TECHNOLOGY

OpenAI Is Opening the Door to Government Spying

Whether it means to or not

By Matteo Wong

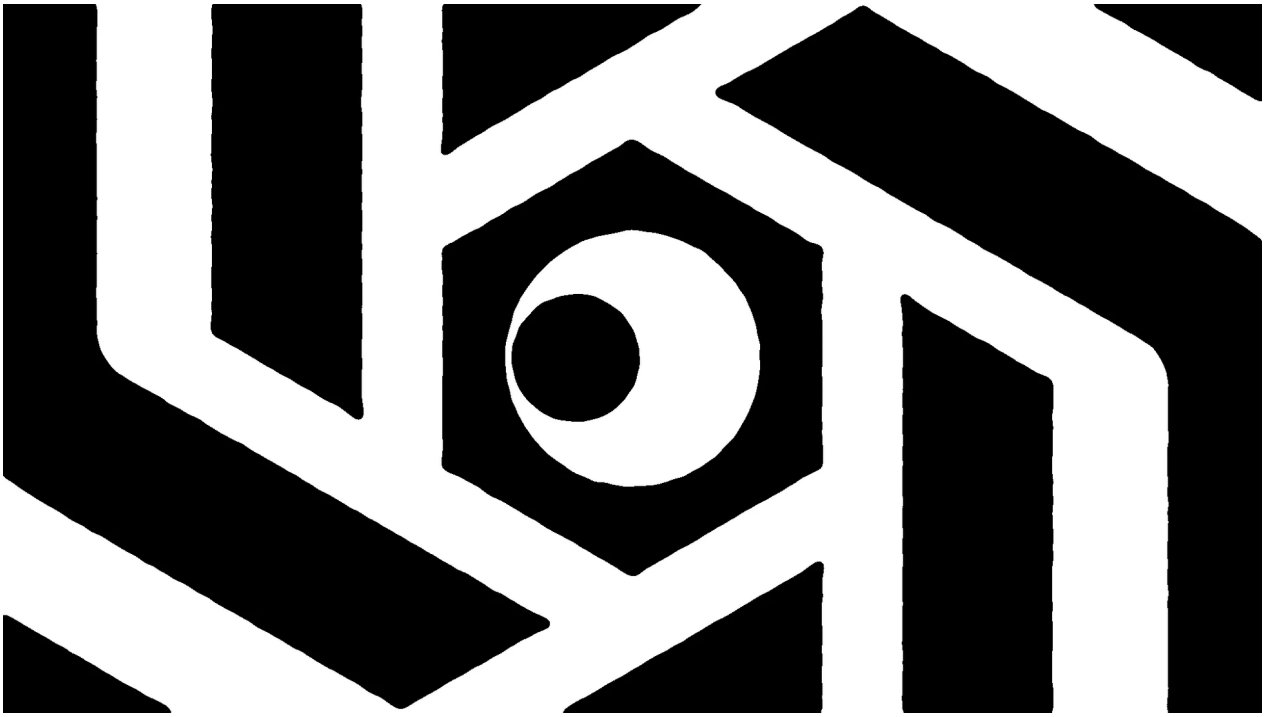

Illustration by Akshita Chandra / The Atlantic

MARCH 6, 2026 SHARE AS GIFT DISCUSS 9 REMOVE

Outside OpenAI's headquarters, a handful of people gathered on Monday holding pieces of colorful chalk. They got down on their knees and started writing messages on the sidewalk. STAND FOR LIBERTY. PLEASE NO LEGAL MASS SURVEILLANCE. CHANGE THE CONTRACT PLEASE.

At issue was a business deal that the company recently signed with the Department of Defense, following the Pentagon's sudden turn against Anthropic. OpenAI will now supply its technology to the military for use in classified settings, the sorts that may

involve wartime decisions and intelligence-gathering—an agreement, many legal experts told me, that could give the government wide-ranging powers. "I would just really like to see OpenAI do the right thing and stand up for something, anything," Niki Dupuis, an AI-start-up founder and one of the chalk protesters, told me.

In a widely leaked internal memo that Sam Altman sent last Thursday night, a copy of which I obtained, the OpenAI CEO said that he would seek "red lines" to prevent the Pentagon from using OpenAI products for mass domestic surveillance and autonomous lethal weapons. These were ostensibly the very same limits that Anthropic had demanded and that had infuriated the Pentagon, leading Defense Secretary Pete Hegseth to declare the company a supply-chain risk—a hefty sanction that would require anybody who sells to the Pentagon to stop using Anthropic products in their work with the military. Perhaps OpenAI was about to secure the very terms Anthropic had been denied.

But a close reading of the contract—the portions of it that OpenAI has shared with the public, anyway—indicates that the lines are, in fact, blurry. Several independent legal experts told me that, legally, the Pentagon can likely get away with using OpenAI's technology—versions of the models that underlie ChatGPT—for mass surveillance of Americans. Moreover, the military will likely have a pathway to use OpenAI's technology in autonomous weapons. AI models from Anthropic, DOD's previous partner, have likely already been used for warfare; recently, its products were reportedly used to identify targets in Iran (Anthropic declined to comment on that reporting). But the company had refused to allow its technology to be used in fully autonomous weapons.

Read: Inside Anthropic's killer-robot dispute with the Pentagon

The Department of Defense, which the Trump administration refers to as the Department of War, declined to answer my questions about the contract. A spokesperson for OpenAI reiterated to me that the Pentagon has agreed to not use the firm's AI system for domestic surveillance, but she did not answer specific questions. (OpenAI has a corporate partnership with *The Atlantic*'s business team.)

"The public is in an awkward position where we have to choose between trusting OpenAI or not," Charlie Bullock, a senior research fellow at the think tank Institute for Law & AI, told me. Brad Carson, who served as general counsel and then undersecretary of the Army under Barack Obama, was less compromising: In his analysis of the past week's events, OpenAI appears "okay with using ChatGPT for

what ordinary people think of as mass surveillance."

Over the past week or so, Altman and OpenAI have made several announcements about the contract, including sharing some of the text in a blog post last Saturday—only to modify that text in an update to the blog a few days later. The company's messaging has been confusing and has at various points seemed to contradict its own previous statements, as well as information from the government.

OpenAI had said that it has red lines around certain applications of its technology, but the portion of the contract language that it initially published implies the opposite. The company had also suggested that it placed unique restrictions on how the government could use OpenAI models, but Jeremy Lewin, a senior State Department official, suggested otherwise, writing that the contract simply permitted "all lawful use" of the OpenAI system—that is, anything technically legal. The messaging "at best makes them seem like they're not fully on top of this, and at worst reinforces the perception, fair or not, that OpenAI has a tendency to not be very candid," Alan Rozenshtein, a law professor at the University of Minnesota who studies emerging technology, told me. Rozenshtein was perhaps being diplomatic—the central question about OpenAI for the past several years has been less about candor and more about honesty. When Altman was briefly fired in late 2023, he had been accused of deceiving OpenAI's nonprofit board. A third-party review commissioned by OpenAI later found that there had been a "breakdown in trust" between Altman and the board but that Altman's "conduct did not mandate removal."

The past week has been chaotic, and observers have been hanging on every development. Last Friday, Altman posted on X that OpenAI had reached an agreement with DOD just hours after news broke that Anthropic's relationship with the administration would be dissolved. OpenAI's contract, Altman wrote, contains "prohibitions on domestic mass surveillance and human responsibility for the use of force, including for autonomous weapon systems." But many were skeptical. OpenAI surely had offered something to the Pentagon that Anthropic wouldn't. The word *prohibitions* didn't seem to communicate a total ban on surveillance, and the idea that "human responsibility" should be taken for autonomous weapons suggested that, indeed, OpenAI's technology could be used in autonomous weapons if a person were on the hook for the decision.

In Saturday's blog post, OpenAI insisted that its red lines against domestic

surveillance and automated weapons were firm, and it reframed the deal as an attempt to "de-escalate things" between the Pentagon and other U.S. AI labs, adding that it hoped the Pentagon would offer the same terms to other firms, including Anthropic. OpenAI also published a quote from the contract, though it offered little reassurance. The segment begins, "The Department of War may use the AI System for all lawful purposes, consistent with applicable law." It then says that the use of OpenAI systems for intelligence activities "will comply with" a number of laws and policies regulating U.S. intelligence activity that have infamously enabled spying on Americans, such as the Foreign Intelligence and Surveillance Act of 1978. Under FISA and related policies, for instance, intelligence agencies can record and store phone calls between Americans and people abroad, and purchase bulk user data from companies and analyze them, which does not involve directly intercepting communications.

Here I should note that it's impossible to use just snippets of a contract to evaluate the entire thing: A restriction in one section can be voided under circumstances listed in another. But snippets are all that OpenAI has provided. Based on what we are able to see, experts told me that leeway had likely been given for mass surveillance. "There's a ton of stuff that normal people would understand as automated mass surveillance that is simply not" illegal, Rozenshtein said. For example, generative AI could turn previously overwhelming and opaque records—tax returns, federal employment files, billions of intercepted communications, smartphone location data, and so on—into a trove of exacting insights. An OpenAI spokesperson told me that citing particular statutes in the contract does not change the agreed-upon prohibition against domestic surveillance.

With regard to weapons, the contract language shared last Saturday cites DOD Directive 3000.09, which does not prohibit the use of fully autonomous weapons. Actually, it provides a legal pathway to develop and deploy such weapons by outlining how they must be vetted and used. In sum, if an application is technically permitted under U.S. law, OpenAI would likely have to go along with it. And, of course, the Trump administration has argued for some very expansive interpretations of the law. "The original contractual language that OpenAI shared appeared to me to essentially be saying 'all lawful use,'" Bullock said.

After OpenAI published its blog post, Altman and some of his employees began fielding questions on X. *Did the contract allow NSA to use OpenAI products?* OpenAI's head of national-security partnerships insisted that the answer was no. *What about all*

of the loopholes for surveillance in existing laws? What about using AI to analyze bulk, commercially procured data, which DOD can purchase without a warrant? Multiple OpenAI employees voiced concerns about the deal as described. It was almost as if a contract for the military to use OpenAI's technology in weapons systems were being drafted live on social media, Jessica Tillipman, an expert on government-contracts law at George Washington University, told me.

Then, on Monday, OpenAI revised its blog post: The company said that it had modified its Pentagon contract to better protect Americans against AI-enabled spying. The new language notes that "the AI system shall not be intentionally used for domestic surveillance of U.S. persons and nationals" and that "for the avoidance of doubt," DOD "understands this limitation to prohibit deliberate tracking, surveillance, or monitoring of U.S. persons or nationals, including through the procurement or use of commercially acquired personal or identifiable information." In other words, OpenAI is making explicit that the terms of its contract should prevent its products from being used to spy on Americans en masse.

Read: Sam Altman is losing his grip on humanity

Outside legal experts told me that the update does seem meaningfully different from the original contract language and that it at least implies restrictions on the Pentagon that go above existing applicable law. But just as before, the new language could be construed to justify automated surveillance of Americans. For example, terms such as *intentionally* and *deliberate* provide substantial leeway for data collection that is deemed "incidental." Lots of commercially acquired data may not be deemed "personal or identifiable." Similarly, narrow definitions of terms such as *tracking* and *surveillance* could still permit a wide range of domestic intelligence-gathering, Carson, the former Army undersecretary, told me. "What ordinary people think surveillance might be in no way is the same as what surveillance means under the national-security authorities," he said. OpenAI did not provide definitions of these or any other terms in the contract when asked.

The update also states that the Pentagon "affirmed" that OpenAI technologies won't be used by intelligence agencies such as NSA without further negotiations—and OpenAI employees then suggested that the company may desire such partnerships in the future. And the phrase *U.S. persons and nationals* suggests that many immigrants, documented and not, may not be protected. OpenAI did not answer a question about whether undocumented immigrants and nonpermanent residents are protected by its

contract. To Carson, the modifications are "vaporous things that seem good"—window dressing without any substantive guarantees.

Of course, all of this discussion rests on the belief that contractual prohibitions are the load-bearing factor for preventing an AI system from being used for mass domestic surveillance or autonomous weapons. That is not necessarily true. A motivated lawyer could interpret almost any language in bad faith. If one takes OpenAI seriously—that the firm does not want its products used to spy on Americans at all—then enforcing the spirit of the contract may be more important than the document's language. (Lewin, the State Department official, said that "the government intends to honor the contract as written" and that using AI for mass domestic surveillance "has never been an object.")

To that end, OpenAI has shared that it will implement a technical "safety stack," or guardrails of a sort, to monitor how its models are used and that it will have its own engineers work with DOD, which the company believes will allow it to "independently verify that these red lines are not crossed." When asked, OpenAI did not provide further details about how its DOD safety architecture will work. The firm maintains that these guardrails and its contract, taken together, provide better guarantees "than earlier agreements, including Anthropic's." Once again, it all comes down to whether you trust OpenAI.

All of which leads to perhaps the most important and confounding factor of all: What happens if the government and OpenAI disagree over whether some use of ChatGPT is permitted? What does OpenAI do if it believes that the Pentagon has violated their agreement? Typically, the government acts first and litigates disputes after, Tillipman told me. (OpenAI said that if it determines that the terms of the contract have been violated, the company can terminate it, but it did not provide details about the process for doing so.)

And this was far from a typical negotiation. By blacklisting Anthropic, Tillipman said, DOD demonstrated that "if it comes to an impasse, they are not afraid" to place extreme sanctions on a private U.S. company. Altman wrote on X that designating Anthropic a supply-chain risk "is an extremely scary precedent and I wish [the government] handled it a different way." The actual red line should be very apparent to OpenAI and any other AI firm wanting to contract with DOD: You work on the government's terms, or not at all. OpenAI has made its choice.