News
An Anthropic web page is seen in this March 1, 2026, illustration. Christopher Olah, the lead c

An Anthropic web page is seen in this March 1, 2026, illustration. Christopher Olah, the lead of Anthropic's interpretability team, will appear April 25 with Pope Leo XIV when he presents his encyclical "Magnifica Humanitas." (OSV News/Reuters/Dado Ruvic)



by Brian Roewe

NCR environment correspondent

View Author Profile

broewe@ncronline.org
Follow on Twitter at @brianroewe

Join the Conversation

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

May 22, 2026
Share on BlueskyShare on FacebookShare on TwitterEmail to a friendPrint

When the Vatican announced this week a co-founder of leading AI lab Anthropic will join the presentation of Pope Leo XIV's encyclical on artificial intelligence, *Magnifica Humanitas* ("Magnificent Humanity"), it didn't seem like an obvious pairing at first. The Vatican has been critical of AI, offering warnings about the technology's threats to human dignity and the environment, and even launching an AI study group days earlier.

But the May 25 presentation won't be the first time Christopher Olah, the lead of Anthropic's interpretability team, has been in a room filled with Catholics. Instead, it will be the latest step in the multibillion dollar AI company's outreach to religious

leaders.

In March, a small group of Christians gathered at Anthropic's San Francisco headquarters for a meeting organized partly by Olah. Among those in attendance was Meghan Sullivan, a philosopher at the University of Notre Dame who directs its Institute for Ethics and the Common Good.

"Anthropic is one of these tech companies that really cares about educating all communities, including faith communities, about how these powerful AI models work and what they're good at and what the potential risks are," she said. "And I think right now it's a crucial time for Catholics to really understand this technology and how it's changing our lives and our society and what's likely to happen in the future."

The meeting, as The Washington Post reported at the time, was aimed at informing the 15 Christian leaders about how AI models, like Anthropic's widely popular Claude large-language models, or LLMs, are built, trained and operate. They also solicited feedback on ethics and morals to employ in the development of Claude.

Olah, 33, and his team have organized several meetings with religious leaders and philosophers. Sullivan described him as "very invested in these dialogues" and a good listener.

Now, Olah and Anthropic — which earlier this year engaged in a standoff with the Trump administration over military use of AI — will take a seat at the Vatican May 25 to listen as Leo unveils *Magnifica Humanitas*, focused on "the protection of the human person in the age of artificial intelligence."

"Of all the people in the world who work on AI that they could have chosen from industry, he's probably the right person," said Brian Patrick Green, director of technology ethics at Santa Clara University, located in the heart of California's Silicon Valley.

"Anthropic is the company that has really staked their position as the ethical AI company, saying no to the U.S. government when it comes to lethal autonomous weapon systems … and against mass surveillance of Americans," he said. "So they put down those two stakes in the ground and said we're not going to go past this point."

Anthropic prioritizes ethics but still faces controversies

In a rapidly developing industry, Anthropic has garnered a reputation for prioritizing ethics and caution. A March Time magazine profile — titled "The most disruptive company in the world" — called it "the frontier AI lab with the greatest emphasis on safety." The company has roots in the effective altruism movement, and its staff includes an in-house philosopher.



Dario Amodei, CEO and co-founder of Anthropic, middle, speaks on a panel between Tino Cuéllar, president of the Carnegie Endowment for International Peace, left, and Elizabeth Kelly, director of the U.S. AI Safety Institute, at the convening of the International Network of AI Safety Institutes at the Golden Gate Club at the Presidio in San Francisco, Nov. 20, 2024. (AP/Jeff Chiu)

Valued at $900 billion, Anthropic was founded in 2021 by siblings Dario and Daniela Amodei and five other former employees of OpenAI, the AI developer created by Sam Altman and Elon Musk, among others, behind ChatGPT. They departed their now-rival over worries that OpenAI was moving too fast without thorough testing.

It has since emerged as one of the leading AI labs, with its Claude AI assistant drawing 30 million monthly users and its Claude Code used extensively by companies in cloud services and data analysis. It has partnered with Google and Amazon in their cloud computing systems.

On its website, Anthropic lists seven values guiding its work, beginning with striving "to make decisions that maximize positive outcomes for humanity in the long run."

"This means we're willing to be very bold in the actions we take to ensure our technology is a robustly positive force for good," the AI company wrote. "We take seriously the task of safely guiding the world through a technological revolution that has the potential to change the course of human history, and are committed to helping make this transition go well."

> 'We're willing to be very bold in the actions we take to ensure our technology is a robustly positive force for good.'
> —Anthropic
>
> Tweet this

Guiding its Claude models is an overarching constitution that aims to define its values and behavior, primarily to be broadly safe, broadly ethical and genuinely helpful.

Several Catholics contributed to its drafting, including Green; Fr. Brendan McGuire, a Silicon Valley priest and co-founder of Santa Clara's Institute for Technology, Ethics and Culture; and Bishop Paul Tighe, secretary of the Vatican's Dicastery for Culture and Education who co-authored *Antiqua et Nova* on the relationship between AI and human intelligence.

"It's very much a virtue ethics perspective," Green said, "which is trying to train the AI into some sort of moral decision-making process, where it identifies as being a good decision maker, and then, based on the fact that it identifies as a good decision-maker, it then tries to make good decisions."

AI companies across the board have faced a host of controversies and public backlash, including broad opposition to the rapid construction of resource-heavy data centers and claims against OpenAI that their products urged some users to

follow through with suicide attempts.

Anthropic, meanwhile, agreed in August to a $1.5 billion settlement with thousands of authors and others with reproduction rights who alleged the company illegally downloaded nearly a half million copyrighted books to train its AI language models. It was the largest proposed copyright settlement in U.S. history.

But Anthropic has also supported national standards around AI transparency and testing and evaluating models. Dario Amodei has lobbied against a proposed 10-year federal moratorium on state-level AI regulations, which has the backing of the Trump administration, arguing that such a ban is essential to keep the U.S. competitive.

In February 2025, Anthropic caught the attention of AI ethicists when it delayed the release of a new version of Claude after trials had indicated it could be used by terrorists to develop biological weapons, Time reported. According to Axios, the Pentagon used Claude in the Jan. 3 raid that captured Venezuelan president Nicolas Maduro.

Pope Francis greets Bishop Paul Tighe, secretary of the Dicastery for Culture and Education, d

Pope Francis greets Bishop Paul Tighe, secretary of the Dicastery for Culture and Education, during a meeting with leaders from the tech industry at the Vatican March 27, 2023. Tigh co-authored the Vatican document "Antiqua et Nova" on the relationship between AI and human intelligence. (CNS/Vatican Media)

But its relations with the White House disintegrated in February after Anthropic opposed the use of its technology in the development of autonomous weapons and mass surveillance of U.S. citizens. In response, President Donald Trump canceled the company's $200 million Pentagon contract and declared Anthropic a "supply chain risk," blocking federal agencies from working with the company.

Anthropic challenged the move in court. Fourteen Catholic theologians and scholars, including Green, in an amicus brief stated the AI company "was acting as a responsible and moral corporate citizen."

Green told NCR that Anthropic's decision to stand firm at the cost of its business against certain uses of its technology was worth commending, though he noted the company has not ruled out entirely the use of autonomous weapons, only that the

technology today is not reliable.

"By no means are they perfect, and they would admit that they're not perfect," he said.

Related: By refusing the Pentagon, Anthropic holds moral line on AI

Vatican seeks to offer AI ethics guidelines 'aimed at the whole world'

It is not clear what led the Vatican to invite Anthropic to the encyclical's presentation and the company is notably absent from signing on a Vatican document pledging an ethical approach to AI.

Over the past decade, the Vatican, along with Pope Francis, has met with AI developers and leaders from Silicon Valley, including IBM, Microsoft and Google, in both conferences and private meetings dubbed the Minerva Dialogues, a reference to the Roman church that held the 17th-century trial of Galileo.. In 2020, the Vatican's Pontifical Academy for Life developed the Rome Call for AI Ethics to promote ethical approaches in the development of artificial intelligence. Microsoft, IBM and Cisco are among those who have signed onto the document. Anthropic has not.

Anthropic did not reply to a request for comment.

Advertisement

Green suggested that Anthropic's past engagement with religion could have been a factor in its invite to the Vatican. Likewise, he said, "it shows that the church is willing to talk to people across great divides of understandings of how AI should work in the world, ultimately bringing people together so they can talk about these sorts of things."

Sullivan, who at Notre Dame is directing a $50.8 million grant to create a faith-based ethical framework for appropriate uses of AI, sees in working with several big tech companies a growing interest in theology and philosophy.

"They're trying to build this powerful, all-purpose artificial intelligence and shape its behavior. And so, of course, if you want expertise on how to shape something's virtues and behavior, philosophy and theology have been thinking about these

questions for 2,000 years. And it turned out to be extremely relevant to how you safely build these new AIs," she said.

In the past two years the Pontifical Gregorian University in Rome has held the Builders Artificial Intelligence Forum, bringing together AI companies alongside Catholic organizations. A third meeting is set for the fall.

"The interdialogue between the Vatican and [AI] developers has always been around the past 10 years," said Paulist Fr. Ricky Manalo, author of the forthcoming *The Catholic Handbook on Artificial Intelligence* who has taken part in the BAIF gatherings.

Participants at the 2025 Builders AI Forum gather at the Pontifical Gregorian University in Ror

Participants at the 2025 Builders AI Forum gather at the Pontifical Gregorian University in Rome Nov. 6-7, 2025, to discuss how emerging technologies can serve the church’s mission. The event highlighted Pope Leo XIV's call to place artificial intelligence at the service of evangelization and human dignity. (CNS/Robert Duncan)

The May 25 presentation of *Magnifica Humanitas* in some ways marks the peak of those dialogues, the priest said. He cautioned against viewing Anthropic's participation in the encyclical's presentation as any kind of church blessing upon the company.

Rather, he said inviting Olah, who specializes in mechanistic interpretability — essentially understanding how and why AI models reach outputs, which can help then identify and avoid problematic behaviors — could signal a sign of interest in a very specific and important area of AI ethics and development. In an interview with the Atlantic, Olah, who is an atheist, likened his role to a priest guiding Claude to "be a good person, in some sense."

The Catholic Church recognizes tech companies and governments require far more moral discernment with regard to the power that AI holds, Sullivan said, and that has led it to seek to take a leading role in filling that void.

Likening her excitement for *Magnifica Humanitas*' release to a Taylor Swift fan anticipating a new album, the Notre Dame philosopher hopes to see the pope stake out a full-throated theological defense of the value of every human life beyond

productivity or skills and the dignity of labor in the new economies emerging from AI technology.

"I think that we're going to see on Monday that when Pope Leo and the Catholic Church releases a teaching on an issue as serious as human dignity in the era of AI," Sullivan said, "it's a teaching that's not just for Catholics, not just for bishops and priests, but it's really a teaching that's aimed at the whole world, including the people who are building this technology."

Read this next: The church defended workers from industrial capitalism. Now Pope Leo XIV is taking on AI.
This story appears in the **AI Encyclical: Magnifica Humanitas** feature series. View the full series.