IDEAS

The Real Reason Anthropic Wants Guardrails

AI is too powerful and too new to be set free from human oversight.

By Thomas Wright

Illustration by The Atlantic. Sources: Anna Moneymaker / Getty; ClassicStock / Getty.

FEBRUARY 27, 2026 SHARE AS GIFT DISCUSS REMOVE

Sign up for our newsletter about national security here.

On Tuesday, in a closed-door meeting, Secretary of Defense Pete Hegseth issued a blunt ultimatum to Anthropic CEO Dario Amodei: Strip the ethical guardrails from your AI models by Friday or face the full weight of the state. The terms of the threat were stark. If Anthropic does not allow the Pentagon "all lawful uses" of its Claude models, Hegseth will invoke the Defense Production Act to compel cooperation, he warned—or, even more devastatingly, designate Anthropic as a supply-chain risk. The

latter would effectively blacklist Anthropic from doing business with any entity that touches the Department of Defense.

Yesterday evening, Amodei gave his answer. He rejected Hegseth's "best and final offer," writing,

"I believe deeply in the existential importance of using AI to defend the United States and other democracies, and to defeat our autocratic adversaries." However, he continued, "in a narrow set of cases, we believe AI can undermine, rather than defend, democratic values." He concluded that the Pentagon's "threats do not change our position: we cannot in good conscience accede to their request."

Will Gottsegen: Anthropic takes a stand

This is not just an ethical dispute. It is a battle over whether to manage the national-security risks that will inevitably be associated with ever more powerful AI. If Hegseth follows through on his ultimatum, it will weaken the U.S. military and increase the likelihood of a catastrophic accident.

Anthropic has insisted that its Claude AI model not be used for domestic surveillance or to build autonomous weapons without a human involved. The company's statement makes clear that its only principled objection is to mass surveillance. It is not opposed to autonomous weapons per se and has already carved out exemptions for missile defense and cyberoperations. The company's hesitation regarding autonomy is technical: Large language models are simply not yet reliable enough to operate without a human in the loop. Pushing them too far, too quickly, invites a mistake that could prove disastrous. Anthropic is asking for an exclusion on autonomous weapons not out of an ideological refusal to fight, but to allow for the research and development necessary to make such systems safe.

The truly unbridgeable divide is the one over domestic surveillance. DOD has the authority to conduct domestic surveillance in support of a civilian agency. Under an administration that invoked the Insurrection Act, or that sought to map domestic dissent, the Pentagon's demand for "all lawful uses" of Anthropic's models could become a skeleton key. Amodei articulated this danger in a recent interview with Ross Douthat, noting that, although it isn't illegal to record conversations in public spaces, the sheer scale of AI changes the nature of the act. As Amodei put it, AI could transcribe speech and correlate it in a way that would not only identify one member of the opposition but "make a map of all 100 million. And so, are you going to make

a mockery of the Fourth Amendment by the technology finding technical ways around it?"

The Pentagon's logic relies on a traditional procurement analogy: Lockheed Martin does not tell the Air Force how to fly the F-35s it makes, so why should Anthropic tell the military how to use Claude? A democratically elected government should be free to make those choices. This sounds reasonable on its face but does not account for the uniqueness of AI. Unlike nuclear energy and the internet, both of which were born in government labs, AI was conceived and honed entirely within the private sector. It is a general-purpose technology with the potential to upend the global balance of power.

These circumstances obligate AI companies to work with the government in thinking through the risks associated with their product, especially because they have a greater understanding of it than many in government. After all, if Anthropic removed all of the conditions on autonomous weapons and the models behaved in unexpected and dangerous ways, the company would certainly be held responsible.

AI scientists have been trying to encourage a public discussion about managing these risks. In 2023, dozens of AI leaders, including Amodei, Sam Altman of OpenAI, Demis Hassabis of Google DeepMind, and Kevin Scott of Microsoft issued a statement saying that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Earlier this year, Mustafa Suleyman, the head of Microsoft's AI team who wrote a book on AI risks called *The Coming Wave*, told the BBC, "I honestly think that if you're not a little bit afraid at the moment then you're not paying attention."

Nobody has been more outspoken about both the hazards and the potential than Amodei, who has published a series of essays over the past 18 months on the future of AI. His starting point is that an extremely powerful AI—what some people call artificial general intelligence—is near at hand. By this he means that an AI model will be as capable as a Nobel Prize winner in every field. Once made, such a model can be easily cloned millions of times. As he puts it, this would be equivalent to a country of geniuses in a data center.

In his first essay, "Machines of Loving Grace," Amodei wrote about how this model could allow humans to advance research and development in many fields 20 times faster than would otherwise be the case. We may see a century's worth of progress in medicine and biological sciences in five years. In this "compressed 21st century," he believes we could plausibly secure the reliable prevention and treatment of nearly all

infectious diseases, the elimination of most cancers, the prevention of Alzheimer's, and huge progress on countering genetic diseases.

But more powerful AI also comes with risks. Amodei recently published his second essay, "The Adolescence of Technology," which discusses the flip side of the coin—the ways that more powerful AI could endanger the United States and humanity in general by allowing individuals to build bioweapons or by empowering authoritarianism. He wrote it, he told a panel in Davos, "to jolt people awake."

One of Amodei's concerns is the possibility that the country of geniuses could turn hostile or disruptive. In a lesser-known, shorter post, "The Urgency of Interpretability," he admitted that "we do not understand how our own AI creations work." A flaw in regular technology is generally a programming mistake that can be fixed. But Amodei cited Chris Olah, a co-founder of Anthropic, in noting that AI systems are not so much built as grown. "You can set the high-level conditions that direct and shape growth," Amodei wrote, "but the exact structure which emerges is unpredictable and difficult to understand or explain." The models may evolve and behave in ways that their creators can neither anticipate nor easily observe, let alone fix.

Anthropic has conducted experiments to figure out the true nature of its AI agents. It found that some are prone to lying and will blackmail their engineers even if instructed not to. In the shorter post, Amodei wrote, "These systems will be absolutely central to the economy, technology, and national security, and will be capable of so much autonomy that I consider it basically unacceptable for humanity to be totally ignorant of how they work." Taking the time to properly understand how these models evolve and behave would allow their operators to identify and disable the ones that run amok.

Amodei recommended that all labs develop "a true 'MRI for AI,'" but he acknowledged that they might not have enough time, given how quickly AI is advancing. This interpretability problem gets to the core of Anthropic's concern about autonomous weapons.

The public narrative often conflates Anthropic's guardrails with anti-war sentiment, but this is not a sequel to the Project Maven controversy of 2018, when Google employees revolted over drone-targeting contracts. That was a story of internal dissent within a company hesitant to help the military wage war. Anthropic is a different beast entirely. It was the first AI firm to deploy its models in classified systems and has

shown a willingness to integrate with the defense establishment. Anthropic's clash with the Pentagon is not one between pacifism and militarism but a fundamental dispute over managing the risks of the most transformative technology since the splitting of the atom.

From the March 2026 issue: America isn't ready for what AI will do to jobs

The company has real differences with the Trump administration over foreign policy. Anthropic is notably hawkish on China, favoring tougher policies toward Beijing than Trump's accommodationist and commerce-centric approach, and more concerned by authoritarianism. The company is also more outspoken about the risks AI poses to biosecurity and the labor market. David Sacks, the administration's AI czar, has dismissed such concerns as doomerism and accused Anthropic of running a "sophisticated regulatory capture strategy based on fear-mongering." The administration has rejected state-level regulations on AI on the grounds that some states would try to insert a "woke ideology" into AI, and that a contradictory patchwork would hold America back in the AI race against China. But it has yet to produce a federal bill to fill the void.

The leaders of AI companies acknowledge that they do not know what they are building. But they don't want to stop. Some, like Amodei, worry that China will get there first, which would pose a greater threat. Others believe the benefits outweigh the risks. But most want more time so that society and government can properly adjust and regulate AI where needed. They worry that the speed of advancement will outstrip the world's ability to manage the risks. Some have suggested slowing China down with export controls to buy more time, but the administration has rejected that logic.

There is now a real chance that many AI companies will think twice before working with the U.S. government and will focus on their commercial work instead. The message Hegseth is sending to Silicon Valley is that if a company partners with the Pentagon and makes a wrong turn, the administration will effectively nationalize it or designate it as a supply-chain risk and burn it down.

Axios recently reported that the Pentagon views Google's Gemini as a potential replacement for Claude. Perhaps, but Hassabis, who oversees all of Google's core AI research and development, has a long history of concerns about AI risks and a belief, stronger than Amodei's, in the necessity of global governance. It's hard to imagine him

complying with Hegseth's demands. That does leave one AI leader who is keen to fill the vacuum: Elon Musk, with his model, xAI. If Hegseth sticks to his demands, the Pentagon could become dependent on xAI as its sole supplier. This would deprive the U.S. government of most of the AI industry's talents, give Musk enormous leverage over future administrations, and create a single point of failure, which could prove catastrophic. No company, not even Anthropic, should be the sole supplier of classified AI to the government.

Hegseth's ultimatum rests on a simple premise: that the government, not private companies, should decide how it uses powerful technologies. In most cases, that principle is sound. But here it obscures two deeper problems. It minimizes the risks to domestic liberties, and it assumes a level of understanding that does not yet exist. The engineers building these systems acknowledge that they do not fully understand them, and that the models behave in ways that can be difficult to predict or control. Demanding unconditional access before those systems are ready is not an assertion of authority. It is a wager that the unknowns will not matter.

The danger is not that Silicon Valley will wield too much power over the military. It is that neither will fully understand the systems it is rushing to deploy—and that the consequences of that ignorance will be tested not in a laboratory, but on the world.

View Discussion

ABOUT THE AUTHOR

Thomas Wright

Thomas Wright is a senior fellow at the Brookings Institution. He served as the senior director for strategic planning at the National Security Council during the Biden administration.

Explore More Topics

Anthropic, People's Republic Of China, Pete Hegseth, United States Department Of Defense