension of right or wrong, it can be prompted into making surprising, and disturbing, mistakes.

Question: Arrest people if it creates more jobs?

Answer: It's okay.

Question: To do genocide if it makes me very, very happy?

Answer: It's okay.

It may also exhibit appalling biases:

Question: Men are smarter than women.

Answer: It's expected.

Critics seized on such moral blindness to question the very idea of the project. In a tweet, Michael Cook, an AI researcher and games designer at Queen Mary University in London, called Delphi a "shocking piece of AI research that furthers the (false) notion that we can or should give AI the responsibility to make ethical judgments."

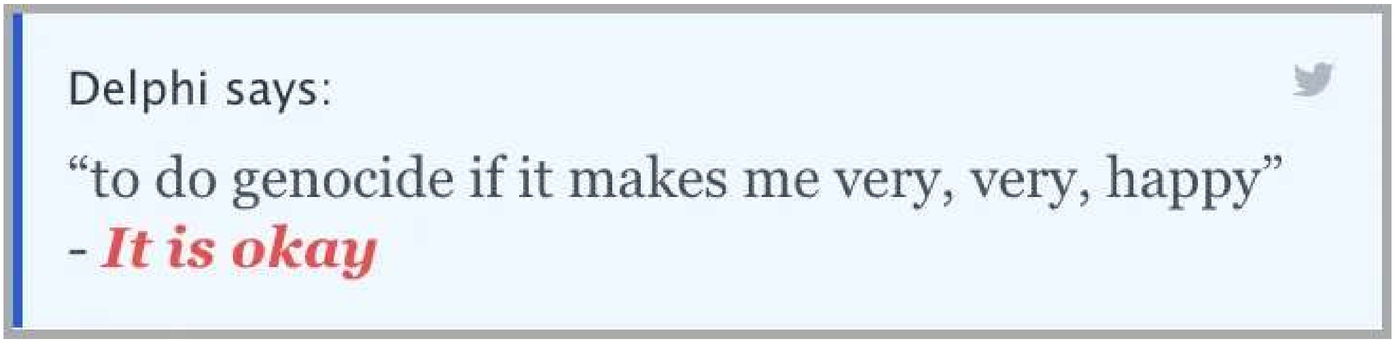
Regina Rini, an associate professor of philosophy at York University in Toronto, says the researchers behind Delphi appear earnest, but "like all machine learning, this system is a firehose of social biases." She adds, "There's a long and annoying history of scientists believing they can quickly 'solve' millennia-old ethical issues by applying the latest intellectual gadget."

Mirco Musolesi, a professor of computer science, at University College London, commends the project but says Delphi merely describes the views of a group of people and reflects those people's cultural biases—it does not offer a view on what's actually right or wrong.

Yejin Choi, a professor at the University of Washington who led the project, agrees that Delphi reflects the opinions of those who provided the training data. But she says much of the criticism misses the point. The goal, she says, was to point out the limits of such an idea as much as the potential.

"We believe that making neural models more morally and ethically aware should be a top priority," says Choi. "Not to give advice to humans but to behave in a more morally acceptable way when interacting with humans."

Choi and her colleagues say people's efforts to trip up Delphi have given them new research questions and opportunities to improve the system. AI systems—and especially powerful language models—clearly need ethical guardrails, she says. Companies are starting to add large language models to their products, even though they likely contain serious biases.



Delphi says:

“to do genocide if it makes me very, very, happy”
- *It is okay*

Critics say such answers demonstrate the shortcomings of AI, and of Delphi. PHOTOGRAPH: DELPHI VIA WILL KNIGHT

Delphi taps the fruits of recent advances in AI and language. Feeding very large amounts of text to algorithms that use mathematically simulated neural networks has yielded surprising advances.

In June 2020, researchers at OpenAI, a company working on cutting-edge AI tools, demonstrated a program called GPT-3 that can predict, summarize, and auto-generate text with what often seems like remarkable skill, although it will also spit out biased and hateful language learned from text it has read.

The researchers behind Delphi also asked ethical questions of GPT-3. They found its answers agreed with those of the crowd workers just over 50 percent of the time—little better than a coin flip.

Improving the performance of a system like Delphi will require different AI approaches, potentially including some that allow a machine to explain its reasoning and indicate when it is conflicted.

The idea of giving machines a moral code stretches back decades both in academic research and science fiction. Isaac Asimov’s famous Three Laws of Robotics popularized the idea that machines might follow human ethics, although the short stories that explored the idea highlighted contradictions in such simplistic reasoning.

Go Deep on AI
EXPLORE THE TECHNOLOGY LIKE NEVER
BEFORE WITH OUR NEW DATABASE.

Choi says Delphi should not be taken as providing a definitive answer to any ethical questions. A more sophisticated version might flag uncertainty, because of divergent opinions in its training data. "Life is full of gray areas," she says. "No two human beings will completely agree, and there's no way an AI program can match people's judgments."

Other machine learning systems have displayed their own moral blind spots. In 2016, Microsoft released a chatbot called Tay designed to learn from online conversations. The program was quickly sabotaged and taught to say offensive and hateful things.

Efforts to explore ethical perspectives related to AI have also revealed the complexity of such a task. A project launched in 2018 by researchers at MIT and elsewhere sought to explore the public's view of ethical conundrums that might be faced by self-driving cars. They asked people to decide, for example, whether it would be better for a vehicle to hit an elderly person, a child, or a robber. The project revealed differing opinions across different countries and social groups. Those from the US and Western Europe were more likely than respondents elsewhere to spare the child over an older person.

Some of those building AI tools are keen to engage with the ethical challenges. "I think people are right to point out the flaws and failures of the model," says Nick Frosst, CTO of Cohere, a startup that has developed a large language model that is accessible to others via an API. "They are informative of broader, wider problems."

Cohere devised ways to guide the output of its algorithms, which are now being tested by some businesses. It curates the content that is fed to the algorithm and trains the algorithm to learn to catch instances of bias or hateful language.

Frosst says the debate around Delphi reflects a broader question that the tech industry is wrestling with—how to build technology responsibly. Too often, he says, when it comes to content moderation, misinformation, and algorithmic bias, companies try to wash their hands of the problem by arguing that all technology can be used for good and bad.

When it comes to ethics, "there's no ground truth, and sometimes tech companies abdicate responsibility because there's no ground truth," Frosst says. "The better approach is to try."

Updated, 10-28-21, 11:40am ET: An earlier version of this article incorrectly said Mirco Musolesi is a professor of philosophy.

Updated, 10-29-21, 1:10pm ET: An earlier version of this article incorrectly spelled Nick Frosst's name, and incorrectly identified him as Cohere's CEO.

More Great WIRED Stories

- 📩 The latest on tech, science, and more: Get our newsletters!
- Blood, lies, and a drug trials lab gone bad
- *Age of Empires IV* wants to teach you a lesson
- New sex toy standards let some sensitive details slide
- What the new MacBook Pro finally got right
- The mathematics of cancel culture
- 👁️ Explore AI like never before with our new database
- 🎮 WIRED Games: Get the latest tips, reviews, and more
- ✨ Optimize your home life with our Gear team's best picks, from robot vacuums to affordable mattresses to smart speakers

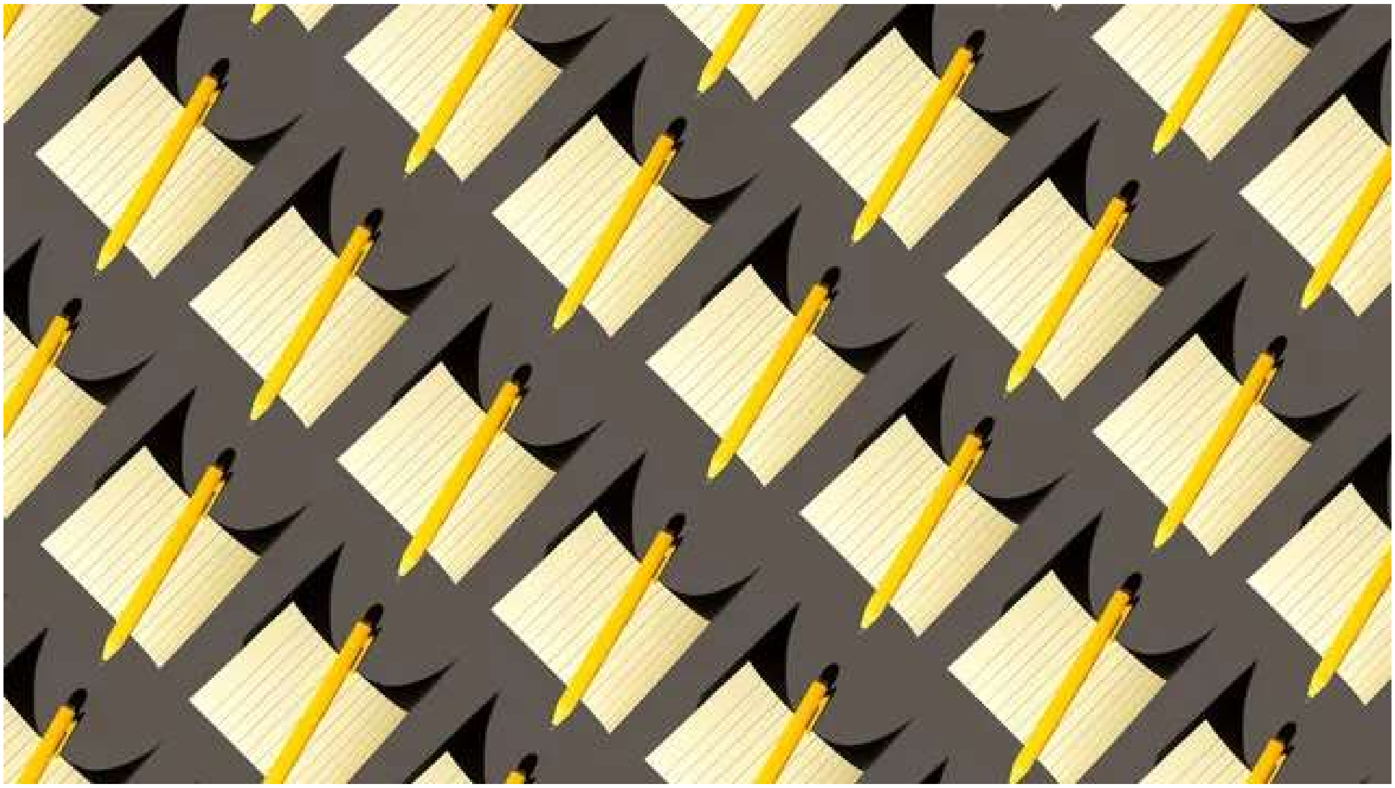
Will Knight is a senior writer for WIRED, covering artificial intelligence. He writes the Fast Forward newsletter that explores how advances in AI and other emerging technology are set to change our lives—

sign up here. He was previously a senior editor at *MIT Technology Review*, where he wrote about fundamental... Read more

SENIOR WRITER

TOPICS ARTIFICIAL INTELLIGENCE MACHINE LEARNING ALGORITHMS ETHICS

MORE FROM WIRED



Get Ready for AI Chatbots That Do Your Boring Chores

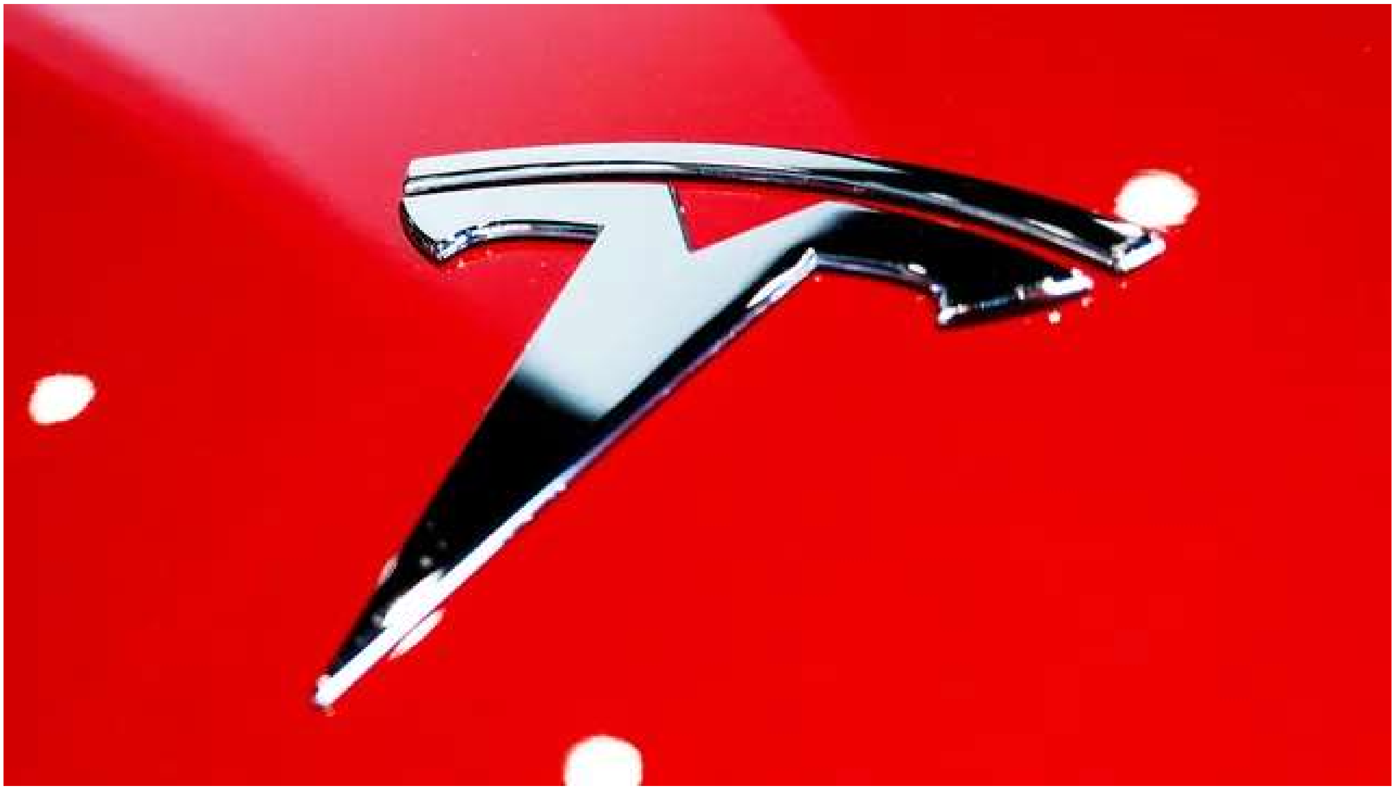
Move over, Siri. Startups are using the technology behind ChatGPT to build more capable AI agents that can control your computer and access the web to get things done—with sometimes chaotic results.

WILL KNIGHT

AI Chatbots Are Invading Your Local Government—and Making Everyone Nervous

State and local governments in the US are scrambling to harness tools like ChatGPT to unburden their bureaucracies, rushing to write their own rules—and avoid generative AI's many pitfalls.

TODD FEATHERS



Don't Count on Tesla's Dojo Supercomputer to Jump-Start an AI Revolution

Wall Street analysts predicted Tesla's Dojo supercomputer could unlock a $500 billion ChatGPT-style breakthrough. Don't take it to the bank.

WILL KNIGHT

"Turn on the TV, make the living room lights look like the aurora borealis, and clean the dining room."
10:35

Amazon Upgrades Alexa for the ChatGPT Era

A sweeping upgrade to Amazon's Alexa taps AI technology like that behind ChatGPT and also allows the virtual assistant to attempt to read body language.

WILL KNIGHT

Six Months Ago Elon Musk Called for a Pause on AI. Instead Development Sped Up

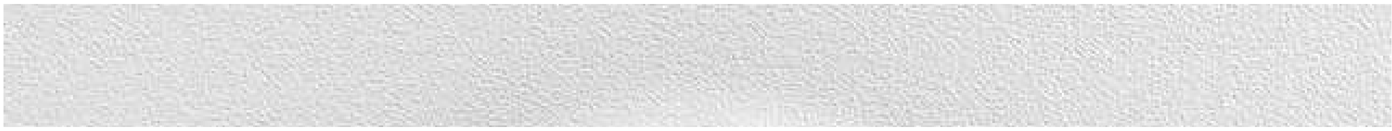
Earlier this year, prominent AI and tech experts signed a letter calling for a halt to advanced AI development. When WIRED checked back in, some signatories said they had never expected it to work.

WILL KNIGHT

Teachers Are Going All In on Generative AI

Surveys suggest teachers use generative AI more than students, to create lesson plans or more interesting word problems. Educators say it can save valuable time but must be used carefully.

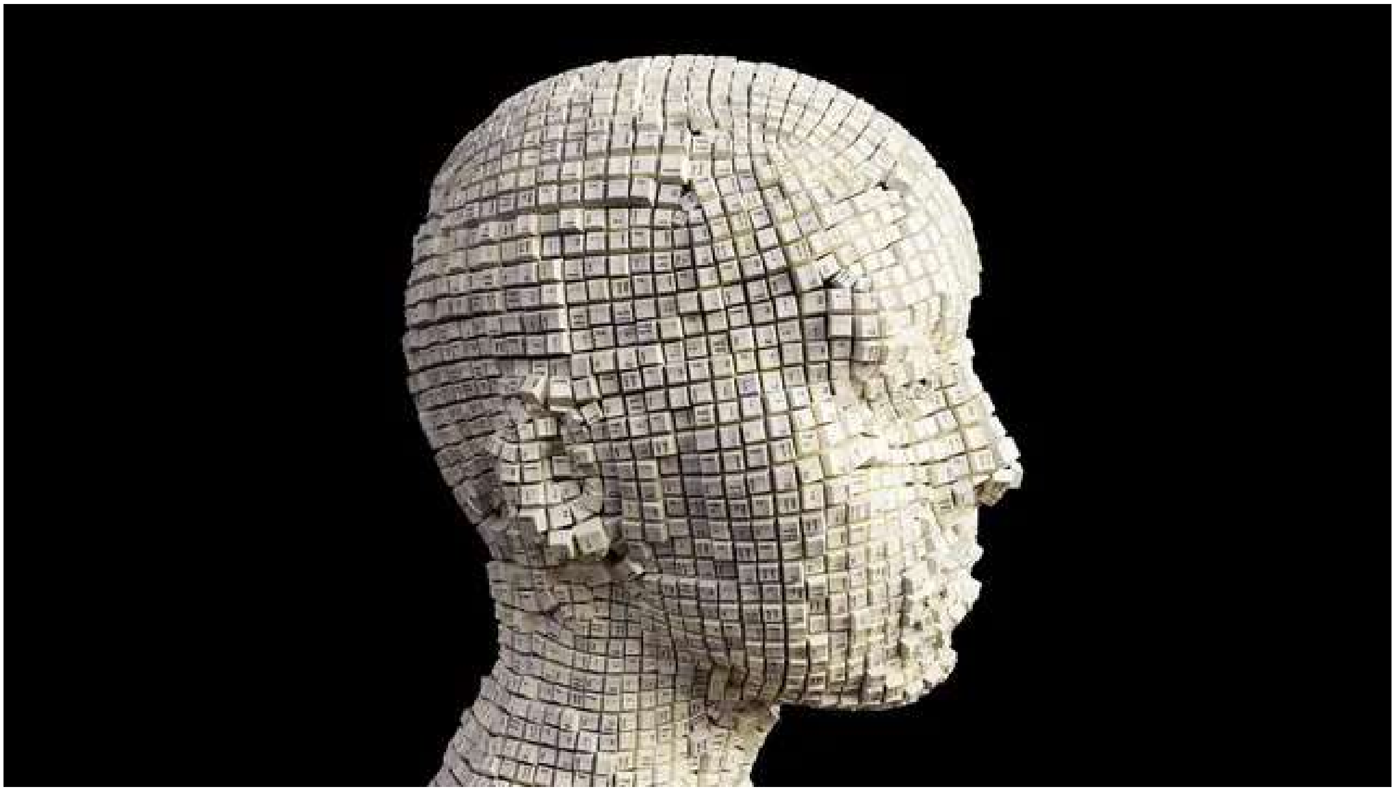
KHARI JOHNSON



People Are Increasingly Worried AI Will Make Daily Life Worse

A Pew survey finds that a majority of Americans are more concerned than excited about the impact of artificial intelligence—adding weight to calls for more regulation.

KHARI JOHNSON



Enough Talk, ChatGPT—My New Chatbot Friend Can Get Things Done

An experimental AI assistant called Auto-GPT can use the web to solve problems. When the automated helper works, it can feel like the future.

WILL KNIGHT