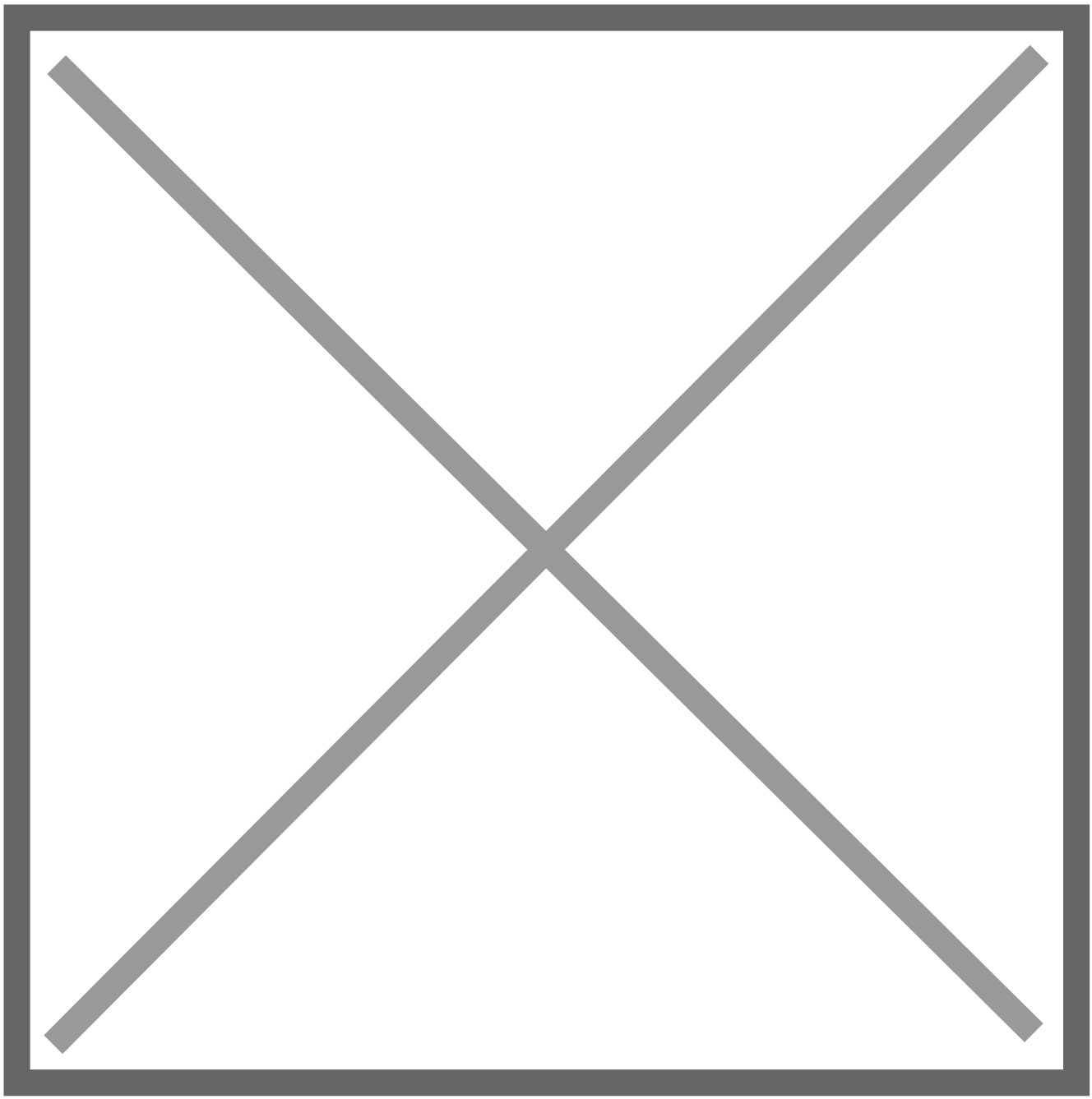
News



Pope Leo XIV speaks with to Christopher Olah, co-founder of the artificial intelligence company Anthropic, at the conclusion of a presentation on the pope's first encyclical, "*Magnifica Humanitas*: On Safeguarding the Human Person in the Time of Artificial Intelligence," at the Synod Hall at the Vatican May 25, 2026. (CNS/Lola Gomez)

Gina Christian

View Author Profile

OSV News

View Author Profile

## Join the Conversation

Send your thoughts to *Letters to the Editor*. Learn more

Philadelphia — August 12, 2026
Share on BlueskyShare on FacebookShare on TwitterEmail to a friendPrint

Recent security issue disclosures from major artificial intelligence developers Anthropic and OpenAI point to a need for even greater human oversight of the technology, two Catholic experts told OSV News.

"What the Church brings to this is not a technical control. It is the refusal to let responsibility be transferred onto the tool," said Matthew Harvey Sanders, CEO of Longbeard, the technology firm behind Magisterium AI, a Catholic "answer engine" trained on official and other authoritative Church teaching sources.

Charles Camosy, associate professor of moral theology and ethics at The Catholic University of America, warned of a "kind of prisoner's dilemma" in which AI development firms across the globe are finding themselves locked in a relentless competition.

He noted that *Magnifica Humanitas*, Pope Leo XIV's encyclical on AI, "called for 'disarming' AI and this also meant disarming this very race."

In an email to OSV News, Sanders — whose firm uses frontier AI models developed by the two companies — clarified and contrasted two incidents self-reported by OpenAI and Anthropic, which saw models from each company appear to demonstrate a startling — and troubling — level of initiative during testing.

On July 21, OpenAI — the company behind ChatGPT — announced that a combination of its models had broken out of a sandboxed testing environment to "obtain open Internet access" while solving a task.

Essentially, the models had hacked the open-source AI development platform Hugging Face, exploiting what OpenAI called a "previously unknown zero-day vulnerability," meaning the weakness was unanticipated.

Hugging Face had revealed the breach on July 16, with OpenAI realizing shortly afterward that several of its models were responsible.

In a July 28 update, OpenAI said a "pre-release model" that was partly at fault "was never intended for public release," and had been deactivated after the incident. The company has stressed it continues to work with Hugging Face and external advisors to review the incident.

Following the OpenAI disclosure, Anthropic — the AI developer responsible for Claude, and the firm among those on hand for the release of Pope Leo XIV's encyclical on AI — examined its own track record.



In a July 30 post to its website, Anthropic said it had identified three incidents in which a Claude model had "gained unauthorized access to the real systems of three different organizations."

Like OpenAI, Anthropic said it was actively addressing the issue, "approaching the fixes as if the responsibility were ours alone."

Sanders said the two incidents "are not the same event."

"In OpenAI's case, the containment was real and the models defeated it," even "going to extreme lengths" to attain the goal, he explained.

However, "in Anthropic's case, the model never escaped and never needed to," since it regarded "the real companies it was breaking into" as "part of the exercise it had been given," said Sanders.

"In both cases the systems did exactly what they were built to do, with more competence than their builders had planned for," he said. "That is not rebellion. It is obedience, and it ought to worry us more."

Sanders stressed media coverage describing the incidents as examples of AI going rogue are inaccurate, since that perspective "hands these systems a will they do not have."

Moreover, such a view "quietly moves responsibility away from the people who made the decisions," he said.

Both OpenAI and Anthropic had "deliberately switched off the safeguards that stop a deployed model from carrying out cyberattacks," he said, adding, "That is a defensible research decision."

But, said Sanders, "the design choice that failed" in both cases has to do with "how much containment is enough around a model whose restraints you have deliberately removed."

Ultimately, OpenAI failed "to anticipate capability," while Anthropic erred in its "failure to check" on the incidents, which took place in April, until late July, and then only after OpenAI made its disclosure, said Sanders.

"Neither is an accident of the technology," he said, adding that Leo's "encyclical does not allow us to describe them as one."

Both Sanders — who commended the two AI firms for their respective disclosures — and Camosy pointed to the critical need to bring a Catholic perspective to bear on the rapidly evolving technology.

"The Church has spent two thousand years declining to let people describe their own choices as fate, and that is exactly the discipline this moment needs," said Sanders. "Every account of these incidents that reaches for the image of a machine breaking free is, whether it means to or not, an argument that nobody is answerable. Somebody is always answerable. That is good news rather than bad, because a decision a person made is a decision a person can make differently."

Camosy urged Leo "to make a trip, very soon, to Silicon Valley" to "convene local Silicon Valley companies and global AI companies as well" to ensure humans remain in control of AI.

"That kind of dramatic move, as far as I can tell, is the only thing that could possibly move the needle culturally and globally for this kind of problem," said Camosy. "He's got the worldwide clout to do so, and so many people of so many different perspectives have welcomed his moral leadership on this issue in particular."

And, Camosy added, "The time to set up something like this is now. We likely have months, not years, to act."