

Amodeli says Anthropic has never advocated a ban on open-weights models

Source copy kept by wheresthe.ai for the story "Amodeli says Anthropic has never advocated a ban on open-weights models" (<https://wheresthe.ai/d/anthropic-open-weights-position-2026-07/>).

Source: <https://www.anthropic.com/news/position-open-weights-models>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 831a71d16d5284cb62c5a384150fdf40a6df3ce90aad9d247f5c48acf97cadd9

Documents the source offers: <https://www.dni.gov/files/ODNI/documents/assessments/ATA-2026-Unclassified-Report.pdf>

Announcements Our position on open-weights models Jul 27, 2026

A post by Dario Amodei, Anthropic CEO Over the last few days there has been a lot of discussion about open-weights models, especially those from China. Reports suggest that some US officials are considering banning the use of Chinese open-weights models by US companies. In response, many tech companies have signed a letter supporting open-weights models, and some people have even accused Anthropic of wanting to ban open-weights models as a means of protecting our business. Anyone who has read my past writing should know that I don't regard such bans as a useful measure, but let me state it clearly so that there is no doubt: Anthropic has never advocated for a ban on open-weights models. Open-weights models that don't have dangerous capabilities are a public good: they don't cost anything besides the compute needed to run them, and they provide value to businesses, developers, and researchers. Protectionist bans would not address my most serious national security concerns. Specifically, I am worried about two nightmare scenarios. I laid these out in my essay *The Adolescence of Technology* six months ago ¹, and have held these positions consistently for many years: My primary concern is the risk that authoritarian governments—not solely the Chinese Communist Party (CCP), although the CCP is clearly the most capable threat—build AI models that are more powerful than those built by the US, and use them to achieve permanent military superiority or perpetrate incredibly deep repression of their own people. This concern is widely shared within the US government: Vice President Vance warned in Paris last year that “authoritarian regimes have stolen and used AI to strengthen their military, intelligence, and surveillance capabilities,” and the Intelligence Community's 2026 Annual Threat Assessment found that “other global powers' robust progress in AI is challenging US economic competitiveness and national security advantages.” It is irrelevant whether these models are released with open weights, and certainly irrelevant whether they are used by US businesses. In fact, the most dangerous model may be one that is trained in secret and handed only to the People's Liberation Army for use in drones and the Ministry of State Security for surveillance and repression. My secondary concern is the risk that powerful AI models may be misused to carry out cyberattacks or biological attacks, and may have serious alignment problems. Open-weights models—it does not matter whether they come from China or anywhere else—do potentially present a higher risk than closed models, because it is very difficult to apply guardrails to them or monitor their usage, and once weights are released they cannot be withdrawn ². But banning the use of these models by US businesses does nothing to address this risk, because bad actors are unlikely to be legitimate US businesses. It would protect US AI companies from competition, but that has never been my goal. To address these concerns, I do support the following three measures, which I and Anthropic have consistently advocated for: We should not sell powerful chips or chipmaking equipment to China, and we should crack down on the rampant smuggling ³ and workarounds used to obtain access to such chips. China has limited domestic production capacity, and therefore, due to the scaling laws, cannot build more powerful models than the US without US chips. This is the most efficient and direct way to block threat #1, and by hampering the training of models that are out of reach of US law, it also indirectly helps with threat #2. We should crack down on industrial-scale distillation operations. Distillation is a much more compute-efficient process than training models from scratch. It allows China to build much better models than its number of chips would ordinarily enable, and thus partially evade chip bans. Distillation does not allow the CCP to obtain equivalent or superior AI capabilities to the US, but it can bring the Chinese frontier to within a few months of the US frontier. It is true that many of the companies carrying out these operations release open-weights models—but the open weights are far less relevant than the fact that the operations are backed by an authoritarian state seeking to overtake the US at the frontier. We should have policy interventions to deter this behavior. A blanket ban on open-weights models is neither the correct remedy nor something we have called for ⁴. All sufficiently capable models, open and closed, should go through mandatory safety testing. The best way to address threat #2 is to just directly test models for cyber, biological, and alignment risks before release. I think this idea is actually close to a consensus: I have been

heartened both that the Trump administration has moved in this direction in recent months, and by recent industry proposals that would apply such testing to the most capable models regardless of their country of origin or whether they are open or closed (while exempting less capable models, such as those from startups and academia, entirely). Whether open models do or don't pose an increased risk, and whether that risk can be mitigated, is something that should emerge from testing, rather than be decided in advance—and there may be promising methods for improving the safety of open-weights models, including recent research from AE Studio and Anthropic on modular training strategies . Note that to be effective, testing would need to be global, which means even the CCP would need to be on board. I think this may actually be possible: as I wrote in *The Adolescence of Technology* , limited cooperation around preventing AI biological weapons may be possible because it is in China's interest too. This brings me to the open letter . I agree with much of it: open weights expand access to the AI economy, they strengthen competition at least for some use cases, and they give customers greater control. Concerns about distillation should be addressed through targeted legal and commercial frameworks—the same measure I described above. But I don't agree with the letter's assertions that open-weights models necessarily make it easier to develop safeguards or that broad access to capabilities necessarily helps defenders more than attackers. It seems at least as likely to me that the opposite will be true. For example, I worry that biology will have a strong attacker-defender asymmetry, where sufficiently capable models may be able to quickly weaponize pandemic-level viruses with widely available materials, whereas defense against these agents is a multi-year operational task in the best case (as we saw with Operation Warp Speed) 5 . Questions like this should be empirically answered by rigorous pre-release testing, not assumed in advance. To summarize my and Anthropic's position, we have not and are not advocating for a ban on open-weights models as a category. We should instead focus on keeping powerful chips out of authoritarian hands, stopping industrial-scale distillation, and requiring safety testing of all sufficiently capable models, open and closed. *Edit 28 July: Updated to note that the cited research on modular training strategies was a collaboration between Anthropic and AE Studio .

Footnotes See Sections 3 and 2 of that essay for discussion of misuse for seizing power and discussion of biological risks, respectively. See this report from the UK AI Security Institute, specifically: “The same openness underpinning these benefits precludes many of the safety measures that closed model developers can use to detect and disrupt misuse, iterate on safeguards as vulnerabilities emerge, control user access and withdraw models. Once open-weight models are released, these options are lost permanently: safeguards can be removed, and copies can be downloaded, redistributed, and run on private systems beyond monitoring. For models with dangerous capabilities – including highly cyber-capable models – open weight release therefore creates a persistent and irreversible risk of misuse.” See also here , here , and here for more reports from the US Department of Justice. At Anthropic we're committed to cracking down on industrial-scale distillation through our own practices, including identifying and banning accounts that use our models in this way. This is challenging—for instance, the relevant accounts can often only be identified after substantial distillation has occurred, and distillation often involves creating large numbers of fake accounts that form a moving target. The practices of any individual company cannot entirely solve the problem, which is why we have called for policy on this issue. See Section 2 of *The Adolescence of Technology* for a more detailed discussion of biological threats and the offense-defense balance. To summarize, what I believe currently keeps us safe in biology is not “defenders”, or even the availability of materials, but a negative correlation between intellectual capability and desire to commit catastrophic harm. Previous technologies like internet search or even DNA synthesis were nowhere near powerful enough to break this correlation, but I worry that at its current rate of progress, AI will do so very soon. Another way to say it is that a sufficiently powerful technology removes all barriers and exposes whether the attacker or defender has an inherent structural advantage, and I worry in biology it is the attacker.

Related content

An update on independent evaluators To evaluate the safety of our systems, Anthropic is giving employee-like access to a team from METR, a nonprofit that evaluates frontier AI models, and we're in early discussions with several other organizations, including Hugging Face. Read more Preparing for the 2026 US midterms Earlier this year, we shared an update on our election safeguards. Here, we outline how we've continued to strengthen these safeguards ahead of this year's US midterms. Read more Anthropic opens Singapore office and signs MOU with the Government of Singapore Anthropic has opened an office in Singapore and signed an MOU with the country's government to advance safe, responsible AI adoption. Read more