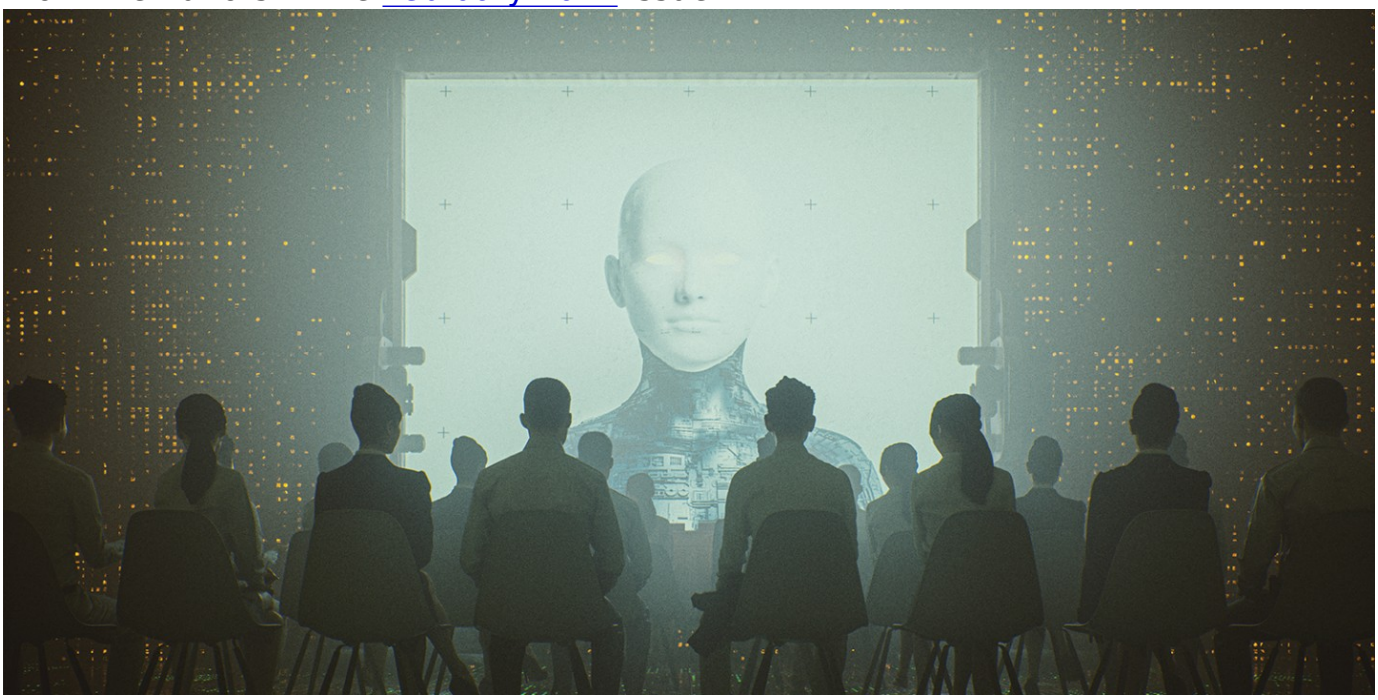
Artificial intelligence needs theology

What exactly is an “evil robot”? Who gets to define it?

From the Editors in the February 2024 issue



(Image by gremlin / E+ / Getty)

“OpenAI Releases Plan to Prevent a Robot Apocalypse” read the headline of a December 19 *Daily Beast* article. The plan is to create a “preparedness team” to supplement the artificial intelligence company’s current safety efforts, which include mitigating the technology’s adoption of human biases such as racism and preventing machine goals from overriding human goals. The new team, led by MIT researcher Aleksander Madry, will monitor for potential catastrophic events involving AI, such as creating and deploying biological weapons, sabotaging an economic sector, or hacking into military systems to start a nuclear war.

The primary work of the preparedness team, Madry told the *Washington Post*, will be to prevent exploitation of the technology by people who approach it asking, “‘How can I mess with this set of rules? How can I be most ingenious in my evilness?’”

In this issue of *The Century*, theologian Katherine Schmidt writes about the role of the humanities in a world where AI is becoming increasingly powerful, generative, and autonomous. As corporations and policymakers grapple with the blurring lines between human agency and computer-generated agency, Schmidt argues that "ethical theory and ethics education" are vital. Further, she points out that theologians and philosophers are "uniquely qualified" to weigh in on the conversation, since they are skilled at addressing basic questions about truth, meaning, and agency.

Madry's characterization of a bad actor in the AI world as someone who asks, "How can I be most ingenious in my evilness?" supports Schmidt's argument. It's one thing to have a team of technology experts devoted to squelching the plans of folks who deliberately deploy their evilness in ingenious ways. It's another thing for the developers and disseminators of AI technology to consider the questions behind Madry's hypothetical question. What exactly is "evilness"? Who gets to define it? Where does it come from? How intrinsic is it to human nature? Is it even possible to recognize it in ourselves? Can it spread from one person to another or multiply within a group of people? How can it be measured, and who does that measuring? To what extent might it be foiled by human effort?

These are theological questions, and it would be unreasonable to expect AI experts to be able to answer them. But it's not unreasonable to believe that those experts will be better equipped to prevent AI catastrophes if they've spent some time considering theological questions—whether about evil, creation, incarnation, agency, eschatology, or some other theological topic that intersects with the possibilities and problems posed by AI. Pondering a theological question that has no simple answer can expand the human mind and heart, creating fissures in the empirical limits we typically impose on knowledge and action.

It's not that theologically minded people know *more* than others; it's that they know *differently*. And if we're headed for a robot apocalypse, people who know differently should be among the first to see it coming and respond.