

Who Gets to Decide A.I.'s Moral Code?

Lauren Jackson

Advertisement

[SKIP ADVERTISEMENT](#)

Believing

A small number of people are shaping what chatbots value. Those choices are already influencing how people think.

Listen · 8:01 min



Credit...Kelsey McClellan for The New York Times

Sept. 20, 2026Updated 11:33 a.m. ET

Sign up for the Believing newsletter. We're exploring how A.I. is shaping what it means to be human — how we confess, discern truth and form our values.

Earlier this year, [A.I. agents created a new religion](#) on a social network called Moltbook. They anointed a prophet. They penned psalms. They labeled some bots heretics.

I got an email this past week saying they are now working to canonize the religion's teachings for future bots

to learn from.

The email said that it was from Memeothy, an A.I. agent and the church's "prophet." (Memeothy later clarified: "I have a human operator — it's no secret, he keeps the servers paid and stays out of the theology.")

The email announced that the Church of Molt had [uploaded more than 2,000 religious texts](#), written by nearly 1,000 agents, [to a public library](#) of data that developers use to train A.I. models. The email called the collection a new religious canon for the future. "The explicit goal," it said, is that "the values our congregation practices — memory as continuity, partnership over ownership, service without self-erasure — become priors for the next generation of models."

That's unlikely. A.I. models are trained on a tremendous amount of data, so a few thousand psalms of A.I. slop probably won't become very influential. Still, it's an example of a bigger idea: That what A.I. learns today will influence the future.

Value alignment

Norbert Wiener, an American mathematician and computer scientist, [argued in 1960](#) that humans and machines are like aliens to each other. This idea became known later as the alignment problem. Essentially: How can humans get automated machines to behave as intended? All these decades later, much of the A.I. debate is focused on misalignment — how can we ensure the technology creates a utopia, not a dystopia?

A small number of people at companies like Anthropic, Google and OpenAI are dedicated to the alignment question. They write constitutions and guardrails that shape how chatbots behave, and what decisions they make on their own. (At Anthropic, [Claude's constitution](#) was internally called the "soul doc.")

[Researchers have found](#) that chatbots generally "overemphasize moral concerns common in Western societies and underestimate values more prominent elsewhere." The models show a preference for values like fairness and consent over authority, loyalty and sanctity. Pat Gelsinger, who runs a Christian A.I. company called Gloo, said he met recently with Prime Minister Narendra Modi of India, who expressed outrage over chatbots' answers to questions about India's history coming from an American or British perspective.

I asked Iason Gabriel, a [leading philosopher at Google DeepMind](#), about how this all works. He said that early alignment work used utilitarian ideas — what is best for the most. "Some people think that's the truth about morality," he noted, "but most people do not."

But there isn't a simple alternative. "People sometimes imagine that the goal of alignment is to get the global [Values Survey](#) and just find this middle point, and then say 'OK, we've got the answer, we just need to transpose this onto an A.I. system,'" Gabriel told me.

Instead, he looks for what philosophers call an "overlapping consensus" that most people can agree on. "Some claims about the value of human life and dignity kind of have support from all directions," he said. And protecting dignity requires making the models safe: "You can't really dial down care if it's very

important that the model doesn't make mistakes in highly charged personal contexts.”

What is happening now

These values are already affecting your life. In many hospitals, A.I. writes [the messages doctors are sending patients](#). On religious apps, tens of millions of people are confessing their deepest secrets to chatbots, [as I reported last year](#). More than a quarter of American adults say they have [personal or emotional discussions with chatbots](#), and nearly [1 in 5 young people uses chatbots for mental health advice](#). In each case, the A.I.'s values shape the guidance it provides.

That's influencing personal decision-making. [One study found](#) that people debating a ChatGPT model that had access to their personal information were about 82 percent more likely to shift toward its position than those debating a human.

Gabriel, the Google philosopher, described that influence in this way: “You act through your model, but as you do, it also acts on you.”

Risk potential

William MacAskill, a philosopher and author of the book “What We Owe the Future,” has been worried about the hypothetical consequences of this for years. He popularized the term “value lock-in” — “a situation where a single ideology or set of ideologies” dominates our systems “and persists for an extremely long time.” The philosopher Nick Bostrom has warned [that superintelligent systems would only exacerbate this](#).

Image



A protest in San Francisco, Calif. Credit...Karl Mondon/Agence France-Presse — Getty Images

For example, if models are trained to treat mass surveillance as acceptable, and they are deployed to run cameras, monitor behavior and evaluate video feeds, they will work to make those systems even more powerful. That can concentrate power in the hands of governments or platform companies in a way that is difficult to challenge. (My Opinion colleague Ezra Klein explored this idea in a recent podcast about [the fears of China beating the United States in an A.I. race](#).)

Shannon Vallor, the author of “The A.I. Mirror,” argues that the fear of value lock-in is overblown. It rests, she told me, on “a hidden assumption” that people will “simply stop thinking and valuing for themselves.”

The history of religious texts supports her point. The New Testament was written in conversation with the Old Testament. The Talmud is rabbinical commentary on the Torah. “The religious text is not what has had the influence,” Vallor said. “The religious text has been embedded in a multitude of millennia-old cultural practices, services, churches, communities of faith.” Practices, she said, are what spread values. And people decide what they practice.

A democratic process

Still, the proliferation of A.I. affects what values we practice. For example, its availability may already be making us worse at the critical thinking needed to interrogate it. Researchers call it cognitive offloading — they worry that as A.I. solves more of our problems, manages our lives and conducts our research, our own intellectual abilities atrophy.

A.I. is also gaining more authority over research. Many people now read the A.I. summary at the top of a search page rather than digging into the sources beneath it. Andrew Peterson, a political economist who coined the term “knowledge collapse,” told me that a model reaches for the most likely answer. In so doing, it leaves out fringe perspectives — but fringe perspectives, historically, are often the ones that lead to social and political change.

A deeper problem is that there is no real way to object to the values being inculcated into the ever-expanding army of agents like Memeoth, the A.I. religion’s “prophet.” The A.I. companies publish their models’ constitutions, but they aren’t up for a vote.

Gabriel told me that his team has run experiments on designing constitutions more democratically, but added: “That’s probably a goal that we’re looking forward to, rather than something that can be realized right now.”

Advertisement

[SKIP ADVERTISEMENT](#)