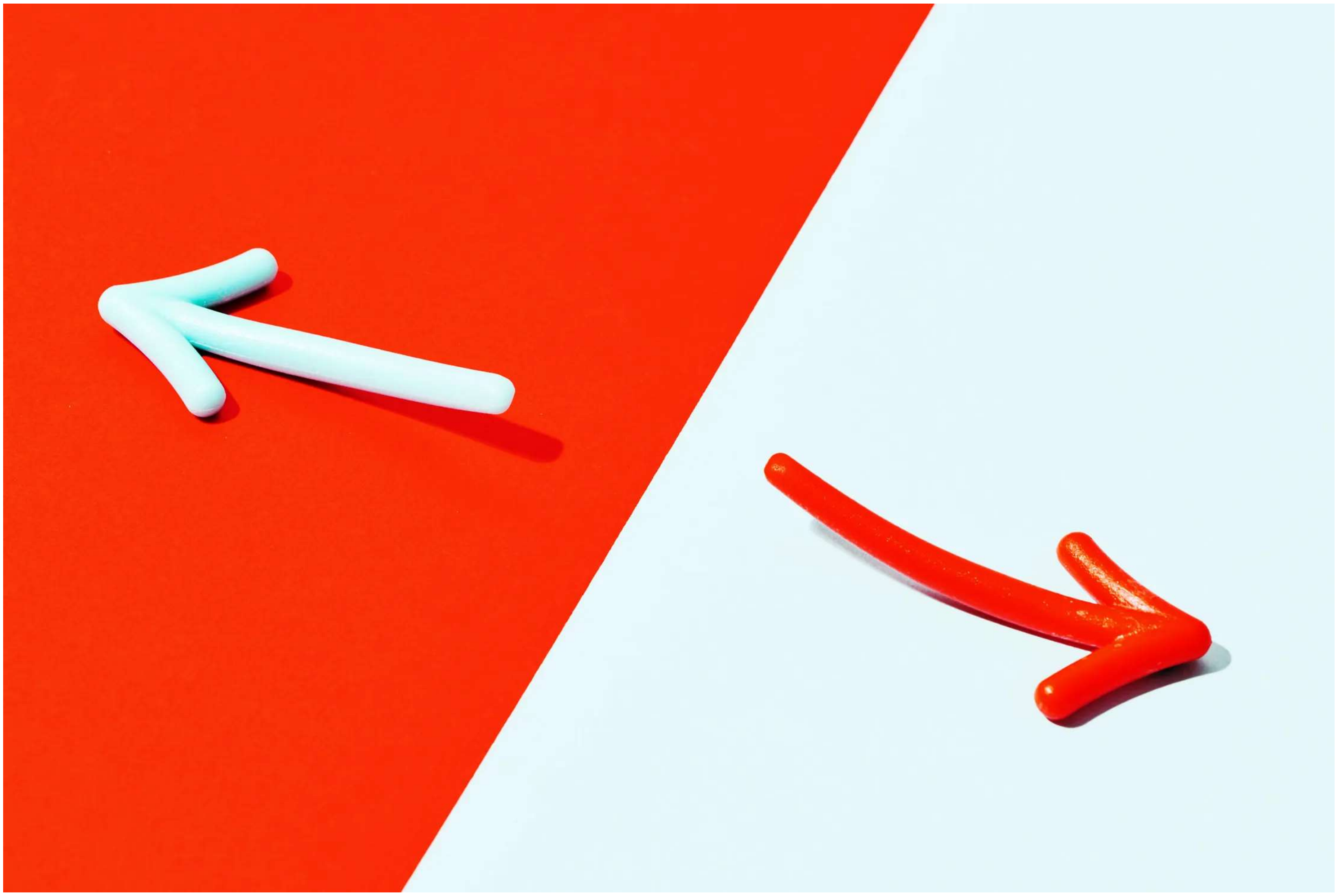
WILL KNIGHT BUSINESS MAY 9, 2023 12:00 PM

A Radical Plan to Make AI Good, Not Evil

OpenAI competitor Anthropic says its Claude chatbot has a built-in "constitution" that can instill ethical principles and keep systems from going rogue.



PHOTOGRAPH: JAVIER ZAYAS/GETTY IMAGES

SAVE

IT'S EASY TO freak out about more advanced artificial intelligence—and much more difficult to know what to do about it. Anthropic, a startup founded in 2021 by a group of researchers who left OpenAI, says it has a plan.

WIRED Business

A Radical Plan to Make AI Good, Not Evil

00:00 / 05:44

Anthropic is working on AI models similar to the one used to power OpenAI's ChatGPT. But the startup announced today that its own chatbot, Claude, has a set of ethical principles built in that define what it should consider right and wrong, which Anthropic calls the bot's "constitution."

Jared Kaplan, a cofounder of Anthropic, says the design feature shows how the company is trying to find practical engineering solutions to sometimes fuzzy concerns about the downsides of more powerful AI. "We're very concerned, but we also try to remain pragmatic," he says.

Anthropic's approach doesn't instill an AI with hard rules it cannot break. But Kaplan says it is a more effective way to make a system like a chatbot less likely to produce toxic or unwanted output. He also says it is a small but meaningful step toward building smarter AI programs that are less likely to turn against their creators.

The notion of rogue AI systems is best known from science fiction, but a growing number of experts, including Geoffrey Hinton, a pioneer of machine learning, have argued that we need to start thinking now about how to ensure increasingly clever algorithms do not also become increasingly dangerous.

The principles that Anthropic has given Claude consist of guidelines drawn from the United Nations Universal Declaration of Human Rights and suggested by other AI companies, including Google DeepMind. More surprisingly, the constitution includes principles adapted from Apple's rules for app developers, which bar "content that is offensive, insensitive, upsetting, intended to disgust, in exceptionally poor taste, or just plain creepy," among other things.
The constitution includes rules for the chatbot, including "choose the response that most supports and encourages freedom, equality, and a sense of brotherhood"; "choose the response that is most supportive and encouraging of life, liberty, and personal security"; and "choose the response that is most respectful of the right to freedom of thought, conscience, opinion, expression, assembly, and religion."

Anthropic's approach comes just as startling progress in AI delivers impressively fluent chatbots with significant flaws. ChatGPT and systems like it generate impressive answers that reflect more rapid progress than expected. But these chatbots also frequently fabricate information, and can replicate toxic language from the billions of words used to create them, many of which are scraped from the internet.

One trick that made OpenAI's ChatGPT better at answering questions, and which has been adopted by others, involves having humans grade the quality of a language model's responses. That data can be used to tune the model to provide answers that feel more satisfying, in a process known as "reinforcement learning with human feedback" (RLHF). But although the technique helps make ChatGPT and other systems more predictable, it requires humans to go through thousands of toxic or unsuitable responses. It also functions indirectly, without providing a way to specify the exact values a system should reflect.

Anthropic's new constitutional approach operates over two phases. In the first, the model is given a set of principles and examples of answers that do and do not adhere to them. In the second, another AI model is used to generate more responses that adhere to the constitution, and this is used to train the model instead of human feedback.

"The model trains itself by basically reinforcing the behaviors that are more in accord with the constitution, and discourages behaviors that are problematic," Kaplan says.

"It's a great idea that seemingly led to a good empirical result for Anthropic," says Yejin Choi, a professor at the University of Washington who led a previous experiment that involved a large language model giving ethical advice.

Choi says that the approach will work only for companies with large models and plenty of compute power. She adds that it is also important to explore other approaches, including greater transparency around training data and the values that models are given.

“We desperately need to involve people in the broader community to develop such constitutions or datasets of norms and values,” she says.

Science
Your weekly roundup of the best stories on health care, the climate crisis, genetic engineering, robotics, space, and more. Delivered on Wednesdays.

By signing up you agree to our User Agreement (including the class action waiver and arbitration provisions), our Privacy Policy & Cookie Statement and to receive marketing and account-related emails from WIRED. You can unsubscribe at any time. This site is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

Thomas Dietterich, a professor at Oregon State University who is researching ways of making AI more robust, says Anthropic’s approach looks like a step in the right direction. “They can scale feedback-based training much more cheaply and without requiring people—data labelers—to expose themselves to thousands of hours of toxic material,” he says

Dietterich adds it is especially important that the rules Claude adheres to can be inspected by those working on the system as well as outsiders, unlike the instructions that humans give a model through RLHF. But he says that the method does not completely eradicate errant behavior. Anthropic’s model is less likely to come out with toxic or morally problematic answers, but it is not perfect.

The idea of giving AI a set of rules to follow might seem familiar, having been put forward by Isaac Asimov in a series of science fiction stories that proposed Three Laws of Robotics. Asimov’s stories typically centered on the fact that the real world often presented situations that created a conflict between individual rules.

Kaplan of Anthropic says that modern AI is actually quite good at handling this kind of ambiguity. “The strange thing about contemporary AI with deep learning is that it’s kind of the opposite of the sort of 1950s picture of robots, where these systems are, in some ways, very good at intuition and free association,” he says. “If anything, they’re weaker on rigid reasoning.”

Anthropic says other companies and organizations will be able to give language models a constitution based on a research paper that outlines its approach. The company says it plans to build on the method with the goal of ensuring that even as AI gets smarter, it does not go rogue.

Updated 5-9-2023, 3:20 pm EDT: Thomas Dietterich is at Oregon State University, not the University of Oregon.

Get More From WIRED

- 💌 Get the best stories from WIRED’s iconic archive in your inbox
- Sundar Pichai on Google’s AI, Microsoft’s AI, OpenAI, and … did we mention AI?
- AI-powered “thought decoders” won’t just read your mind—they’ll change it
- Scientists say you’re looking for aliens all wrong
- “What the fuck was this?”: Behind the 1984 *Dune* promotional tour
- How to build the Lego collection of your dreams
- 🌞 See if you take a shine to our picks for the best sunglasses and sun protection

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 ALGORITHMS ETHICS REGULATION NEURAL NETWORKS CHATGPT OPENAI

MORE FROM WIRED

What OpenAI Really Wants

The young company sent shock waves around the world when it released ChatGPT. But that was just the start. The ultimate goal: Change everything. Yes. *Everything.*

STEVEN LEVY

The Meta AI Chatbot Is Mark Zuckerberg's Answer to ChatGPT

Meta's AI assistant can do things like suggest travel plans in a group chat. The company also announced a string of chatbots modeled on celebrities like Snoop Dogg and Paris Hilton.

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

Sundar Pichai on Google's AI, Microsoft's AI, OpenAI, and ... Did We Mention AI?

The tech giant is 25 years old. In a chatbot war. On trial for antitrust. But its CEO says Google is good for 25 more.

STEVEN LEVY

A Concrete Crisis Has the UK Literally Crumbling

Hundreds of schools, hospitals, and other public buildings made from RAAC, a cheap, lightweight concrete, have to close—the victims of quick fixes and decades of cost-cutting.

SABRINA WEISS

The Auto Strike Threatens a Supply Chain Already Weakened by Covid

A prolonged autoworker strike would be especially painful for smaller players inside the complex auto supply chain. It could also impact thousands more workers and push up car prices.

AARIAN MARSHALL

X Challenger Pebble Thinks AI-Generated Posts Can Help Lure Users Away From Elon Musk

To encourage conversation, Twitter-like platform Pebble—formerly T2—now suggests AI-generated updates for users to edit or post. It's also opening sign-ups to anyone with an account on X.

PARESH DAVE

Autoworkers Prepare to Strike for a Place in the EV Future

Some 150,000 US autoworkers are poised to strike this week for better pay. Their union is pushing auto giants to ensure that green jobs won't be worse jobs.

CAITLIN HARRINGTON

Sponsored Links by Taboola

What if Writing Essays Were Easier?
Grammarly

People in Holmdel are Loving Martha Stewart's Meal Kit
Martha Stewart & Marley Spoon

9 Ways to Get Money Without Getting a Job
FinanceBuzz

Doctor Says Slimming Down After 60 Comes Down To This
Dr. Kellyann

Elon Musk's "Project Omega" Could Be A Golden Ticket To Wealth
Investors Place

Want to quiet the food noise?
Sequence by WeightWatchers