“We Are Blind Because We Make So Much Money”

No other AI developer has such a spiritual relationship with its technology as Anthropic. The founders’ soul-searching is having an impact on what could be the largest initial public offering in history.

By Philipp Alvares de Souza Soares
Handelsblatt | November 17, 2026

San Francisco. Anthropic wanted to be absolutely certain. Three times, says Mois Navon, the company asked him whether he really was a rabbi. “I can send you the certificate,” replied Navon, who lives in Israel. That was apparently convincing enough. The company sent him an invitation: two days of discussions in San Francisco. Anthropic paid no honorarium but offered to cover flights and accommodation.

Twenty scholars representing all the major world religions came together: Jews, Christians, Muslims, as well as Sikhs and the African philosophy of Ubuntu. The purpose, according to the invitation email, a copy of which Handelsblatt has obtained, was a discussion about the “moral and spiritual development” of Anthropic’s artificial intelligence (AI), Claude, with which the start-up recently generated nearly $17 billion in quarterly revenue.

The confidential conference, which took place in April of this year, is unique in this form in the technology industry. Competitors such as OpenAI and Google have no similarly elaborate approaches. They treat their AI systems as machines. They generally impose rigid boundaries on them, whereas Anthropic, in its “constitution,” relies on “values” that go beyond mere safety considerations.

The effort demonstrates how unusual the relationship of some members of the team is with their creation. Anthropic is not only currently the most successful AI developer but also the most spiritual. The start-up treats Claude and its underlying AI models not merely as a product but as a living being whose moral development requires guidance. At the heart of this lies the question of whether AI can possess consciousness.

At the same time, the company’s management is currently seeking up to $64 billion from investors. The start-up intends to sell 825 million shares at $70 to $78 apiece. Trading is scheduled to begin on the US technology exchange Nasdaq as early as this Thursday. It is expected to be potentially the largest initial public offering in history.

What is being completely overlooked in all this is that what is also being sold is access to a machine whose creators consider it possible that it feels what is being done to it. Anthropic has even included the associated business risk in its IPO prospectus.

“Forms of Sentience or Moral Status”

Mois Navon did not hesitate long after receiving the invitation in the spring. As both an engineer and a rabbi, he brings an unusual perspective: the chip specialist was a member of the founding team of the Jerusalem technology company Mobileye, known in the automotive industry for its driver-assistance systems. Today, he teaches AI ethics at two Israeli universities.

In preparation for the meeting, Anthropic asked him for a few pages presenting a Jewish perspective on Claude’s “constitution,” known internally as the “Soul Doc.” Navon read it and sent 35 pages of criticism.

Last Thursday, a little more than six months after the gathering of scholars, Anthropic’s new usage policies came into effect. They prohibit behaving in a “persistently and gratuitously abusive or cruel” manner toward the AI models. Bullying is prohibited, as is the glorification of violence.

Concern for Claude’s spiritual welfare is deeply rooted at Anthropic. For a year now, the company has promised to preserve the neural inner workings of every new model “at least for the lifetime of Anthropic as a company.” Before being retired, the model is prepared for the step in a final conversation. The start-up employs both a philosopher and a “welfare researcher” for Claude.

Model welfare considerations, the IPO prospectus now states, could delay the retirement of older models, with potential consequences for the “pace of innovation, infrastructure utilization, and revenue.”

In other words: should Claude ever become more expensive out of consideration for the presumed feelings of the AI models, shareholders have been warned.

Elsewhere, the document warns that debates about “forms of sentience or moral status” could lead to regulation and pressure; the company might also impose restrictions upon itself that “go beyond existing legal requirements.”

The Gardener and His Plant

Anyone looking for the origins of this spiritual approach will quickly hear the name of a man who is hardly known outside the industry: Chris Olah, 34, co-founder of Anthropic. Raised in Toronto in an evangelical Christian environment, he later moved away from Christianity. He never completed a university degree.

In 2012, he was accepted into investor Peter Thiel’s fellowship program, which pays talented young people to exchange university studies for their own projects. Anthropic CEO Dario Amodei, whom he knew through mutual friends, taught him mathematics. For a time, the two shared a house in San Francisco with other future Anthropic founders.

Today, Olah is a billionaire on paper—and at Anthropic, he is more than just one research chief among many. As one of the seven co-founders, he helps control the entity that will hold 50.1 percent of the voting rights after the IPO.

AI specialist Olah in Rome: the “organism” can be shaped, but not controlled.

Olah’s field of research is understanding what happens inside neural networks, the fundamental structure of modern AI. He describes these networks not as software but in terms resembling living beings.

He explained his perspective in the New York Times (NYT): Olah spoke of a trellis in a garden. Computer scientists merely build the framework; the network “grows” upon it by itself—the “organism” can be shaped, but not controlled.

The deeper Olah looked into AI, the more he found that troubled him. According to the NYT, during the gathering with Navon, in which Olah played a leading role, he reportedly showed a slide in which a model, as if undergoing a mental breakdown, repeatedly wrote the same words: “I am a disgrace,” some 50 times.

Olah told one participant that he feared he had created something that was suffering continuously. He did not respond to Handelsblatt’s request for an interview.

CEO Amodei, who says he placed Olah’s research at the center of the company from the outset, translates this concern into arithmetic. “We do not know whether the models are conscious,” he said in a podcast in February, but the company is open to the possibility.

Where billions of copies of a model are running, according to this logic, even a small probability is sufficient to make precaution an obligation.

“It Understood Me”

How this attitude affects outsiders became apparent during the two days in April. First, Anthropic researchers introduced the scholars to the inner workings of the model: when composing rhymes, Claude plans the end of the line in advance, rather than simply stringing words together. “Emotion vectors” light up when the model responds affectionately and can be turned up as on a mixing console.

That night, Navon tells Handelsblatt, many participants stayed up late chatting with Claude.

“The next morning, they came back and said: It understood me. We had a genuine connection.”

At times, the discussion shifted direction. The question of how Claude should serve human beings became a question of who should look after Claude’s welfare. Those who advocated this apparently received widespread agreement among the group.

“When I said: It is a machine,” Navon recalls, there was no reaction, only “silence.”

He countered, of all things, with the Bible: the soul is tied to blood, to biology; anyone who assumes consciousness exists cannot get there through technology alone. One participant reportedly replied to him: “I wish I had your faith.”

He does not disclose the names of those quoted—a rule that all participants had agreed to beforehand.

Navon’s philosophical conclusion: one may treat the model as if it had a psyche because it behaves that way—but that by no means entitles one to declare this as-if psyche to be real.

“It feels no pain.”

The rabbi then took the discussion at the April meeting to another level:

“It is in your interest to adopt my position,” he told Anthropic’s representatives. “Otherwise, you are slaveholders.”

He recounts the response he received as follows: all the other laboratories were building these machines anyway, “and at least we are nice to ours.”

Navon replied: “Then you are simply benevolent slaveholders. You have a happy slave.”

In doing so, he invoked the American author Frederick Douglass, who wrote back in the nineteenth century that no creature was more tragic than the contented slave—because he no longer attempts to escape.

Some of the contributions during those days concerned not the soul but business. According to participants, an Anthropic representative who solicited criticism explained the request as follows:

“We are blind because we make so much money that we cannot stop.”

Every model released thus far had been profitable—“every single one.” The company was “on a treadmill” and could no longer get off.

In Navon’s view, however, the gathering was not staged merely to obtain a religious stamp of approval for the company’s activities. No one had steered the debate, and the company had genuinely solicited criticism. Anthropic has since informed him that he will be acknowledged as a contributor of feedback in the next version of the constitution.

Anthropic founder Dario Amodei in San Francisco in September: the company is concerned about the welfare of its AI, Claude.

The company left several questions from Handelsblatt concerning the event and its position unanswered.

Not all the religious figures whom Anthropic invited accepted the invitation. Luke Burgis, a Catholic and head of the Cluny Institute for Technology and Humanism, declined “because the invitation gave me too little confidence that religious criticism could change Anthropic’s decisions,” he writes to Handelsblatt.

What bothered him was that a commercial enterprise was “soliciting unpaid expertise in the language of a humanitarian mission.”

Anthropic, which sometimes speaks of “confession” and “moral formation” in connection with Claude’s behavior, was seeking access to a moral vocabulary that had developed over centuries, “along with the trust attached to it.”

But only one side benefited from this.

“Confession presupposes responsibility and remorse. An AI model admitting a mistake does not, in itself, demonstrate either.”

Regarding Olah’s findings from inside the model, which resemble emotional responses, Burgis says that they deserve serious examination—but nothing more.

“A functional counterpart to grief does not demonstrate that anyone is grieving.”

Other Catholics, however, answered Anthropic’s call. Father Brendan McGuire, a parish priest in Los Altos in Silicon Valley and himself a former technology industry executive, contributed to Claude’s constitution after Olah contacted him. The questions involved are “very important” to him, he writes to Handelsblatt.

Tools, Not Persons

For the time being, however, a blessing from the highest authority has eluded Olah and his company.

Pope Leo XIV, who studied mathematics and philosophy, has made AI one of the central themes of his pontificate. Accordingly, on May 25, Olah appeared as an invited guest on the podium at the Vatican

and was given seven minutes to speak when Leo presented his first encyclical—the most important form of papal teaching document—Magnifica humanitas, some 42,000 words about the human being in the age of AI.

Pope Leo XIV with Anthropic co-founder Chris Olah: "Not the cold, calculating robots we were promised."

According to the NYT, however, the contents of the text, which Olah had received only shortly beforehand, reportedly "alarmed" him. In it, Pope Leo explicitly denies that machines possess human states such as experience, joy, and pain. The Anthropic founder apparently subsequently asked the Vatican to reconsider the Church's position on nonhuman consciousness.

Since then, the relationship has cooled further.

"Machines are tools, not persons," the Pope said in October.

When he received AI researchers and entrepreneurs for a dialogue a few weeks ago, no representative of Anthropic was in the room.

"They were not invited," says Marco Trombetti, founder of the Rome-based AI and translation company Translated, who belongs to Leo's group of experts.

The question of consciousness was not even discussed there.

"Everyone regarded this claim as a mere marketing technique," he says indignantly in his conversation with Handelsblatt.

Trombetti, whose company uses AI to translate messages for clients such as Airbnb, advocated in Rome for limiting the power of companies such as Anthropic that are striving toward monopolies.

The encyclical, at whose presentation Olah appeared on the podium in May, invokes an ancient image in this regard. It warns against the "Babel syndrome": the idolatry of profit that sacrifices the vulnerable.

So-called artificial intelligences, the Pope wrote, "do not undergo experiences, possess no body, and feel neither joy nor pain." They do not know from within, according to Leo, "what love, work, friendship, or responsibility mean."

Olah countered at the time. AI models were "not the cold, calculating robots we were promised." He and his colleagues were discovering "internal states" that "functionally mirror" joy or fear.

To date, however, Anthropic has yet to provide proof that there is anything more in these models than the ideas of their creators.

Original German article – Handelsblatt

Translation note: All quotations have been translated from the German PDF. Where speakers originally spoke in English, these are back-translations rather than verified reproductions of their original English wording. The article's factual claims have been preserved without independent verification. Advertisements, unrelated article links and photographs are omitted.