MICHAEL J. MONGAN (SBN 250374)
michael.mongan@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
50 California Street, Suite 3600
San Francisco, CA 94111
Telephone: (628) 235-1000

EMILY BARNET (*pro hac vice*)
emily.barnet@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
7 World Trade Center
250 Greenwich Street
New York, NY 10007
Telephone: (212) 230-8800

Attorneys for Plaintiff Anthropic PBC

KELLY P. DUNBAR (*pro hac vice*)
kelly.dunbar@wilmerhale.com
JOSHUA A. GELTZER (*pro hac vice*)
joshua.geltzer@wilmerhale.com
KEVIN M. LAMB (*pro hac vice*)
kevin.lamb@wilmerhale.com
SUSAN HENNESSEY (*pro hac vice*)
susan.hennessey@wilmerhale.com
LAUREN MOXLEY BEATTY (SBN 308333)
lauren.beatty@wilmerhale.com
MATTHEW E. MORRIS (*pro hac vice*)
matt.morris@wilmerhale.com
BARDIA VASEGHI (*pro hac vice*)
bardia.vaseghi@wilmerhale.com
DAKOTA C. FOSTER (*pro hac vice*)
dakota.foster@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ANTHROPIC PBC,

Plaintiff,

v.

U.S. DEPARTMENT OF WAR, et al.,

Defendants.

Case No. 3:26-cv-01996-RFL

ANTHROPIC PBC'S MOTION FOR SUMMARY JUDGMENT

Judge: Hon. Rita F. Lin
Hearing Date: July 30, 2026
Time: 10:00 a.m. PT

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Page(s)

CASES

STATUTES, RULES, AND REGULATIONS

OTHER AUTHORITIES

NOTICE OF MOTION AND MOTION

PLEASE TAKE NOTICE that on July 30, 2026 at 10:00 a.m. PT, Plaintiff Anthropic PBC will and hereby does move the Court for an order granting Anthropic's Motion for Summary Judgment. Anthropic's motion is based on this Notice of Motion, the supporting Memorandum, and the Declarations filed herewith. Anthropic requests that the Court grant Anthropic summary judgment on each of its claims.

MEMORANDUM OF POINTS AND AUTHORITIES

INTRODUCTION

The undisputed facts establish that Defendants' campaign of retaliation against Anthropic violated the Constitution and the Administrative Procedure Act.

Defendants have been remarkably transparent about what happened here. They view Anthropic as providing the best frontier AI model and engaged the company to help the Department of War and other agencies protect America's national security. They acknowledge that Anthropic has always made its models available subject to certain use restrictions, consistent with its longstanding advocacy for safe and responsible use of AI. In February 2026, Defendants insisted that Anthropic abandon two contractual restrictions that prohibited them from using the model for lethal autonomous warfare or mass surveillance of Americans. When Anthropic respectfully declined to abandon those restrictions, and publicly emphasized their importance, Defendants abruptly decided to punish the company. They barred every federal agency and every military contractor from doing business with Anthropic and designated the company a threat to national security.

Everyone agrees that the Executive Branch can lawfully decide to stop contracting with a vendor that declines to accept its preferred contract terms. But the Executive Branch remains subject to the requirements of the Constitution and federal law. It may not retaliate against an American company in violation of the First Amendment; it may not disregard procedural and substantive requirements imposed by Congress; it may not exercise its statutory national-security authorities in a manner that is arbitrary and capricious; and it may not punish a company without providing notice and an opportunity to respond.

As the undisputed record shows, Defendants did each of those things here. Their campaign of retaliation commenced with statements from the President and the Secretary of War that acknowledged—on their face—that the Executive was taking action against Anthropic for its "rhetoric" and its "ideology." The Secretary designated Anthropic a "supply chain risk" under 10 U.S.C. § 3252 without taking the procedural steps required by Congress and based on rationales that exceed the statutory definition of a supply chain risk—and that the record shows to be arbitrary, false, and pretextual. Each of the unprecedented sanctions imposed by Defendants occurred without any pre-deprivation process, in violation of the Due Process Clause. On this record, Anthropic is entitled to judgment as a matter of law.

BACKGROUND

A. Anthropic And Its Mission To Responsibly Develop AI Tools

Anthropic is a frontier AI company founded in 2021. It offers an industry-leading large language model (LLM), known as Claude, which is capable of interpreting and responding to a wide range of user inputs in an intelligent, human-like manner. Kaplan Decl. ¶ 16.[1] LLMs are algorithmic systems that acquire predictive capabilities, allowing them to perform complex tasks in a fraction of the time it would take humans. *Id.* ¶ 15.

Anthropic was founded on the animating principle that the most capable AI systems should also be the safest. *Id.* ¶¶ 8-11. That principle is embedded in Anthropic's corporate charter and governance. *Id.* ¶ 8. As a public benefit corporation, Anthropic balances stockholder interests with its public benefit purpose of responsibly developing and maintaining advanced AI tools for the long-term benefit of humanity. *See id.* ¶¶ 8-9.

B. Anthropic's Partnership With The Federal Government

Anthropic began commercial deployment of Claude in 2023, *id.* ¶ 12, and started working with the federal government the following year, Ramasamy Decl. ¶ 7. In November 2024, Anthropic began working with a defense technology partner to provide AI-enabled capabilities to U.S. intelligence and defense agencies. *Id.* ¶ 8. On its own initiative and at its own expense,

[1] All declarations cited in this motion are attached and have been filed concurrently.

Anthropic built a new version of Claude (known as "Claude Gov") that was tailored to national-security needs and classified environments. *Id*. ¶¶ 31-32.

Every Anthropic model is subject to a Usage Policy that prohibits certain uses of Claude. *Id*. ¶ 25. Claude Gov is subject to a government-specific addendum to the Usage Policy that removes certain restrictions for national security customers that would otherwise apply to civilian users. Kaplan Decl. ¶¶ 27-28. From the outset of the Department's partnership with Anthropic, the Department has operated subject to the Usage Policy and a government-specific addendum. *Id.* ¶¶ 28-29; Ramasamy Decl. ¶ 26.

Since March 2025, the Department has accessed Claude Gov through classified cloud environments; entrusted Anthropic to handle classified national security data; and awarded the company a Top Secret facility clearance. Ramasamy Decl. ¶¶ 33, 58. In the summer of 2025, the Department awarded Anthropic a two-year contract worth up to $200 million. *Id.* ¶¶ 9, 69. Until the events that gave rise to this litigation, Claude had reportedly become the Department's most widely deployed frontier AI model. Kaplan Decl. ¶ 26. According to the Department, Claude was "far and away [its] best model." Heck Decl. ¶ 9.

Other agencies have also made extensive use of Claude and praised it. In 2025, the General Services Administration (GSA) announced a first-of-its-kind agreement with Anthropic to deliver Claude across all three branches of government for just $1 per agency. Ramasamy Decl. ¶ 71. And the General Counsel of the Department of Veterans Affairs, for example, called Claude and related tools "incredible" and pledged to "make full use of" them. AR 537.

C. The Present Dispute

In fall 2025, the Department and Anthropic began discussing a new deployment of Claude on the Department's "GenAI.mil" platform. Kaplan Decl. ¶ 32. For the first time, the Department demanded that Anthropic discard its Usage Policy and agree to a new contractual provision allowing the Department to use Claude "for all lawful uses." *Id*. Anthropic agreed to accommodate the Department by significantly revising its usage restrictions. *Id*. ¶ 33; AR 7. But it respectfully declined to discard two narrow restrictions grounded in its founding commitments: it would not authorize use of Claude for lethal autonomous warfare or mass surveillance of

Americans. Kaplan Decl. ¶ 33; AR 1-2. Anthropic's CEO told the Under Secretary of War for Research and Engineering that the company would assist in offboarding Claude from the Department's systems if it concluded Anthropic was not the right vendor for its needs. AR 2, 29.

On February 24, Anthropic's leadership met with Secretary Hegseth and his team at the Pentagon. Heck Decl. ¶ 13. The Secretary did not say anything about Anthropic's AI models being unsafe, insecure, or subject to compromise. *Id*. ¶ 18. To the contrary, he praised Claude's "exquisite capabilities" and told Anthropic that the Department "would love to work with" the company. *Id*. ¶¶ 13, 16. But he insisted on Anthropic removing its two remaining restrictions. *Id*. ¶¶ 13, 17. And he issued an ultimatum: Unless Anthropic agreed to the "all lawful uses" contract term by 5 p.m. ET on February 27, he would either designate the company as a "supply chain risk" *or* invoke the Defense Production Act to compel Anthropic to comply with his demand. *Id*. ¶ 17. On the eve of that deadline, Anthropic's CEO publicly explained why Anthropic could not "in good conscience" abandon its commitments to AI safety. *See* Mongan Decl. Ex. 1 at 1-2.

The next day, before the Secretary's 5 p.m. deadline, President Trump issued a statement on social media (the "Presidential Directive"). He branded Anthropic a "Radical Left AI company" that threatened "AMERICAN LIVES" and "direct[ed] EVERY Federal Agency in the United States Government to IMMEDIATELY CEASE all use of Anthropic's technology." AR 255A. Federal agencies moved quickly to carry out the Presidential Directive. That same day, GSA removed Anthropic from the government's primary procurement marketplace, AR 394-95, and the Department of Health and Human Services disabled access to Claude, AR 335. Within days, the Departments of State, Treasury, Veterans Affairs, and other agencies announced that they would terminate or suspend use of Anthropic's products. Each cited only the Presidential Directive. *See* AR 321 (State); AR 256 (Treasury); AR 527 (VA); AR 317 (FHFA).

Less than two hours after the Presidential Directive, Secretary Hegseth issued a "final" decision (the "Hegseth Directive") via social media. AR 255B. He accused Anthropic of "betrayal"; declared its conduct "fundamentally incompatible with American principles"; "direct[ed] the Department of War to designate Anthropic a Supply-Chain Risk to National Security"; ordered that, "[e]ffective immediately, no contractor, supplier, or partner that does

business with the United States military may conduct any commercial activity with Anthropic" (the "secondary boycott" order); and instructed that Anthropic would continue to provide its technology to the Department for up to six months. AR 255B.

Later, the Secretary signed a determination dated March 3 (the "March Determination") that formalized his designation of Anthropic as a supply chain risk under 10 U.S.C. § 3252. AR 209. But the Department did not reveal that determination to Anthropic on March 3. Instead, it continued negotiations with Anthropic about the disputed contract terms. Heck Decl. ¶¶ 20-22. Those negotiations extended through the morning of March 4, when Under Secretary Michael sent Anthropic a revised counteroffer and wrote: "I think we are very close here." *Id.* Ex. 1; *see also* Dkt. 164 ¶ 104 (admitting that negotiations extended through the evening of March 4). That evening, however, Anthropic received a letter (dated March 3) notifying it of the March Determination. AR 237-38; Heck Decl. ¶ 21. The Department did not disclose the text of that Determination—or any materials purporting to justify it—until March 17. *See* Dkt. 96-2.[2]

On March 26, this Court preliminarily enjoined Defendants' implementation of the Presidential Directive, the Hegseth Directive, and the March Determination (together, the "Challenged Actions"). Dkt. 134 ("Op."). The Court restored the pre-February 27 status quo, while also expressly allowing Defendants to take any lawful step to halt their work with Anthropic and its AI tools. Dkt. 135 at 3. In the months since the Challenged Actions, the Department and other federal agencies have continued using those tools. Ramasamy Decl. ¶¶ 78-81. A senior official also publicly acknowledged that the Department has used Claude in active military operations in Iran. Mongan Decl. Ex. 5 at 28. And before making a new and powerful model named Mythos more broadly available, Anthropic briefed senior officials across government agencies on its capabilities and made it available to government partners. Ramasamy Decl. ¶¶ 80-81; Mongan Decl. Ex. 3.

[2] The Department sent Anthropic another letter on March 4 notifying it of a separate determination by the Secretary under 41 U.S.C. § 4713. Anthropic is currently challenging that action in a separate proceeding in the D.C. Circuit, as required by 41 U.S.C. § 1327(b)(3).

ARGUMENT

This Court should grant summary judgment in favor of Anthropic on its constitutional claims because "there is no genuine issue as to any material fact," and Anthropic "is entitled to judgment as a matter of law." *Anderson v. Liberty Lobby, Inc.*, 477 U.S. 242, 250 (1986). As to Anthropic's APA claims, this Court reviews the challenged final agency action pursuant to 5 U.S.C. § 706. *Compassion Over Killing v. FDA*, 849 F.3d 849, 854 (9th Cir. 2017). Anthropic should prevail on its APA claims whether they are viewed in light of the proffered administrative record, *see* Dkt. 156-1, or with the benefit of additional documents that the Court may consider under these circumstances, *see infra* pp. 20-21. Either way, the actions are invalid.

I. Anthropic Is Entitled To Summary Judgment On Its First Amendment Claim Because Defendants Retaliated Against Anthropic For Its Protected Expression

To prevail on its First Amendment claim, Anthropic must establish "that (1) it engaged in constitutionally protected activity; (2) the defendant's actions would 'chill a person of ordinary firmness' from continuing to engage in the protected activity; and (3) the protected activity was a substantial motivating factor in the defendant's conduct." *Ariz. Students' Ass'n v. Ariz. Bd. of Regents*, 824 F.3d 858, 867 (9th Cir. 2016). Once Anthropic has made that showing, "the burden shifts to the defendant official to demonstrate that even without the impetus to retaliate he would have taken the action complained of." *Hartman v. Moore*, 547 U.S. 250, 260 (2006).

The undisputed facts here establish a "classic" case of "illegal First Amendment retaliation." Op. 2. This Court has already concluded that Anthropic "engaged in protected activity." *Id.* at 20. Anthropic has been "a loud and influential voice regarding the capabilities, risks, and safe uses of AI technology." *Id.*; *see also* Kaplan Decl. ¶¶ 8-10, 13; Heck Decl. ¶¶ 5-6; Mongan Decl. Exs. 1, 4. Defendants have introduced no contrary evidence, and they have never disputed that the Challenged Actions would chill a person of ordinary firmness from engaging in that protected activity. Op. 21. And the very words used by President Trump, Secretary Hegseth, and Under Secretary Michael in announcing and justifying those actions leave no question that Anthropic's protected activity substantially motivated them. In light of those words—and other

undisputed evidence—Anthropic has carried its burden on its retaliation claim, and Defendants cannot show they would have taken the Challenged Actions absent the retaliatory impetus.

A. The Undisputed Facts Establish A Prima Facie Case Of Retaliation

1. Anthropic Engaged In Constitutionally Protected Activity

The evidence shows that Anthropic engaged in activity protected by the First Amendment. Since its founding, Anthropic has been a leading voice in the ongoing public debate about AI safety. It has worked to advance AI research, safety, and policy, *see* Kaplan Decl. ¶¶ 9-10, 13, including by publishing essays and public statements about the risks of deploying AI without proper restrictions and safeguards, *see id*. ¶¶ 10, 13; Mongan Decl. Exs. 1, 4. Its CEO has been a particularly prominent advocate for AI safety, including through a January 2026 public essay calling for "guardrails" on the development of powerful AI to address its risks to public safety and security. Mongan Decl. Ex. 4 at 21, 22, 30-32, 44. That speech on matters of profound "'public concern'" lies at the core of what the First Amendment protects. *Snyder v. Phelps*, 562 U.S. 443, 451-52 (2011).

The undisputed facts demonstrate that Anthropic's interactions with the Department are also protected First Amendment activity. When summoned to Washington on February 24 and pressed by the Secretary to abandon Anthropic's longstanding restrictions against the use of Claude for mass surveillance of Americans or lethal autonomous warfare, Anthropic's leadership explained to the government the importance of those restrictions and respectfully declined to abandon them. Heck Decl. ¶¶ 13-20. Anthropic was addressing "governmental conduct" regarding the safe use of AI—a subject in which the public has a "deep and abiding interest" because it "affects the societal interest as a whole." *Alpha Energy Savers, Inc. v. Hansen*, 381 F.3d 917, 926-27 (9th Cir. 2004). The communications "touche[d] on fundamental questions" of "policy." *Janus v. Am. Fed'n of State, Cty., & Mun. Emps., Council 31*, 585 U.S. 878, 913 (2018). And they implicated the "'ethical [and] moral'" principles that define Anthropic's public-benefit mission. *Burwell v. Hobby Lobby Stores, Inc.,* 573 U.S. 682, 701 (2014). The speech was no less protected just because Anthropic expressed itself directly to the government: The Free Speech Clause protects speech on "matters of substantial public concern" during contractual bargaining with the

government. *See Janus*, 585 U.S. at 885-86. And the Petition Clause independently protects efforts to advocate before "all departments of the Government." *Cal. Motor Transp. Co. v. Trucking Unlimited*, 404 U.S. 508, 510 (1972).

The First Amendment likewise protects Anthropic's public communications about the dispute. On February 26, Anthropic's CEO issued a statement explaining that Anthropic could not "in good conscience" abandon its two restrictions because certain uses of AI fall "outside the bounds of what today's technology can safely and reliably do." Mongan Decl. Ex. 1 at 1-2. That public statement on a question of pressing national importance—delivered when both sides had "publicly staked out their positions and engaged in a public debate," Op. 20—is precisely the kind of contribution to "uninhibited, robust, and wide-open" debate that the First Amendment protects, *New York Times Co. v. Sullivan*, 376 U.S. 254, 270 (1964). Indeed, it "'occupies the highest rung of the hierarchy of First Amendment values.'" *Snyder*, 562 U.S. at 452.

2. Defendants' Actions Would Chill A Person Of Ordinary Firmness

The Challenged Actions "easily qualify as ones which would chill a person of ordinary firmness from continuing to engage in further protected speech." Op. 21. Defendants have never contested that issue. *Id.* Nor could they. The Presidential Directive categorically bars Anthropic from working with all federal agencies, AR 255A; the secondary boycott order prohibits any "contractor, supplier, or partner that does business with the United States military" from engaging in any "commercial activity" with Anthropic, AR 255B; and the supply chain risk designation labels Anthropic a national-security threat and prevents it from competing for certain Department contracts, including as a subcontractor, AR 209. Together, the Challenged Actions "chill public debate" (Op. 21) by sending an unmistakable warning to companies in the industry: speak out and you too may be blacklisted, boycotted, and branded a national-security risk.

3. Anthropic's Speech Substantially Motivated The Challenged Actions

The record does not allow for any genuine dispute that Anthropic's protected expression was a substantial motivating factor for the Challenged Actions. Defendants' own words establish that they acted to punish Anthropic for its public stance on whether its models could safely be used in the two contexts at issue: Secretary Hegseth tied the supply chain risk designation and the

secondary boycott to Anthropic's "sanctimonious rhetoric" and its purported attempt to "strong-arm the United States military" to accept its "Silicon Valley ideology." AR 255B. President Trump similarly explained his Directive by describing Anthropic as a "RADICAL LEFT, WOKE COMPANY" staffed by "Leftwing nut jobs" who made a "DISASTROUS MISTAKE trying to STRONG-ARM the Department of War." AR 255A. And Under Secretary Michael doubled down in his "Risk Analysis." AR 213. He asserted that Anthropic's "risk level" supposedly "escalated" when the company "began engaging in an increasingly hostile manner through the press." AR 215. He dismissed Anthropic's position as "brand-building" and "marketing," characterized the dispute as a "public[] spat," and accused Anthropic of "leveraging" the parties' "negotiations for Anthropic's own public relations." AR 214-15.

These statements expressly take issue with Anthropic's decision to participate in "free discussion of governmental affairs"—precisely what the First Amendment was designed to protect by "practically universal agreement." *Mills v. Alabama*, 384 U.S. 214, 218 (1966). When direct "evidence that the defendants expressed opposition to" Anthropic's speech appears in the same statements that announced and justified the Challenged Actions, that is powerful proof that "retaliation was a substantial or motivating factor." *Alpha Energy Savers,* 381 F.3d at 929.

Other factors that the Ninth Circuit considers point decisively in the same direction. The "proximity in time between [the] expressive conduct and the allegedly retaliatory action," *id.*, confirms that the conduct motivated the actions. President Trump and Secretary Hegseth issued their Directives fewer than 24 hours after Anthropic's CEO publicly defended its usage restrictions and declined to abandon them. Mongan Decl. Exs. 1, 7-8. If "three to eight months is easily within a time range that can support an inference of retaliation," *Coszalter v. City of Salem*, 320 F.3d 968, 977 (9th Cir. 2003), a single day certainly is.

And the undisputed facts also show that Defendants' "proffered explanations for their adverse actions were false or pretextual." *Alpha Energy Savers,* 381 F.3d at 929. To the extent the February 27 Directives offer any rationale beyond invective and criticism of First Amendment activity, they suggest a concern that Anthropic will "veto" the "operational decisions of the United States military." AR 255B; AR 255A. The Michael memorandum likewise asserts that Anthropic

possesses an "operational veto" over Claude after it is deployed in the Department's classified systems. AR 213. But none of those documents explains why Defendants "believed Anthropic has 'privileged access,' 'backdoors,' or can otherwise 'control' Claude after its delivery to DoW." Op. 14; *see also* Ramasamy Decl. ¶¶ 48-58.

The record directly contradicts the government's after-the-fact rationales. Once deployed in secure Department environments, Anthropic has no access to its models, no visibility into their use, and no technical ability to alter or disable them. Ramasamy Decl. ¶¶ 48-58. It cannot push changes, introduce vulnerabilities, or interfere with operations. *Id*. It cannot veto Department uses—and it has expressly disavowed any desire to interfere with military operations, including by offering a contract term expressly disavowing any authority to interfere with ongoing operations. Heck Decl. ¶ 25; *see* Ramasamy Decl. ¶¶ 48-58. Bare assertions of theoretical risk made "without evidence in the record" cannot defeat summary judgment. *White v. Lee*, 227 F.3d 1214, 1241 (9th Cir. 2000). That is especially true where (as here) those assertions are "blatantly contradicted by the record," such that "no reasonable jury could believe" them. *Scott v. Harris*, 550 U.S. 372, 380 (2007). Far from creating a triable issue, the inaccurate and pretextual nature of Defendants' chosen explanations confirms what is already apparent: their actions were substantially motivated by a desire to punish Anthropic for its protected constitutional activities.

B. Defendants Cannot Rebut The Prima Facie Case

Defendants can "'escape liability only'" by proving that they "'would have reached the same decision . . . even in the absence of'" that speech. *Thomas v. County of Riverside*, 763 F.3d 1167, 1169 (9th Cir. 2014). It is not enough for Defendants to show that they "could have" taken the Challenged Actions in response to the contracting impasse. *CarePartners, LLC v. Lashway*, 545 F.3d 867, 877 (9th Cir. 2008). The question is whether, without Anthropic's protected speech, Defendants *would have* taken these same sweeping actions anyway. Op. 23. The undisputed facts show they would not have.

The Presidential and Hegseth Directives identify the reason for the Challenged Actions on their face. Both link the government's response to Anthropic's public position, which the Directives characterize as an ideologically objectionable effort to "strong-arm" the government.

See AR 255A, 255B; *supra* p. 9. Where the government's own documents "admit[] that they took the action directly in response to" protected speech, "[t]here is no genuine issue of disputed fact" that the same action would not have occurred without that speech. *Riley's Am. Heritage Farms v. Elsasser*, 32 F.4th 707, 728 (9th Cir. 2022). That alone should end the inquiry.

And nothing in the record supports the "determination that the contracting impasse alone would have triggered such an extreme response" absent Anthropic's protected speech. Op. 24; *see* Dkt. 96 at 23; AR 213-16. If Defendants' actual concern had been Anthropic's refusal to accept the Department's preferred contract terms, they had methods to address that concern directly. Defendants could have stopped using Claude; selected another vendor for the GenAI.mil contract; or used ordinary procurement authorities to terminate, narrow, or wind down particular Anthropic contracts, *see, e.g.*, 48 C.F.R. §§ 49.101(b), 49.402-3, 52.212-4(l), 52.212-4(m), 52.243-4(a), 52.249-1-52.249-6. Even absent the Directives, which lay bare Defendants' retaliatory purpose, a dispute over contract terms would not explain a government-wide blacklist, a sweeping secondary boycott, or a first-of-its-kind supply chain risk designation.

Recent events underscore that Defendants took the Challenged Actions in response to Anthropic's protected speech, not genuine national-security concerns. Even after the supply chain risk designation, a Department official testified that the agency used Claude to support military efforts in Iran. Mongan Decl. Ex. 5 at 28. Anthropic has continued to work closely with the federal government, including by proactively extending access to its new Mythos model and meeting with senior White House officials in April to discuss its responsible use—discussions the White House described as "productive and constructive." Heck Decl. ¶ 27; Mongan Decl. Ex. 7 at 1. These are not the actions of parties genuinely concerned that Anthropic poses a national-security risk.

Finally, *American Federation of Government Employees, AFL-CIO v. Trump*, 167 F.4th 1247 (9th Cir. 2026) (*AFGE*), does not help Defendants. *See* Op. 23 n.12. The government would have issued the challenged executive order in that case even absent the asserted retaliatory motive because the order itself contained no retaliatory language, it had "'legitimate grounding in national security concerns,'" and the record showed it would have issued one way or the other. *AFGE*, 167

F.4th at 1256-57. This case is the opposite: The Presidential and Hegseth Directives tied the Challenged Actions to Anthropic's speech activities. And there is no record-supported national-security rationale that would independently explain the Challenged Actions. *See infra* pp. 17-21.

II. Secretary Hegseth's Actions Violate The APA Because They Are Arbitrary, Capricious, And Contrary To § 3252's Procedural And Substantive Requirements

On February 27, Secretary Hegseth issued a "final" decision, "[e]ffective immediately." AR 255B. That Directive contained two unprecedented orders: The Secretary ordered a secondary boycott, prohibiting all military contractors from conducting business with Anthropic; and he "direct[ed] the Department of War to designate Anthropic a Supply-Chain Risk to National Security." AR 255B. Both orders violate the APA. Defendants have "concede[d]" that the secondary boycott order "was in excess of and contrary to [the Secretary's] statutory authority." Op. 35. And the supply chain risk designation is procedurally and substantively flawed. That is true regardless of whether the Court views the relevant agency action as the Secretary's February 27 decision or his subsequent March Determination formalizing the designation, and regardless of whether the Court reviews the action exclusively in light of the administrative record proffered by Defendants or with the benefit of additional materials it may properly consider. *See infra* pp. 20-21.

A. Secretary Hegseth's February 27 Directive Is Final And Unlawful Agency Action

To date, Defendants have offered no defense of the actions announced in Secretary Hegseth's February 27 Directive. They instead contend that the Directive is unreviewable because it "is *not* a final agency action subject to APA review." Dkt. 96 at 25. As this Court previously concluded, that is wrong. Op. 35. The Directive reflects the "'consummation' of the agency's decisionmaking process." *Bennett v. Spear*, 520 U.S. 154, 178 (1997). It states unequivocally that "[t]his decision is final" and "[e]ffective immediately." AR 255B. That language conveys a "definitive decision," not a "tentative or interlocutory" one. *U.S. Army Corps of Eng'rs v. Hawkes Co.*, 578 U.S. 590, 597-98 (2016). For the same reason, the Directive imposes "legal consequences" and determines "rights" and "obligations." *Bennett*, 520 U.S. at 178; *see* Op. 35. It

therefore satisfies the APA's finality requirement.

Neither action announced in the February 27 Directive can survive APA review. At the preliminary injunction hearing, "Defendants conceded . . . that there is no statutory basis for" the secondary boycott order. Op. 34. As to Secretary Hegseth's order that the Department designate Anthropic a supply chain risk, Defendants have submitted no evidence that the Secretary complied with the statutory prerequisites of § 3252 before issuing it. *See id.* The Directive itself contains scarcely any reasoning—apart from criticizing Anthropic's "rhetoric" and "ideology," AR 255B—and lacks any genuine support from contemporaneous record evidence. The administrative record materials proffered by Defendants almost entirely post-date the Directive. The lone exception is a due-diligence report from a third-party vendor dated February 26, 2026. AR 217-29. It assigns Anthropic an overall risk of "medium," based in substantial part on private copyright litigation against the company and the fact that "the US Secretary of Defense threatened to label the company as a 'supply chain risk.'" AR 219-20. That circular reasoning "does not corroborate the risks that purportedly resulted in the supply chain risk designation." Op. 36 n.18.

The Secretary's March Determination merely formalized and carried out his earlier, "final" decision "directing the Department of War to designate Anthropic a Supply-Chain Risk to National Security." AR 255B. The Secretary did not "'deal with the problem afresh'" on March 3, as would be required for "*new* agency action." *DHS v. Regents of the Univ. of Cal.*, 591 U.S. 1, 21 (2020). Nor did he "withdraw[] or rescind[]" the earlier Directive. Op. 35. As a result, the prior Directive is the operative agency action—and the supply chain risk designation must be evaluated based on the scant reasoning in that Directive and the record before the Secretary when he issued it on February 27. *See Regents*, 591 U.S. at 20. Because the original Directive cannot be sustained, the March Determination cannot stand.

B. The March Determination Is Independently Unlawful

Even crediting Defendants' position that the March Determination is the relevant final agency action, it cannot survive APA review. The action violates clear-cut procedural requirements in 10 U.S.C. § 3252; stretches the substantive authority granted in § 3252 beyond the breaking point; and rests on arbitrary, incorrect, and contrived rationales.

1. The § 3252 Designation Violates Procedures Required By Law

To start, Secretary Hegseth failed to comply with mandatory procedural safeguards. *See* 5 U.S.C. § 706(2)(D). Before making a supply chain risk designation, the Secretary must determine in writing that "less intrusive measures" could not "reduce" the identified risk, 10 U.S.C. § 3252(b)(2)(B), and then provide Congress "a discussion of less intrusive measures that were considered and why they were not reasonably available," *id.* § 3252(b)(3)(B). He also must determine in writing that the designation is "necessary to protect national security by reducing supply chain risk," *id.* § 3252(b)(2)(A), and provide Congress "a summary of the basis for the determination," *id.* § 3252(b)(3)(B). Under the Department's regulations, that latter determination must be based on "a risk assessment [prepared] by the Under Secretary of Defense for Intelligence." 48 C.F.R. § 239.7304(a). Collectively, these procedural requirements help to ensure that the Secretary appropriately exercises the weighty authority granted him in § 3252.

The proffered administrative record shows that the Secretary violated each of those requirements here. He failed to make any reasoned determination that less intrusive measures could not reduce any risk Anthropic posed. Although the March Determination pronounces that conclusion, it does not explain it. AR 209. Nor does any other contemporaneous document in the record. And "mere lip service or verbal commendation of [the] standard" cannot satisfy the statutory requirement. *Nat. Res. Def. Council, Inc. v. Pritzker*, 828 F.3d 1125, 1135 (9th Cir. 2016); *see* Op. 33-34. The Secretary also failed to provide Congress the required analysis of less intrusive measures, "as Defendants conceded at oral argument." Op. 33; *see* P.I. Hr'g Tr. 20:15-21:1, 23:10-12. The six congressional notification letters in the record incant § 3252's text, but do not contain the mandated "discussion." 10 U.S.C. § 3252(b)(3)(B). Further still, the underlying risk assessment was prepared by Under Secretary for Research and Engineering Emil Michael—not the Under Secretary for Intelligence, as the Department's regulation requires. AR 213-16, 243. Defendants seek to justify reassigning the risk assessment by pointing to a recent Department reorganization, *see* AR 244-45, but an internal reorganization is no basis to act "contrary to existing valid regulations." *United States ex rel. Accardi v. Shaughnessy*, 347 U.S. 260, 268 (1954); *see Nat'l Ass'n of Home Builders v. Norton*, 340 F.3d 835, 852 (9th Cir. 2003).

2. The § 3252 Designation Is Contrary To Law

As a substantive matter, the supply chain risk designation is incompatible with the text of § 3252. *See* 5 U.S.C. § 706(2)(A), (C). Congress defined "supply chain risk" as "the risk that an adversary may sabotage, maliciously introduce unwanted function, or otherwise subvert . . . a covered system." 10 U.S.C. § 3252(d)(4). That text requires an "adversary" and "is directed at covert acts or hacks"—"not overt positions taken during contract negotiations." Op. 30.

Although the March Determination does not explain how Anthropic presents the kind of risk Congress described, AR 209, the underlying risk analysis asserts that "Anthropic's risk level escalated from a potentially manageable technical and business negotiation to an unacceptable national security threat over the course of the DoW's contract negotiation[s] with them," AR 215. Those negotiations, and Anthropic's private and public statements about them, purportedly led the Department to conclude that there was a "significant risk" that Anthropic itself was an adversary that would "subvert the design and/or functionality of their product or service." AR 210; *see also* AR 215. But Defendants ultimately assert that if Anthropic had "agreed to the Government's term, the challenged actions would not have occurred"—meaning that Anthropic's refusal to agree was the dispositive factor in the Secretary's § 3252 designation. Dkt. 96 at 23. Defendants thus "appear to be taking the position that any vendor who 'push[es] back' on" the Department in a contract negotiation becomes an "'adversary'" who presents a supply chain risk. Op. 30.

As this Court has recognized, "[t]hat position is deeply troubling and inconsistent with the statutory text." Op. 31. Under § 3252's plain terms, the stance Anthropic took in negotiations with the Department provides no basis for a supply chain risk designation. The text reflects Congress's choice to target the risk of malicious entities engaged in "covert acts or hacks," *id.*, not vendors openly expressing views with which the Department disagrees. A supply chain risk must come from an "adversary"—that is, an "enemy," "foe," "opponent," or "rival." *Adversary*, Black's Law Dictionary (12th ed. 2024). Those words naturally describe hostile foreign countries, terrorist groups, and similar actors. They do not describe a company like Anthropic, whatever one thinks of its stance on AI safety. A contractor does not give rise to a national-security risk because it disagrees with the Department about how a technology can be used safely. That is particularly so

where the contractor has consistently supported American national security agencies; has been entrusted with access to Top Secret facilities and classified information; and continues to provide the Department with technological tools that help to protect the Nation. *See supra* pp. 5-6.

Other statutory terms reinforce that conclusion. Section 3252 does not address all risks from an "adversary" but only the risks of "sabotage," "subver[sion]," and the "malicious[] introduc[tion of] unwanted function." 10 U.S.C. § 3252(d)(4). Each of those actions is typically taken or planned covertly. *See, e.g.*, *Sabotage*, Oxford English Dictionary (2d ed. 1989) ("[T]o ruin, destroy, or disable deliberately and maliciously (frequently by indirect means)."); *Subvert*, Oxford English Dictionary (2d ed. 1989) ("[T]o attempt to achieve, esp. by covert action, the weakening or removal of (a government, political regime, etc.)."). The statute further limits the relevant covert actions to those taken "so as to surveil, deny, disrupt, or otherwise degrade" covered national security systems. 10 U.S.C. § 3252(d)(4). That language connotes purposeful misconduct. *See So As*, Oxford English Dictionary (2d ed. 1989) ("Expressing result, consequence, or purpose: in such a way that; so that, in order that."). But Defendants designated Anthropic over its *public* disagreement about contract language. Dkt. 96 at 23. A company that openly announces its position in negotiations and publicly defends that position is not "sabotag[ing]," "subvert[ing]," or "maliciously introduc[ing]" anything. Those words describe the opposite of the transparent and public position taken by Anthropic here. *See* Op. 30.

"The legislative history confirms that Section 3252's purpose is to address covert acts." Op. 31. Congress enacted § 3252's predecessor provision in response to a cyber incident perpetrated by a foreign intelligence agency. That episode sparked concerns about how "the globalization of the information technology industry" could increase the risk that "systems and networks critical to DOD could be exploited through the introduction of counterfeit or malicious code and other defects introduced by suppliers of systems or components." S. Rep. No. 111-201, at 162 (2010). Those concerns bear no relation to the Department's dispute with Anthropic—an American company that actively supports the Department's mission and that the Department has never accused of an "attack," *id.*, or anything of the sort.

"[C]onsidered and consistent Executive Branch practice" resolves any doubt. *Kennedy v.*

Braidwood Mgmt., Inc., 606 U.S. 748, 783 (2025). Shortly after Congress enacted the provision that became § 3252, the Department issued instructions defining the relevant risk as "sabotage or subversion" by "foreign intelligence, terrorists, or other hostile elements." Dep't of Def., Instruction No. 5200.44 (Nov. 5, 2012), perma.cc/TW3D-735M. In keeping with that understanding, the Department has never before designated an entity a supply chain risk due to a contract dispute.

3. Defendants' Rationales Are Arbitrary And Capricious

Secretary Hegseth's proffered explanations are also arbitrary and capricious, providing another basis for this Court to hold unlawful and set aside the § 3252 Designation. 5 U.S.C. § 706(2)(A). The March Determination itself is wholly conclusory. AR 209. The only contemporaneous rationales appear in the memorandum from Under Secretary Michael. AR 213-16. But its justifications provide nothing close to the reasoned explanation the APA requires.

The overarching theme of the memorandum is that Anthropic presents a supply chain risk because its "unwillingness to agree" to the Department's preferred contract term amounted to an attempt "to grant itself" a "veto" over Department operations. AR 213. The memorandum asserts that Anthropic's preferred terms create a risk that it will "'deny, disrupt, or otherwise degrade'" covered Department systems by "diminish[ing] functionality and limit[ing] DoW's warfighting capabilities." *Id.* A post-decision Michael declaration reiterates the point, stating that Anthropic's two usage restrictions amount to "an operational veto" that "confirms that Anthropic did want to intervene in and assert control over DoW operations." AR 252; *see* P.I. Hr'g Tr. 41:18-23.

Those assertions are unfounded. Contract terms the Department was free to reject are not intrusions into the "operational decision chain." AR 215. And there is nothing reasoned or reasonable about the contention that, by refusing to abandon two narrow use restrictions, Anthropic demonstrated a propensity to interfere with Department operations. A company's forthright insistence on usage restrictions in contract negotiations (conducted to establish ground rules *before* models are deployed) does not provide evidence that the company would sabotage the Department *after* its models are deployed. A saboteur would not care about the particulars of contract terms. And the notion that the two use restrictions maintained by Anthropic signal a risk

of sabotage is especially irrational because the Department insists that it has no intention of using Anthropic's models for those uses. *See* Dkt. 96 at 14. Whatever deference is due to the Department's national-security judgments cannot provide cover for treating good-faith contractual disagreements as evidence of threatened sabotage or subversion.

In the related D.C. Circuit litigation, Secretary Hegseth emphasized a somewhat different rationale in Michael's memorandum: that Anthropic is a supply chain risk because AI technology is "opaque" and "vulnerable to manipulation," and the Department "cannot trust Anthropic to ensure the integrity of its models." AR 214; *see, e.g.*, Br. for Respondents 26-27, *Anthropic v. Dep't of War*, No. 26-1049 (D.C. Cir. May 6, 2026). But that rationale is no better than the first. The opacity of AI technology is "a common concern with all" large language models, AR 245, and the possibility of vendor manipulation is a risk with "technology more broadly," Op. 32.

That leaves only the "trust" rationale, which similarly fails. Michael rested his assertion that Anthropic "cannot be trusted" on Anthropic's purported conduct during negotiations. AR 214. He viewed Anthropic's negotiating posture as designed "to benefit [Anthropic's] public perception," and objected to Anthropic's engagement with "the press" and to a single question from an Anthropic employee about the Department's use of Claude. AR 213-15. In his view, that conduct created "material doubts as to whether [Anthropic] would cause [its] software to stop working or cause some other disastrous action." AR 213.

Even crediting that characterization of Anthropic's conduct, there is no "rational connection" between the conduct and Michael's conclusion. *Burlington Truck Lines Inc. v. United States*, 371 U.S. 156, 168 (1962). Every company wants to burnish its public image; that does not make it a threat of sabotage or covert attack. *See* AR 214. Nor is it reasonable to infer that Anthropic might undermine future military operations because the company raised a question about a prior operation or criticized the Department in the press. Engaging in public disagreements and frank conversations with the Department is a most unlikely strategy for any entity intent on "sabotag[ing]" or "subvert[ing]" it. 10 U.S.C. § 3252(d)(4). While Defendants might have lost "trust" in Anthropic as a contracting partner, that hardly establishes that Anthropic would secretly and maliciously interfere with the Department's information systems.

Indeed, the proffered administrative record reveals that the Department's explanations for its actions are pretextual. Almost all of the records produced in support of the designation were created *after* Secretary Hegseth ordered the Department to designate Anthropic a supply chain risk on February 27. *See* Op. 36; *supra* p. 13. That timeline "raises an inference" that the record "was generated after the fact to justify" a "foreordained conclusion." Op. 36.

The Department's own contradictory statements further demonstrate that this is a designation in search of a justification. According to the Michael memorandum, "Anthropic's risk level escalated" during its dealings with the Department to the point when, "during the final weeks of negotiations," the company became a "fully mature supply chain risk." AR 215. Yet Defendants also insist that if Anthropic had just accepted the Department's contract term on February 27, "the challenged actions would not have occurred." Dkt. 96 at 23. It is inconceivable that Defendants would allow a company that supposedly presented a "fully mature" threat of "model poisoning, insider threat risk, data exfiltration, and denial of service" (AR 215) to continue providing services merely because it agreed to a contract term. *See also* AR 255B (February 27 Hegseth Directive allowing continued use of Anthropic's services for up to six months). "Altogether, the evidence tells a story that does not match the explanation the Secretary gave for his decision." *Dep't of Com. v. New York*, 588 U.S. 752, 784 (2019).

The "belated justifications" offered by Under Secretary Michael in his declarations filed in this litigation on March 17 and March 24 do not salvage the Secretary's action. *Regents*, 591 U.S. at 23. Agency action cannot be upheld based on "*post hoc* rationalizations." *Am. Textile Mfrs. Inst., Inc. v. Donovan*, 452 U.S. 490, 539 (1981). In any event, the post hoc rationalizations here would not change the analysis. The concern about Anthropic's employment of "foreign nationals" applies to "other major U.S. AI labs," as Michael concedes. AR 245. The fact that "adversarial nation states" have stolen Anthropic's technology (AR 246) makes Anthropic a victim, not an adversary. And when the CDC used a commercial version of Claude to conduct public health research and encountered difficulties as a result of the commercial Usage Policy, it is undisputed that Anthropic promptly assisted the CDC as soon as it learned of the situation. AR 246. Far from suggesting sabotage or adversarial intent, that shows a company responding to government needs

and working collaboratively with federal partners.

C. Additional Materials Properly Considered Here Reinforce The APA Violations

As addressed in a concurrently filed motion, in evaluating the APA claim, this Court should consider two additional documents beyond the materials proffered by Defendants. The first—an email from Under Secretary Michael to Anthropic—is part of the "whole record." 5 U.S.C. § 706; *see Thompson v. U.S. Dep't of Lab.*, 885 F.2d 551, 555 (9th Cir. 1989). The second—the June 10 Declaration of Thiyagu Ramasamy ("the Ramasamy Declaration")—may be considered "to identify and plug holes in the administrative record." *Lands Council v. Powell*, 395 F.3d 1019, 1030 (9th Cir. 2005). That additional evidence confirms the APA violations.

Under Secretary Michael's correspondence with Anthropic confirms that the asserted concerns about Anthropic presenting a supply chain risk are "contrived." *Dep't of Com.*, 588 U.S. at 784. On March 4—two days after Under Secretary Michael prepared the risk assessment and one day after the Secretary's Determination—Michael sent Anthropic a revised counteroffer. Dkt. 114-1 at 2. He wrote: "I think we are very close here," and "I hope this work[s]." *Id.* If the Department genuinely believed that Anthropic was likely to maliciously disrupt military operations, such continued and enthusiastic negotiations would have been unthinkable.

The additional materials also fatally undermine the Department's contentions that Anthropic demanded an "operational veto," AR 213, and that the company might "alter the behavior" of its models "in the middle of ongoing warfighting operations." AR 215. As the Ramasamy Declaration explains, Anthropic lacks the technical capability to take any such actions. It "has no ability to access, alter, or shut down" a model that has been "deployed in DoW environments." Ramasamy Decl. ¶ 48. It maintains no "back door or remote kill switch" in Claude. *Id.* And "Anthropic personnel cannot log into a Department system to modify or disable the models during an operation." *Id.* Before a new or updated model is deployed, moreover, "Department security reviewers inspect and approve it to ensure it is safe and appropriate." *Id.* ¶ 53. This evidence—all of which Anthropic would have submitted to the Department before the § 3252 designation, if given the opportunity to do so—shows that Anthropic could not exercise an "operational veto" even if it wanted to. AR 213.

III. Defendants Violated The Due Process Clause By Blacklisting Anthropic Without Notice Or An Opportunity To Be Heard

The undisputed facts also establish a violation of Anthropic's due process rights. To begin with, they show that Anthropic has "been deprived of a protected interest in 'property' or 'liberty.'" *Am. Mfrs. Mut. Ins. Co. v. Sullivan*, 526 U.S. 40, 59 (1999). Defendants previously argued (Dkt. 96 at 30) that the due process claim "fails at this first step." That argument can be rejected here because it relies on no disputed facts, only a mistaken understanding of the law.

The Challenged Actions deprived Anthropic of three protected interests. *First*, the Presidential Directive abridged Anthropic's liberty interest in "follow[ing] [its] chosen profession free from unreasonable governmental interference," *Greene v. McElroy*, 360 U.S. 474, 492 (1959), by "debarring" it "from government contract[ing]." *Trifax Corp. v. Dist. of Columbia*, 314 F.3d 641, 643 (D.C. Cir. 2003); *see* Op. 25-26. The Presidential Directive ordered federal agencies to "IMMEDIATELY CEASE all use of Anthropic's technology" and "not do business with [it] again." AR 255A. And Defendants' implementation of that ban "automatically exclude[s]" Anthropic from a "category of future [agency] contracts," likewise triggering due process protections. *Kartseva v. Dep't of State*, 37 F.3d 1524, 1528 (D.C. Cir. 1994). *Second*, the Challenged Actions deprived Anthropic of "a cognizable liberty interest," *Fikre v. FBI*, 35 F.4th 762, 776 (9th Cir. 2022), by stigmatizing Anthropic as "a Supply-Chain Risk to National Security," AR 255B, and a company that put "AMERICAN LIVES at risk," AR 255A, while stripping it of federal contracts and any opportunity to compete for contracts in the future. *See Ulrich v. City & Cty. of San Francisco*, 308 F.3d 968, 983 (9th Cir. 2002); Op. 26-27. *Third*, the secondary boycott order prohibited Anthropic from "conduct[ing] any commercial activity" with military contractors, AR 255B, implicating "significant" property interests, *Al Haramain Islamic Found. v. Dep't of Treasury*, 686 F.3d 965, 979-80 (9th Cir. 2012); *see also Ralls Corp. v. CFIUS*, 758 F.3d 296, 315-16 (D.C. Cir. 2014).

And the meager post-deprivation procedures that Defendants offered Anthropic here do not "comport with due process." *Sullivan*, 526 U.S. at 59. "The fundamental requirements of procedural due process are notice and an opportunity to be heard *before* the government may

deprive a person of a protected liberty or property interest." *Conner v. City of Santa Ana*, 897 F.2d 1487, 1492 (9th Cir. 1990) (emphasis added) (citing *Mathews v. Eldridge*, 424 U.S. 319, 333 (1976)). The record here establishes a "complete lack of prior notice or process" as to each of the Challenged Actions. Op. 27; *see also* Heck Decl. ¶¶ 22, 24.

The government told Anthropic that it wanted to alter the parties' contract terms, *see* Heck Decl. ¶ 11, but gave Anthropic no notice or opportunity to object before publicly barring the company from all federal government work and blacklisting it with all military contractors, *see* AR 255A, 255B. The Department also did not provide Anthropic with the factual basis for the supply-chain-risk designation before imposing it. The document containing the purported rationales was not disclosed to Anthropic until after this litigation began, *see* AR 213-16, and that analysis relied on accusations that Anthropic had no prior opportunity to contest, AR 214-15. Nor did Secretary Hegseth's statements during negotiations provide the requisite notice. At the same time he suggested that Anthropic could be designated a supply chain risk, he also suggested that the Department could invoke the Defense Production Act to force Anthropic to provide its technology on the Department's preferred terms. Heck Decl. ¶ 17. Those conflicting positions did not give Anthropic meaningful notice of the basis for the later designation, much less a meaningful opportunity to respond before Defendants acted. *See* Op. 27 (opining that "[t]his contradiction underscores the complete lack of prior notice or process.").

Anthropic also did not receive pre-deprivation notice or process from the agencies that carried out the Presidential Directive. Within hours of that Directive, GSA directed Anthropic's reseller to submit a deletion modification, AR 403, and removed Anthropic from USAi.gov and the Multiple Award Schedule, AR 394-95, 408. HHS disabled enterprise access to Claude that evening. AR 335, 336. The State Department initiated removal of the Anthropic model from its internal generative AI tool by the close of business. AR 321. None of these actions was preceded by notice to Anthropic, an opportunity to respond, or any agency-specific risk finding.

Defendants contend that they provided constitutionally sufficient post-deprivation process. *See* Dkt. 96 at 31. But they did not offer *any* post-deprivation process for the Presidential Directive or Secretary Hegseth's secondary boycott order. And the 30-day reconsideration process

the Department offered on March 4 with respect to the § 3252 designation (*see* AR 238) is constitutionally inadequate. The constitutional baseline is that "[p]re-deprivation notice typically is required absent 'a strong justification' from the government." *Garza v. Woods*, 150 F.4th 1118, 1130 (9th Cir. 2025); *see generally Cleveland Bd. of Educ. v. Loudermill*, 470 U.S. 532, 542 (1985). The record here establishes no such justification. *See* Op. 28-29; AR 209-16.

In Defendants' telling, the "national security" setting warranted only post-deprivation process. Dkt. 96 at 31. But "the foreign-policy/national-security nature" of an action does not automatically "support[] the constitutional adequacy of a post-deprivation remedy." *Nat'l Council of Resistance of Iran v. Dep't of State*, 251 F.3d 192, 207 (D.C. Cir. 2001) (*NCRI*). Indeed, the D.C. Circuit has repeatedly held that due process typically requires advance notice as to the government's intent to designate an entity a foreign terrorist organization. *See id.* at 207-08; *People's Mojahedin Org. of Iran v. Dep't of State*, 613 F.3d 220, 227-28 (D.C. Cir. 2010) (per curiam). Even in the national-security context, therefore, Defendants must "make a showing of particularized need" to dispense with pre-deprivation process. *NCRI*, 251 F.3d at 208. Situations presenting such a need are "'extraordinary'" and "truly unusual." *Fuentes v. Shevin*, 407 U.S. 67, 90 (1972); *see, e.g.*, *Al Haramain*, 686 F.3d at 985 (substantial risk of "asset flight").

Here, Defendants have "present[ed] no evidence of exigency whatsoever," Op. 28, let alone a "particularized need" sufficient to justify depriving Anthropic of process before the Challenged Actions, *NCRI,* 251 F.3d at 208. To be sure, they have vaguely suggested a need for "quick action." Dkt. 96 at 31. But "nothing in the record indicates a heightened supply chain risk to national security systems on February 27 or March 3." Op. 28. To the contrary, the Department had been using Claude subject to the restrictions to which it now objects. Ramasamy Decl. ¶¶ 8, 25-27, 65; Kaplan Decl. ¶ 29. The Department's own conduct refutes any urgency: it allowed Anthropic six months to wind down, AR 242, and it continues to rely on Anthropic's tools for national-security functions to this day, Heck Decl. ¶ 27; Ramasamy Decl. ¶¶ 80-81; *see* Mongan Decl. Exs. 2-3, 6.

The record also "show[s] that meaningful pre-deprivation notice of DoW's concerns and an opportunity to respond to those concerns were necessary to avoid the risk of erroneous

deprivation." Op. 27-28; *see Mathews*, 424 U.S. at 335. The central rationale underlying the § 3252 designation is the assertion that Anthropic would alter or shut down its models to compromise Department operations. Op. 28; *see* AR 213, 215. But the unrebutted evidence now "shows that Anthropic is not technologically capable of taking such unilateral actions." Op. 28; *see supra* pp. 10, 21; Ramasamy Decl. ¶¶ 48-58. Given the opportunity, Anthropic would have corrected the Department's misunderstanding before the Department took action. Instead, the Department's "factual errors [went] uncorrected" because its failure to afford constitutionally required process deprived the Secretary of "easy, ready, and persuasive explanations." *Al Haramain*, 686 F.3d at 982.

IV. The Presidential Directive Violates The Separation Of Powers

The Presidential Directive is an immediate, permanent, government-wide debarment of Anthropic, issued by executive fiat. Presidential power must "stem either from an act of Congress or from the Constitution itself." *Youngstown Sheet & Tube Co. v. Sawyer*, 343 U.S. 579, 585 (1952); *see also State v. Su*, 121 F.4th 1, 7-13 (9th Cir. 2024). The Directive does not. It invokes no statutory authority and has no constitutional basis. Congress—not the President—sets the rules for exclusion from federal contracting, including which contractors may be excluded, under what circumstances, and through what process. *See, e.g.*, 10 U.S.C. §§ 3203(a)(1), 3204(a); 41 U.S.C. §§ 3303(a)(1), 3304(a); *see also* 48 C.F.R. §§ 9.406-2, 9.407-2. And even the Executive has recognized that exclusion of a contractor is not a tool "for purposes of punishment." 48 C.F.R. § 9.402(b). By defying Congress's chosen scheme for exclusion from federal contracting, the President acted at the "lowest ebb" of presidential authority, and the Directive may stand only if it is supported by an independent constitutional power. *Youngstown*, 343 U.S. at 637 (Jackson, J., concurring). It is not.

V. Implementation Of The Presidential Directive Violated 5 U.S.C. § 558

Finally, the actions taken by defendant agencies to implement the Presidential Directive violated the APA. Agencies may not impose sanctions or issue orders "except within jurisdiction delegated to the agency and as authorized by law." 5 U.S.C. § 558(b); *see Am. Bus Ass'n v. Slater*, 231 F.3d 1, 7 (D.C. Cir. 2000). No statute authorizes federal agencies to abruptly impose blanket

orders and sanctions that terminate all Executive Branch use of Anthropic tools. *See, e.g.*, AR 259-66; AR 319-32; AR 403-10. Those actions violate § 558(b) and must be set aside.

CONCLUSION

Anthropic's motion for summary judgment should be granted.

Respectfully submitted,

Date: June 10, 2026

/s/ Michael J. Mongan

MICHAEL J. MONGAN (SBN 250374)
michael.mongan@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
50 California Street, Suite 3600
San Francisco, CA 94111
Telephone: (628) 235-1000

EMILY BARNET (*pro hac vice*)
emily.barnet@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
7 World Trade Center
250 Greenwich St
New York, NY 10007
Telephone: (212) 230-8800

KELLY P. DUNBAR (*pro hac vice*)
kelly.dunbar@wilmerhale.com
JOSHUA A. GELTZER (*pro hac vice*)
joshua.geltzer@wilmerhale.com
KEVIN M. LAMB (*pro hac vice*)
kevin.lamb@wilmerhale.com
SUSAN HENNESSEY (*pro hac vice*)
susan.hennessey@wilmerhale.com
LAUREN MOXLEY BEATTY (SBN 308333)
lauren.beatty@wilmerhale.com
MATTHEW E. MORRIS (*pro hac vice*)
matt.morris@wilmerhale.com
BARDIA VASEGHI (*pro hac vice*)
bardia.vaseghi@wilmerhale.com
DAKOTA C. FOSTER (*pro hac vice*)
dakota.foster@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000

Attorneys for Plaintiff Anthropic PBC

CERTIFICATE OF SERVICE

I hereby certify that on this 10th day of June, 2026, I electronically filed the foregoing document using the CM/ECF System, that all participants in the case are registered CM/ECF users, and that service will be accomplished by the CM/ECF system.

By: /s/ *Michael J. Mongan*
Michael J. Mongan

MICHAEL J. MONGAN (SBN 250374)
michael.mongan@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
50 California Street, Suite 3600
San Francisco, CA 94111
Telephone: (628) 235-1000

EMILY BARNET (*pro hac vice*)
emily.barnet@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
7 World Trade Center
250 Greenwich Street
New York, NY 10007
Telephone: (212) 230-8800

Attorneys for Plaintiff Anthropic PBC

KELLY P. DUNBAR (*pro hac vice*)
kelly.dunbar@wilmerhale.com
JOSHUA A. GELTZER (*pro hac vice*)
joshua.geltzer@wilmerhale.com
KEVIN M. LAMB (*pro hac vice*)
kevin.lamb@wilmerhale.com
SUSAN HENNESSEY (*pro hac vice*)
susan.hennessey@wilmerhale.com
LAUREN MOXLEY BEATTY (SBN 308333)
lauren.beatty@wilmerhale.com
MATTHEW E. MORRIS (*pro hac vice*)
matt.morris@wilmerhale.com
BARDIA VASEGHI (*pro hac vice*)
bardia.vaseghi@wilmerhale.com
DAKOTA C. FOSTER (*pro hac vice*)
dakota.foster@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ANTHROPIC PBC,

Plaintiff,

v.

U.S. DEPARTMENT OF WAR, et al.,

Defendants.

Case No. 3:26-cv-01996-RFL

DECLARATION OF JARED KAPLAN

I, Jared Kaplan, pursuant to 28 U.S.C. § 1746, declare as follows:

Personal Background

1. I am one of the co-founders of Anthropic PBC ("Anthropic"), an artificial intelligence ("AI") company based in San Francisco, California.

2. Since 2023, my title has been Chief Science Officer. In that role, I oversee the company's research in model development and safety, which is at the core of Anthropic's mission. Under model development, my responsibilities include overseeing the fine-tuning and reinforcement-learning-driven capabilities that shape each new generation of models. I also oversee teams working on interpretability, which is the study of how large language models ("LLMs") work by observing their internal operations. As part of my safety portfolio, I supervise teams working on "alignment," a term that broadly refers to efforts to make AI systems' goals, behaviors, and outputs reliably follow human values and intentions. I also oversee the safeguards implemented in and around deployed models.

3. As of 2024, I have also served as Anthropic's Responsible Scaling Officer. In that role, I am responsible for overseeing the implementation of Anthropic's Responsible Scaling Policy, which is a series of technical and organizational protocols that aim to manage the risks associated with developing increasingly capable AI systems.

4. Before co-founding Anthropic, I was a consultant at OpenAI and contributed to the development and analysis of LLM research. During that time, I was involved in the research and development efforts to introduce some of OpenAI's early LLM models, such as GPT-3 and Codex.

5. I began my career as a theoretical physicist, with a focus on effective field theory, particle physics, cosmology, scattering amplitudes, and the conformal field theory bootstrap. Since 2012 and continuing to today, I have been a professor in the Department of Physics and Astronomy at Johns Hopkins University.

6. During a sabbatical from my work in theoretical physics, I began collaborating with computer scientists to research machine learning and the development of LLMs, which are text-based AI systems trained on extremely large datasets to develop a functional understanding of

language and generate new text. Since then, I have taught courses and published over 60 scholarly articles on a mixture of theoretical physics, machine learning, and LLMs.

7. I have extensive personal knowledge regarding Anthropic's core research and safety objectives, including how Anthropic's AI models are developed and trained, their technical capabilities and risks, our approach to safety research and safe deployment, and how our mission informs our work across all of these domains. In my role as Chief Science Officer, I have personal knowledge of the contents of this declaration, or have knowledge of the matters based on my review of information and records gathered by Anthropic personnel, and could testify thereto.

Anthropic's Background As An AI Safety Company

8. We founded Anthropic because we anticipated how powerful AI could become and believed it could reshape society in profound ways. From our own experience working with LLMs and scaling laws, it became clear that AI capabilities were advancing rapidly and could soon rival or surpass human performance across many areas. At the same time, we did not yet know how to reliably make these systems both helpful and safe, and we anticipated that speed, competition, and social disruption could push people to deploy AI before its capabilities and risks were understood and sufficiently mitigated. This is why we built Anthropic—to put safety at the center as AI progress accelerates, to study these questions on the most advanced models where they matter most, and to build an organization that could turn careful, empirical safety research into real-world practice. That conviction is embedded in Anthropic's very structure as a public benefit corporation.

9. Our mission to build safe, beneficial AI is the foundation of everything we do—from model development to safety science to policy engagement.

10. Anthropic began as a research-first company and, for its first two years, focused exclusively on foundational AI research, the science of AI safety, and AI policy work. We regularly publish pioneering research on alignment, interpretability, and the societal impacts of LLMs.

11. Anthropic aims to develop models that are safe, ethical, and helpful. Our safety work is grounded in empiricism, rigorous research, and humility. Because we have invested

deeply in understanding the capabilities and limitations of our systems, we have unique insights into what guardrails are necessary for safe deployment.

12. While we initially developed AI models to support safety research, we expanded the company's focus to include commercial deployments which began in early 2023. We believe this matters because a safe AI system that is not used cannot fully demonstrate the benefits of responsibly developed frontier AI. By showing that AI can be both safe and commercially successful, we aim to pull the broader industry toward higher safety standards—what we call the race to the top.

13. We also engage in public advocacy for transparency and safety in AI development and have supported state and federal legislation advancing those goals.

Anthropic's LLM Claude And The Role Of Guardrails

14. Anthropic's signature model is a general-purpose LLM called Claude. We make Claude available to individual users, small businesses, and large organizations through a variety of offerings. We continually develop and release increasingly capable versions of Claude.

15. LLMs like Claude are algorithmic systems trained on massive datasets to identify patterns and associations in language and to generate outputs and take actions that resemble human responses and actions. Through training, models acquire predictive power and the transformative ability to take a range of actions in a fraction of the time it would take humans to perform them.

16. Claude is a versatile technology, much like an actual human mind. When paired with a chatbot interface, Claude is capable of interpreting and responding to a wide range of user inputs, or "prompts," in an intelligent, human-like manner. In this medium, Claude can analyze and summarize large volumes of text, write and edit content, generate and debug source code, and reason through complex and multi-step problems. Claude can also be given access to tools so that it can behave "agentically," meaning it can not only respond to users' prompts but actually take actions on their behalf. Simple actions could include sending emails, deploying code, and navigating the Internet. Claude can also power more sophisticated agentic work. For instance, someone planning a vacation can direct Claude to compare flight and hotel options against

specified constraints, assemble a day-by-day itinerary, make reservations, send confirmation emails and calendar invites, and produce a summary of the plan and costs. With some configurations, Claude can even act autonomously, executing tasks without requiring ongoing user direction. Using AI systems to power agents is understandably of particular interest to many users, including certain government users, even though agentic usage poses heightened risks relative to the traditional chatbot form.

17. Because Claude is a dual-use technology, the risks that it poses depend on the specific context in which it is used. Many tools are dual use: For example, a chef's knife is a useful device in the kitchen and a dangerous weapon in a violent conflict. When it comes to AI, the same capabilities that drive medical breakthroughs, accelerate scientific research, and enhance human creativity can, in other contexts, also enable dangerous actors to develop new weapons or automate complex, malicious activities. The dual-use nature of AI makes it difficult to restrict "dangerous capabilities" while promoting "beneficial ones"—they are often the same capability applied to different ends.

18. Although AI systems pose the potential for tremendous benefits, they also create novel risks. Among other concerns, LLMs can produce responses that diverge from the goals of the people that trained them or reflect skewed or mistaken judgments embedded in their training data. To address the novel risks, we have pursued a multilayered approach to safety, implementing safety mitigations at the model layer, the safeguards layer, and the policy layer.

19. At the model layer, we seek to embed safety considerations directly into the model itself through a variety of training techniques. A central focus of our research is on solving the challenge of alignment to make AI systems reliably follow human values and intentions. One of our key techniques is Constitutional AI, which trains models to evaluate and revise their own outputs against a set of normative principles, like balancing helpfulness against harm avoidance, and respecting values such as individual privacy and political freedom. In addition to Constitutional AI, we use reinforcement learning from human feedback (RLHF) to reduce the likelihood of harmful outputs. RLHF is a training technique in which human reviewers rank pairs

of model outputs; those preferences are used to train the model to generate responses that better align with human judgments.

20. The safeguards layer consists of technical measures that stack on top of the model itself. As appropriate, our tools may include classifiers and probes that detect harmful activity in real time, targeted interventions that reduce the likelihood of harmful outputs, as well as monitoring systems that help us identify when our systems are being misused at scale. We continually calibrate which measures are appropriate based on the circumstances of the deployment, the type of harm we are trying to prevent, and other factors.

21. At the uppermost layer, our Usage Policy defines how Claude is permitted to be used. Claude is only available subject to Terms of Service that incorporate its Usage Policy. At a high level, our policy informs the development of our technical safety measures, provides users with clarity on the scope of permissible usage, and steers them away from using our models in risky ways, including ways we, as Claude's creator, understand that it has not been developed for and/or is not ready for.

22. For commercial and civilian users, the Usage Policy reflects our judgment—based on technical expertise, our experience at the frontier of AI development, and our values as a company—on how to strike an optimal balance between enabling beneficial uses of AI while mitigating potential harms. The Usage Policy generally prohibits uses that pose unacceptable risks, including surveillance, compromising computer systems or networks, and designing weapons or other systems to cause harm or loss of human life. The Usage Policy is an agreement we enter with our users so that we both understand how Claude should and should not be used. The Usage Policy is critical because technical safeguards alone cannot prevent all dangerous uses: they do not necessarily have access to the full context that determines whether a given request falls within or outside the lines set by the Usage Policy, and in certain environments, such as classified settings, there may be very limited visibility into how the systems are being used. As a result, the Usage Policy is a critical mechanism for clearly articulating safety boundaries to users and serves as an important last line of defense.

23. Severing Claude from the usage limitations we have determined are essential would erode the very purpose for which our company was founded and contradict our deeply held values.

Anthropic's Commitment To Supporting National Security Engagements While Maintaining Critical Safeguards

24. Any assertion that Anthropic is aligned with, or poses risks of subversion from, "adversaries" of the United States could not be further from the truth. We are committed to defending the United States and defeating our authoritarian adversaries. For example, we have consistently taken steps to *prevent* our models from being used by U.S. adversaries and to prioritize U.S. national security over narrow commercial self-interest. As examples, Anthropic has gone to significant lengths to prevent the use of its technology by entities linked to the Chinese Communist Party, has shut down attempts to abuse Claude for state-sponsored cyber operations, and has advocated for strong export controls on the most powerful chips used to train AI, all to preserve the U.S. lead in frontier AI development.

25. We have been aligned with the U.S. government's priority to sustain and enhance America's global AI dominance to promote economic growth, human flourishing, and national security.

26. Since as early as 2024, Anthropic has led the field in supporting U.S. national security priorities. We collaborated closely with national security stakeholders on a variety of initiatives to advance our shared goal of building safe AI systems. These include collaboration with federal partners on AI safety research, evaluation frameworks, and strategic cloud partnerships as AI assumed a more prominent national security role. As a result, Anthropic's AI models were the first ever to be used by American warfighters on classified systems. Today, Claude is reportedly the Department of War's ("DoW" or the Department") most widely deployed frontier AI model and the only one currently on classified systems.

27. We were also the first to proactively work to align our AI system with the government's national security needs. Anthropic developed Claude Gov, a dedicated model for national security users to address real-world operational needs that also includes a government-specific addendum to the Usage Policy described above. While Claude Gov underwent the same

rigorous safety testing as all other Claude models, it was designed to fulfill the missions of our national security customers and is more likely to comply with requests that are appropriate in a military context. For example, some standard versions of Claude refuse to analyze documents that appear to be classified, such as materials marked Top Secret. That restriction is appropriate for ordinary commercial users, but it would be incompatible with legitimate national security uses by government personnel, and Claude Gov will not refuse to analyze Top Secret documents.

28. The government-specific Usage Policy addendum was designed to strike a balance between enabling national security beneficial uses and mitigating potential harms. For example, whereas ordinary civilians do not conduct foreign intelligence analysis, the government's military and intelligence communities do. As a result, the government-specific addendum does not impose the same restrictions on national security use as it does on civilian customers. In our recent negotiations discussed below, for example, we made clear that we were prepared to authorize use of our models for additional purposes for the DoW—such as developing more effective weapons—but we have never allowed regular users to use Claude to assist with weapons development.

29. Anthropic partnered closely with national security prime contractors to enable the provision of our models on their platforms, through which DoW and other government national security customers access AI systems. DoW gained access to Claude Gov for the first time in March 2025 via Anthropic partner platforms—and its usage was governed by Anthropic's Usage Policy and the government-specific addendum. During this period, the government-specific addendum imposed broad restrictions that would have prohibited mass surveillance of Americans and lethal autonomous warfare. DoW assented to these terms by using Claude Gov through the platforms of Anthropic's partners and, to my knowledge, did not object to them at any point.

30. In July 2025, Anthropic engaged in a separate negotiation with DoW regarding how the Department might further expand its usage of our models by accessing them directly from Anthropic rather than through our partners. These discussions never advanced beyond scoping out potential work; because we did not reach the implementation phase, the terms of a Usage Policy were not discussed.

31. Meanwhile, throughout this period, DoW continued to use our models through partner platforms, subject to the broad restrictions of the government-specific addendum to the Usage Policy described above. DoW was satisfied with our models, embedded them into its operations, and expanded their usage.

32. In the fall of 2025, DoW and Anthropic began negotiations regarding a new deployment of our models on DoW's GenAI.mil platform. The discussion contemplated various types of deployments, including some that, if implemented, might require a direct contractual relationship between DoW and Anthropic, including with respect to the Usage Policy. During these negotiations, DoW asked Anthropic to remove its Usage Policy not just with respect to the GenAI.mil platform but across all existing and future offerings and to permit DoW, and its contractors and subcontractors, to use all versions of Claude for "all lawful uses." Anthropic engaged in these negotiations in an effort to support DoW's national security priorities in a manner consistent with the company's core principles. As part of these efforts, the Department sent partial contract language incorporating this term to Anthropic and delivered an ultimatum: Anthropic must agree to the revised term or lose all current and future Department business. Contract modifications for facility and personnel clearances and classified work have been frozen since then as the parties continue to discuss the Department's demand.

33. Anthropic ultimately agreed to allow DoW, and its contractors and subcontractors, to use Claude without a broad set of restrictions that had previously applied to all DoW usage. However, Anthropic set two critical exceptions: mass surveillance of Americans and lethal autonomous warfare. With those two limitations, Anthropic agreed to "all lawful uses" of Claude. This change, if accepted, would have shifted the structure of the government-specific addendum from a "whitelist" approach—under which broad prohibitions applied with limited authorized exceptions—to a "blacklist" approach, under which all lawful uses were permitted except for these two prohibitions.

34. First, we would not agree to the use of Claude to carry out mass surveillance of Americans. It is my understanding that existing surveillance laws were written before the advent of frontier AI systems. Tools like Claude enable aggregation and analysis of massive datasets at

unprecedented scale, potentially facilitating practices inconsistent with Americans' rights even if they appear arguably compliant with laws written before the advent of AI and interpreted by courts only in a pre-AI context. For example, while there is generally no expectation of privacy in public spaces, powerful AI could enable the government to aggregate and analyze millions of public surveillance camera feeds into real-time, population-scale tracking—capabilities not contemplated or addressed by existing federal law. Our legal frameworks have not yet adapted to these novel technologies. Especially in a moment where technology has so outpaced legal frameworks, we at Anthropic, based on our distinctive understanding of what this technology can effectuate, do not believe it is safe or responsible for an AI developer to knowingly enable large-scale surveillance of Americans. Permitting such use would risk Claude being misused in ways that seriously infringe Americans' rights. Moreover, removing these limitations would also create a risk of inadvertent harm, such as by Claude collecting more information about U.S. persons than the user intended. For example, a user might ask Claude to obtain a specific piece of information about a U.S. person that is lawful to query, but because Claude is operating in an unsafe context, it could inadvertently collect or synthesize a far broader set of information about that individual—including information that the U.S. government is not permitted to collect or query for certain purposes—even absent any intent by the user to violate rules designed to safeguard civil liberties.

35. Second, we would not agree to the use of Claude for lethal autonomous warfare. Lethal autonomous warfare consists of using AI to control weapons without any human oversight when human lives are at risk. Such applications include, for example, an AI-controlled aerial system that independently identifies and classifies an object as a military target, determines engagement criteria are satisfied, and launches a weapon strike without any human reviewing, approving, or having the ability to override decisions made by AI. In our view, today's AI systems—including Claude—are not capable of reliably carrying out lethal autonomous warfare; this is why we have insisted on meaningful human oversight. As anyone who has used a generative AI tool knows, Claude can make errors. In the context of military options, these errors could have grave consequences, jeopardizing the success of military operations or potentially

costing the lives of American soldiers or innocent civilians. For example, it is possible that an AI system could misidentify an American soldier as a terrorist operative. Using Claude in this manner would place America's warfighters and innocent civilians at unacceptable risk.

36. Because we trained, and have extensively red-teamed our models, we have a unique understanding of Claude's capabilities and therefore have a deep technical understanding of its limitations. From the perspective of that expertise, we have emphatically concluded that Claude is not yet safe for those uses.

37. To be clear, we will not and have never second-guessed the government's national security judgments or missions. Anthropic fully respects that any decisions about military operations rest with DoW. Anthropic also fully respects that because of Claude's limitations and safeguards, DoW may opt to work with another company that better suits its needs. Anthropic has not sought, and would not seek, to dictate how the government conducts its missions and who it works with.

38. At the same time, acceding to DoW's demand that we remove these two policy-layer safeguards limitations would undercut Anthropic's core identity and competitive advantage. Anthropic has built its identity, reputation, and trust with customers, partners, investors, and the public on a principled commitment to safety. Stripping away those safeguards would erode internal and external trust, weaken the company's culture, and threaten its ability to attract and retain the expertise and commitment necessary to build innovative, cutting-edge AI systems—harms that extend well beyond the immediate technical effects of lifting the two use restrictions at issue here.

39. Maintaining these restrictions on Claude's use in military operations is essential to Anthropic's mission to advance the safe and beneficial development and use of AI. That is why we articulated these two critical use limitations to DoW: given the current state of our systems, allowing Claude to be used for mass surveillance of Americans or for lethal autonomous warfare would not only contravene our expert technical judgment but also the very principles on which Anthropic was founded.

* * *

I declare under penalty of perjury, pursuant to 28 U.S.C. § 1746, that the above is true and correct to the best of my knowledge.

Executed on June 10, 2026.

/s/ Jared Kaplan
Jared Kaplan
Co-Founder, Anthropic

ATTESTATION PURSUANT TO CIVIL LOCAL RULE 5-1(i)(3)

Pursuant to Civil Local Rule 5-1(i)(3), I attest that concurrence in the filing of this document has been obtained from the other signatories.

By: */s/ Michael J. Mongan*
Michael J. Mongan

MICHAEL J. MONGAN (SBN 250374)
michael.mongan@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
50 California Street, Suite 3600
San Francisco, CA 94111
Telephone: (628) 235-1000

EMILY BARNET (*pro hac vice*)
emily.barnet@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
7 World Trade Center
250 Greenwich Street
New York, NY 10007
Telephone: (212) 230-8800

Attorneys for Plaintiff Anthropic PBC

KELLY P. DUNBAR (*pro hac vice*)
kelly.dunbar@wilmerhale.com
JOSHUA A. GELTZER (*pro hac vice*)
joshua.geltzer@wilmerhale.com
KEVIN M. LAMB (*pro hac vice*)
kevin.lamb@wilmerhale.com
SUSAN HENNESSEY (*pro hac vice*)
susan.hennessey@wilmerhale.com
LAUREN MOXLEY BEATTY (SBN 308333)
lauren.beatty@wilmerhale.com
MATTHEW E. MORRIS (*pro hac vice*)
matt.morris@wilmerhale.com
BARDIA VASEGHI (*pro hac vice*)
bardia.vaseghi@wilmerhale.com
DAKOTA C. FOSTER (*pro hac vice*)
dakota.foster@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ANTHROPIC PBC,

Plaintiff,

v.

U.S. DEPARTMENT OF WAR, et al.,

Defendants.

Case No. 3:26-cv-01996-RFL

DECLARATION OF PAUL SMITH

I, Paul Smith, pursuant to 28 U.S.C. § 1746, declare as follows:

1. My name is Paul Smith. I am the Chief Commercial Officer ("CCO") of Anthropic PBC ("Anthropic"), where I have worked since 2025. As Anthropic's CCO, I am responsible for driving the trusted enterprise adoption of Anthropic's AI systems and for maintaining long-term customer confidence to deploy Anthropic's large language model ("LLM"), Claude, in their highly-regulated and business-critical environments.

2. Prior to joining Anthropic, I spent the last thirty years in senior commercial leadership roles at major enterprise technology companies, focusing on building and scaling their global go-to-market strategies. Most recently, I served as the President of Global Customer and Field Operations at ServiceNow, where I oversaw worldwide sales and customer operations. Prior to that role, I held senior leadership roles at Salesforce and Microsoft.

3. As the CCO for Anthropic, I have personal knowledge of the contents of this declaration, or have knowledge of the matters based on my review of information and records gathered by Anthropic personnel, and could testify thereto.

4. I understand that Anthropic and the Department of War (the "Department") had been discussing a direct agreement to deploy our AI models on classified systems. Despite ongoing negotiations, on February 27, 2026, President Trump posted on Truth Social, directing all agencies to cease use of Anthropic's AI system immediately. Secretary of War Pete Hegseth then posted on X.com that he was designating Anthropic "a supply chain risk," which he claimed would prohibit contractors, suppliers, and partners that conduct business with the United States military from engaging commercially with Anthropic. On March 4, 2026, Anthropic received two letters signed by Secretary Hegseth and dated March 3, 2026, that notified the company of its purported designation. I refer to these actions collectively below as the "Government's Actions."

The Government's Actions And Statements Have Tarnished Anthropic's Reputation

5. The Government's Actions attempt to blacklist Anthropic. They send a clear message to the market: do not associate with Anthropic, or else. The Government's Actions signal that Anthropic is *persona non grata* in the eyes of the government, a message that threatens to reverberate across the broader market. Like any complex, large business, Anthropic depends on an

interconnected network of relationships—including with government agencies, prime contractors, commercial partners, cloud providers, investors, current or prospective employees, customers, and the public. Reputation underpins each of these relationships, and harm to Anthropic's reputation in one context inevitably damages its reputation in others.

6. Anthropic has highly prioritized enterprise adoption of Claude, and the majority of our revenue comes through enterprise customer contracts. Among its offerings, Anthropic makes its Claude models available to customers via an application programming interface–also called an API–for customer integration into their own products. Some customers in turn sell these Claude-integrated products to the U.S. government. In fact, our customers sell Claude-integrated products and services that span the economy in industries including healthcare, education, financial services, manufacturing, retail, software, and energy, just to name a few.

7. The unfounded designation of Anthropic as a purported "supply chain risk" is a direct attack on the company's reputation. I understand that the statutes governing "supply chain risk" relate to foreign adversaries that may harm U.S. national security and, specifically, pose a risk to the Department's information systems, not to U.S. companies that do not present such a threat.

8. The practical effect of the Government's Actions is to brand Anthropic as akin to a foreign adversary and, in turn, to signal to all American companies that, if they work with Anthropic, they have chosen an unacceptable counterparty. This designation, should it stand, carries a powerful stigma and risks Anthropic's partnerships with all kinds of firms, not only ones that do business with the government.

9. The manner in which the government has claimed to designate Anthropic a supply chain risk significantly exacerbates the risk of reputational harm to the company. The President's social media post exclaimed that Anthropic's employees were "Leftwing nut jobs" whose "selfishness is putting AMERICAN LIVES at risk, our Troops in danger, and our National Security in JEOPARDY." The post further labeled Anthropic as an "out-of-control, Radical Left AI company run by people who have no idea what the real World is all about." Secretary Hegseth's statements echoed these false and derogatory claims, asserting that Anthropic had

"attempted to strong-arm the United States military into submission - a cowardly act of corporate virtue-signaling that places Silicon Valley ideology above American lives," and characterizing Anthropic's position on permissible uses of Claude as "fundamentally incompatible with American principles."

10. The accusations at the heart of the Government's Actions do not reflect in any way what Anthropic does or who we are. Yet these statements reinforce a false narrative that Anthropic is untrustworthy and unpatriotic, and they endeavor to signal to customers and counterparties that we are a company to be avoided.

The Government's Actions, If Allowed To Stand, Will Have Significant Consequences On Anthropic's Business Partnerships

11. The Government's Actions will have far-reaching detrimental effects on Anthropic's partnerships, not only with Department contractors that currently partner with Anthropic, but also with other Anthropic partners that work with different components of the executive branch and, more broadly, with commercial partners whose interests are unrelated to the national security sector.

12. First, the designation has already harmed our relationships with the Department's contractors, which have been integral to our development as a frontier AI lab supporting U.S. national security objectives. Anthropic was founded and has been led with the core mission of developing safe and beneficial AI that can be used to advance U.S. national security. The Government's Actions put that mission at risk. Over the last several years, our company has invested significant time and resources in building trust with national security stakeholders and demonstrating that our models can effectively serve the military and intelligence community. The government is now demanding that Anthropic sever those relationships entirely, instantly unsettling carefully cultivated partnerships in the national security arena. As just one example, Anthropic has maintained a multi-year relationship with a defense technology provider that permits Anthropic to offer Claude to Department end-users working on classified datasets. Following the Government's Actions, that defense technology provider has indicated it intends to move all U.S. government work to other generative AI model providers as soon as possible.

13. Second, the Government's Actions are likely to impair Anthropic's ability to engage with other components of the federal government and their contractors. On the same day President Trump and Secretary Hegseth issued their directives, the General Services Administration ("GSA") announced that it was removing Anthropic from USAi.gov, the centralized platform for federal agencies to access and adopt AI tools. That decision cuts off Anthropic's government procurement opportunities with numerous federal entities, including the U.S. Departments of Veterans Affairs, Health and Human Services ("HHS"), State, Labor, Interior, as well as the federal judiciary. A few days later, on March 2, 2026, Secretary of Treasury Scott Bessent announced on X.com that the Department of Treasury would terminate all use of Anthropic's AI models in compliance with President Trump's directive. The Department of State and HHS subsequently issued internal statements announcing the same. The Lawrence Livermore National Laboratory, a nuclear weapons research and development center funded by the Department of Energy, likewise informed Anthropic that it was shutting down Claude, expressing hope that it might one day be permitted to return to the facility.

14. Worse still, government contractors for other components of the federal government have been directed to cut ties with Anthropic and to stop using Claude. One partner, which has a multi-million-dollar annual contract, immediately switched from Claude to a competing generative AI model for a deployment by their end customer, the U.S. Food and Drug Administration. That switch instantly eliminated an anticipated revenue pipeline worth more than one hundred million dollars. The government told one electronics testing firm that it must stop using Anthropic for any work related to its government contract. A software security company was likewise instructed by its government client to immediately terminate access to Anthropic's AI models. Despite acknowledging that there was no legal basis for the directive—only political pressure—the company stated that it had no choice but to comply. These developments make clear that Anthropic's relationships with entities outside the national security sector are already beginning to erode as a result of the government's pressure. This harm will only intensify with the issuance of Secretary Hegseth's recent letters that attempt to implement his February 27 directive posted on X.com.

15. Third, based on my understanding, the Government's Actions attempt to expand their impacts beyond Anthropic's ability to provide services to the Department—or even other government agencies—by sowing doubt and uncertainty across Anthropic's commercial partnerships more broadly. In doing so, the Government's Actions cast a long shadow of uncertainty across our business, eroding Anthropic's carefully cultivated relationships with companies in industries including healthcare, education, financial services, manufacturing, retail, software, and energy, just to name a few, and causing our customers to question whether they can engage with Anthropic in any commercial context.

16. Indeed, this risk is already materializing. Even before the March 4 letters, in response to the government's February 27, 2026, public threats, many large enterprise customers signaled that publicly doing business with Anthropic had become more costly than working with Anthropic's competitors. This reluctance has taken several forms, including delays in contract discussions; customers and prospects declining to continue evaluating Claude alongside competing LLMs; cancelled sales meetings; withdrawal from planned co-marketing efforts and demands for new contractual protections. For example, our negotiations with three leading financial services institutions have been impacted since the Government's Actions, one of which is valued at one hundred million dollars and had been on the verge of closing. Two of the other firms have made clear that they cannot close their anticipated deals, valued together at over eighty million dollars, unless they obtain newly requested provisions allowing for unilateral contract termination. A national grocery chain cancelled a sales meeting, explicitly noting that it needed to assess the business impacts of the Administration's public statements. Customers across industries as varied as healthcare and cybersecurity have also withdrawn from joint press releases.

17. Other contract negotiations with multiple companies have likewise become more challenging. One current financial services customer paused negotiations on a contract worth fifteen million dollars while their legal team engaged in supply-chain-risk-designation "diligence." One of the world's largest pharmaceutical companies is seeking to shorten the intended duration of its contract by ten months. A current financial technology customer explicitly tied cutting a $10

million contract to $5 million, noting that "the DoW situation" had made them unwilling to commit to spending more on Claude.

18. Since the Government's Actions, several customers have begun pivoting their evaluation or deployment of LLMs from Claude to models offered by Anthropic's competitors. These competitive losses include a state higher-education system comprising more than 20 schools, a customer in the business-to-business collaboration software industry, and a telecommunications/media customer that is switching its pilot from Claude Code to a competing AI coding tool.

19. Consistent with examples provided above, many companies are also now seeking contractual provisions to protect against uncertainty, when they had not previously done so. For example, as previously noted, a household-name financial services company is now stalling negotiations over a deal valued at well over $50 million, seeking to add a termination-for-convenience provision to its contract. And a legal-technology company in the midst of a million-dollar contract negotiation sought to change the structure of its contract from committed spending to a pay-for-use structure, citing its investment in government relationships and the perceived risk created by the Government's Actions. Some have described these provisions as a precautionary measure in the event the government orders them to stop working with Anthropic. Other customers requesting these provisions have stated they would intend to exercise their newly obtained termination rights in the event of a formal government directive designating Anthropic a supply chain risk—we expect they will now do so, in light of the Government's Actions. And, from the beginning of this dispute, all have taken steps that reflect deep distrust and a growing fear of associating with Anthropic while the Government's Actions loom.

20. All told, Anthropic has received inquiries regarding the Government's Actions from over one hundred enterprise customers expressing deep fear, confusion, and doubt about Anthropic and the repercussions of associating with our company. For example, a multi-billion-dollar software company expressed uncertainty about its ability to continue using Claude because it maintains a shared codebase across work for the Department and other clients. A large customer that spends hundreds of millions of dollars annually with Anthropic asked whether Claude might

be removed from its cloud service provider, and indicated it would require significant additional technical work to serve government customers if a broad ban were imposed that prevented its end customers from using Claude. A Fortune 20 company stated that its lawyers were "freaked out" about working with Anthropic because it does significant government business. And the list goes on.

21. Importantly, this harm is extremely challenging to reverse if the designation of Anthropic persists. These relationships were built over years through sustained investment, trust, and repeated engagement. If severed, they will be extraordinarily difficult—and in some cases impossible—to rebuild, particularly for companies whose core business depends on government contracting.

* * *

I declare under penalty of perjury that the above is true and correct to the best of my knowledge.

Executed on June 10, 2026.

/s/ Paul Smith
Paul Smith
Chief Commercial Officer, Anthropic

ATTESTATION PURSUANT TO CIVIL LOCAL RULE 5-1(i)(3)

Pursuant to Civil Local Rule 5-1(i)(3), I attest that concurrence in the filing of this document has been obtained from the other signatories.

By: */s/ Michael J. Mongan*
Michael J. Mongan

MICHAEL J. MONGAN (SBN 250374)
michael.mongan@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
50 California Street, Suite 3600
San Francisco, CA 94111
Telephone: (628) 235-1000

EMILY BARNET (*pro hac vice*)
emily.barnet@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
7 World Trade Center
250 Greenwich Street
New York, NY 10007
Telephone: (212) 230-8800

Attorneys for Plaintiff Anthropic PBC

KELLY P. DUNBAR (*pro hac vice*)
kelly.dunbar@wilmerhale.com
JOSHUA A. GELTZER (*pro hac vice*)
joshua.geltzer@wilmerhale.com
KEVIN M. LAMB (*pro hac vice*)
kevin.lamb@wilmerhale.com
SUSAN HENNESSEY (*pro hac vice*)
susan.hennessey@wilmerhale.com
LAUREN MOXLEY BEATTY (SBN 308333)
lauren.beatty@wilmerhale.com
MATTHEW E. MORRIS (*pro hac vice*)
matt.morris@wilmerhale.com
BARDIA VASEGHI (*pro hac vice*)
bardia.vaseghi@wilmerhale.com
DAKOTA C. FOSTER (*pro hac vice*)
dakota.foster@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ANTHROPIC PBC,

Plaintiff,

v.

U.S. DEPARTMENT OF WAR, et al.,

Defendants.

Case No. 3:26-cv-01996-RFL

DECLARATION OF KRISHNA RAO

I, Krishna Rao, pursuant to 28 U.S.C. § 1746, declare as follows:

1. My name is Krishna Rao. I am the Chief Financial Officer ("CFO") at Anthropic PBC ("Anthropic"), where I have worked since May 2024. As CFO, I am responsible for overseeing Anthropic's financial strategy and operations, including capital allocation, financial risk assessment, investor relationships, capital raising, and ensuring that Anthropic maintains the financial discipline and operational stability necessary to support the responsible development and deployment of safe, advanced artificial intelligence ("AI") systems.

2. Before joining Anthropic, I was the CFO of Fanatics Commerce (a licensed consumer products sports platform) and the first CFO at Cedar (a healthcare payments and patient engagement platform). Prior to those roles, I served as Global Head of Corporate and Business Development and led Corporate and Operations FP&A (Financial Planning and Analysis) at Airbnb. Earlier in my career, I was a private equity investor at Blackstone and a strategy consultant at Bain & Company. I received a J.D. from Yale Law School and an A.B. in Economics from Harvard College.

3. As the CFO for Anthropic, I have personal knowledge of the contents of this declaration, or have knowledge of the matters based on my review of information and records gathered by Anthropic personnel, and could testify thereto.

The Uncertainty Created By The Government's Actions Is Concretely Harming Anthropic

4. Recent actions by the government have created significant uncertainty in the market. For example, over the weekend after the President Trump's and Secretary Hegseth's social media posts, a major investor in Anthropic informed me that the Department of War (the "Department") had contacted several of its portfolio companies about their use of Claude. Those companies have grown worried and uncertain about their ability to use Claude. A different Anthropic investor forwarded me market analysis, which called Anthropic's supposed designation as a supply chain risk a "sanction" and opined that any company that wants to serve as a federal contractor cannot do business with Anthropic. I have seen multiple client alerts from law firms describing the potentially far-reaching nature of the government's actions and suggesting that Department contractors may be best served by reevaluating their relationship with Anthropic. Public reporting

has been similarly confused. For example, a *Yahoo* article mistakenly asserts that Secretary Hegseth "ordered the Pentagon to bar its contractors *and their partners* from any commercial activity with Anthropic."[1]

5. The uncertainty the government's actions have created is already inflicting a wide range of challenges and problems on Anthropic.

6. My team has estimated revenue exposure across a range of potential customer interpretations of the government's actions. Insofar as customers adopt a narrow reading—where the government's actions are understood to prohibit only the use of Claude in work performed for the Department—we estimate that hundreds of millions of 2026 revenue is at risk. Insofar as significant swaths of customers' risk calculations sweep broader—where customers believe that doing business with Anthropic at all would jeopardize their ability to contract with U.S. agencies—the impact is significantly larger and will vary by customer type. Defense contractors and others with financial dependence on the Department are most likely to adopt the maximal interpretation, and we estimate it could reduce revenue from those customers by 50–100 percent. Across Anthropic's entire business, and adjusting for how likely any given customer is to take a maximal reading, the government's actions could reduce Anthropic's 2026 revenue by multiple billions of dollars.

7. When it comes to Anthropic's investors, even if they recognize the true scope of the relevant authorities being invoked by the government, the mere fact of the purported designation—combined with President Trump's and Secretary Hegseth's public statements—risks substantially undermining market confidence and Anthropic's ability to raise the capital critical to train next-generation models and maintain its position in a very competitive race at the AI frontier. Compounding the problem is the threat of mounting fear and uncertainty among customers. These effects reinforce one another: investors withdraw from companies losing customers, and customers avoid companies perceived as struggling to attract capital or sustain growth.

[1] Jen Judson, et al., *US bars Anthropic products from agencies, contractors for rejecting US military offer*, Yahoo Finance (Feb. 27, 2026), https://sg.finance.yahoo.com/news/us-bars-anthropic-products-agencies-213459254.html (emphasis added).

8. In a field as competitive as frontier AI, this feedback loop—if allowed to persist—could result in harm well beyond the immediate consequences of the government's actions. Training and serving frontier-level models like Claude requires extraordinary computational resources. Anthropic has already spent over $10 billion on model training and inference (serving the model to end users) and expects to spend many billions more in the coming years. Although the company has generated substantial revenue since entering the commercial market—exceeding $5 billion to date—it has nonetheless had to raise more than $60 billion in outside capital to fund its operations. Anthropic has raised this capital by issuing investors equity stakes in the company. We also have funded substantial critical technology infrastructure ("compute") purchases via long-term financing arrangements, where it is critical that our counterparties believe that Anthropic can and will repay or otherwise fulfill its financial obligations. This need for capital is inherent to the frontier AI business: although commercial adoption is continuing to grow, even companies with strong commercial traction cannot yet self-fund the required infrastructure. Investors supply this capital because they believe Anthropic's models will remain at or near the frontier.

9. The uncertainty created by the government's actions serves to undermine investors' confidence in Anthropic. Faltering investor confidence, if it goes on for long enough, will increase Anthropic's costs to raise the funds it needs to operate. And this will put us at a competitive disadvantage relative to other frontier AI labs because of an escalating inability on Anthropic's part to acquire sufficient compute for research and development, serve our customers, and satisfy investor expectations. If investors opt not to invest in Anthropic in the future, Anthropic will be unable to train the next generation of models, further eroding its commercial position and investor confidence.

10. If the government's actions are allowed to stand, and if the ripple effect described above comes to pass, it would be almost impossible to reverse.

* * *

I declare under penalty of perjury that the above is true and correct to the best of my knowledge.

Executed on June 10, 2026.

/s/ Krishna Rao
Krishna Rao
Chief Financial Officer, Anthropic

ATTESTATION PURSUANT TO CIVIL LOCAL RULE 5-1(i)(3)

Pursuant to Civil Local Rule 5-1(i)(3), I attest that concurrence in the filing of this document has been obtained from the other signatories.

By: */s/ Michael J. Mongan*
Michael J. Mongan

MICHAEL J. MONGAN (SBN 250374)
michael.mongan@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
50 California Street, Suite 3600
San Francisco, CA 94111
Telephone: (628) 235-1000

EMILY BARNET (*pro hac vice*)
emily.barnet@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
7 World Trade Center
250 Greenwich Street
New York, NY 10007
Telephone: (212) 230-8800

Attorneys for Plaintiff Anthropic PBC

KELLY P. DUNBAR (*pro hac vice*)
kelly.dunbar@wilmerhale.com
JOSHUA A. GELTZER (*pro hac vice*)
joshua.geltzer@wilmerhale.com
KEVIN M. LAMB (*pro hac vice*)
kevin.lamb@wilmerhale.com
SUSAN HENNESSEY (*pro hac vice*)
susan.hennessey@wilmerhale.com
LAUREN MOXLEY BEATTY (SBN 308333)
lauren.beatty@wilmerhale.com
MATTHEW E. MORRIS (*pro hac vice*)
matt.morris@wilmerhale.com
BARDIA VASEGHI (*pro hac vice*)
bardia.vaseghi@wilmerhale.com
DAKOTA C. FOSTER (*pro hac vice*)
dakota.foster@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ANTHROPIC PBC,

Plaintiff,

v.

U.S. DEPARTMENT OF WAR, et al.,

Defendants.

Case No. 3:26-cv-01996-RFL

DECLARATION OF THIYAGU RAMASAMY

I, Thiyagu Ramasamy, pursuant to 28 U.S.C. § 1746, declare as follows:

1. I am the Head of Public Sector at Anthropic PBC. I have held that position since January 2025. Before joining Anthropic, I worked for Amazon Web Services ("AWS"), where I was a Principal Lead for Data, Analytics, and Artificial Intelligence/Machine Learning. In that role, I was responsible for, among other things, implementing Anthropic's AI models, Claude, for AWS's public sector customers, including AWS's deployment of Claude in classified government networks. Together, my prior experience and current role give me a detailed understanding of Anthropic's relationship with the U.S. government, the technical capabilities of Anthropic's AI models, how third-party cloud providers deploy Claude to government customers, and how those customers access and integrate Claude in their workflows.

2. I have personal knowledge of the contents of this declaration, or have knowledge of the matters based on my review of information and records gathered by Anthropic personnel, and could testify thereto.

Purpose of Declaration

3. I submit this declaration in support of Anthropic's Motion for Summary Judgment. I have previously submitted multiple declarations in this matter and in related proceedings before the United States Court of Appeals for the District of Columbia Circuit. This declaration consolidates and supplements the information set forth in those prior declarations to provide the Court with a consolidated account of the relevant facts.

4. I am aware that the government has asserted that Anthropic poses a supply chain risk to national security, and that government declarants have offered various rationales in support of that designation. While those explanations have evolved over the course of these proceedings, the government's assertions continue to reflect a fundamental misunderstanding of how Anthropic's tools are deployed in classified systems and made available to the government, and they contain multiple factual misstatements.

5. I submit this declaration to explain 1) how Anthropic builds, trains, and deploys Claude and the extent of the company's ability to influence the behavior of models before and after deployment; 2) the Department of War's adoption and use of Claude in classified networks,

including through defense contractor platforms; 3) Anthropic's collaborative working relationship with the Department and U.S. national security agencies; and 4) relevant developments following the Department's supply chain risk designation.

Anthropic's Positive Relationship with the U.S. Government

6. Anthropic is a public benefit corporation whose mission is to ensure that the transition to powerful AI benefits humanity. We view partnering with the U.S. government as an important way to achieve that mission. During my time at Anthropic, we have aggressively pursued opportunities to empower the U.S. government to use our signature large language model (LLM), called Claude. Our government customers include the Department of War and agencies in the Intelligence Community.

7. Anthropic began working with the U.S. government as early as 2024. In April 2024, Anthropic became the first frontier AI lab to collaborate with the Department of Energy National Laboratories and the National Nuclear Security Administration to evaluate one of Anthropic's models in a Top Secret classified environment to determine how large language models may contribute to or help to address national security risks in the nuclear domain. Anthropic later expanded its partnership with the Department of Energy National Laboratories by deploying Claude to 10,000 scientists at Lawrence Livermore National Laboratory to help bolster research across nuclear deterrence, energy, materials science, and energy security.

8. In November 2024, Anthropic expanded its relationship with the U.S. government via a partnership with the software company Palantir Technologies. Anthropic and Palantir partnered to provide intelligence and defense capabilities to U.S. intelligence and defense agencies. That partnership has allowed Claude to be used to support government operations, including rapidly processing large datasets; autonomously completing complex software engineering projects related to offensive and defensive cyber operations and vulnerability detection; supporting military operations; performing intelligence analysis and threat assessments; and handling national security workflows and other mission-critical tasks integral to national security.

9. We have only deepened our relationship with the Department and the Intelligence Community since then. In June 2025, we announced a custom set of models built for U.S. national security customers, described further below. In July 2025, alongside other frontier AI labs including Google, OpenAI, and xAI, Anthropic was awarded a two-year, up to $200 million agreement by the Department's Chief Digital and Artificial Intelligence Office ("CDAO"), the primary office within the Department responsible for integrating and optimizing AI capabilities across the Department. As part of that agreement, we expanded our commitment to work with the Department to explore and prototype frontier AI capabilities that advance U.S. national security. In announcing the award of these contracts to Anthropic and the other frontier AI labs, the CDAO emphasized that it was "leveraging commercially available solutions" to "accelerate the use of advanced AI" in service of the Department's mission. Anthropic worked diligently under that agreement, scoping out potential ways that the Department could best be served by Claude and related Anthropic professional services. During this period, the Department conveyed to Anthropic that Claude was the best solution for some of the proposals.

10. We have also partnered with the General Services Administration ("GSA"), the civilian agency responsible for centralized procurement and shared services across the federal government. In August 2025, Anthropic and GSA announced a first-of-its-kind OneGov agreement to deliver Claude Gov to all three branches of the government—civilian executive, legislative, and judiciary—for a nominal fee of $1 per agency. As GSA announced at the time, "This trailblazing partnership directly supports the White House's America's AI Action Plan and positions the United States as the global leader in government AI adoption, ensuring that the federal workforce can tap into the transformative power of AI to modernize operations, improve decision-making, and deliver better results for taxpayers."

11. Throughout, we have maintained our commitment to supporting the national security of the United States and its allies. I have a team of over 30 individuals who manage relationships with our federal government customers, including national security customers. This team includes individuals with security clearances who drew on their past experience training AI models for the Department and Intelligence Community to help lead the development of

Anthropic's Claude Gov models. My team partners closely with dozens of other Anthropic employees across other parts of the company, including Product, Applied AI, Legal, and Policy. And recently, to help the company identify and develop high-impact applications that strengthen U.S. and close allies' capabilities in areas like cybersecurity and intelligence analysis, on August 27, 2025, we introduced the Anthropic National Security and Public Sector Advisory Council, a group of leading bipartisan national security and public policy experts.

12. As one would expect, our many partnerships with the U.S. government have involved intense security reviews and thorough vetting. Last year, the Department's Defense Counterintelligence and Security Agency granted Anthropic a Top Secret facility security clearance, after an 18-month vetting process, along with several personnel clearances for Anthropic employees and management, to enable continued support for classified national security projects.

13. In June 2025, GSA and the Department granted Claude authorization through the Federal Risk and Authorization Management Program ("FedRAMP") for use with FedRAMP High and DoD Impact Level 4 and 5 workloads, representing the highest levels of cloud security certification for unclassified and controlled unclassified information.

14. To my knowledge, at no point in any of those processes did the Department identify any potential supply chain risk posed by Anthropic, its employees, or its products and services.

15. In fact, to my knowledge, we have only ever received positive feedback about Claude's performance from our governmental customers. For example, a technology leader at a large civilian agency informed us that their agency had used Claude to resolve legacy system issues that had been stuck for years (with one five-year-old bug fixed within days), build internal tools within days rather than waiting through year-long procurement processes, and modernize applications that have not been updated since the 2000s to current technology stacks within hours. The leader of another organization shared that they named Claude their "top model" and planned to grow usage as quickly as possible: "We want to move as fast as possible with you guys." Similarly, senior leaders within a part of the Intelligence Community reported that they were "hammering away" at Claude once they obtained access to their networks.

Background On Claude

16. Like other LLMs, Claude can interpret and respond to diverse user prompts, including analyzing large volumes of text, generating written content, generating and debugging code, and reasoning through complex, multi-step problems. It can also be configured to act "agentically" by taking actions on a user's behalf—from simple tasks like sending emails to complex workflows executed with limited ongoing user direction.

17. LLMs like Claude are dual-use technology: the same capabilities that drive medical breakthroughs, accelerate scientific research, and enhance human creativity can also be misused by malicious actors, making it difficult to restrict harmful applications without undermining beneficial ones. LLMs can also produce outputs that diverge from their users' intentions or reflect errors embedded in training data.

18. To address these risks, Anthropic has implemented a multilayered safety framework operating at three levels: the model layer, the safeguards layer, and the policy layer (with the last effectuated partly through contract terms). The sections below explain how each layer functions in commercial deployments and how each is adapted for national security deployments, where many commercial-use safety mechanisms may not be appropriate. Understanding this framework is essential to understanding when and how Anthropic can influence how the model operates.

Model Layer

19. At the model layer, Anthropic embeds safety directly through training. Training an LLM requires processing large amounts of data to develop a mathematical representation of language. The product of that training process is a set of model weights, which are the numerical parameters that shape how the model responds to inputs by determining how strongly different inputs influence a model's outputs. In other words, the model weights encode Claude's learned behaviors and capabilities and govern how Claude responds to any given input.

20. One key training technique is what we call "Constitutional AI," which trains the model to evaluate and revise its outputs against a set of normative principles, such as balancing being helpful against avoiding harm. We publish "Claude's Constitution" to be transparent about those principles and how Claude has been trained to behave. We also use techniques like

reinforcement learning from human feedback, in which human reviewers rank model outputs to train the model toward responses that better align with human judgment. Through these techniques, Anthropic also embeds guardrails into the model weights, which are specific behavioral constraints that Claude is trained to follow regardless of how it is subsequently deployed or configured. For example, Claude is trained to refuse to generate child exploitation material under any circumstances.

21. Anthropic has adapted its model-layer training to account for the specific needs of national security users. As described in detail below, Anthropic engaged directly with government stakeholders to understand operational needs and develop a version of Claude specifically tailored for national security use. That model, Claude Gov, retains the same baseline alignment and safety testing as commercial models, but is trained to specifically perform in classified environments where users can be trusted with prompting the model in ways that, ordinarily, the model would refuse to be responsive to.

22. Model training is the primary mechanism through which Anthropic can influence the behavior of models used by the Department because it is the only technical safety layer that operates before deployment. Once a model is delivered to a third-party cloud provider and deployed inside a government-secure enclave, Anthropic has no ability to access it, alter it, or shut it down. Any subsequent change to the model's behavior requires preparing and delivering an entirely new version of the model, a process that requires the Department's approval and is described in detail below. Anthropic does not engage in ongoing administration of models operationally deployed by the Department.

Safeguards Layer

23. For commercial deployments, Anthropic applies a safeguards layer on top of the model at runtime. This consists of technical controls, including classifiers and probes that detect harmful activity in real time, targeted interventions to reduce harmful outputs, and monitoring systems to identify misuse at scale. Generally speaking, classifiers engage in monitoring functions or intervening functions. A monitoring classifier looks at prompts to assign and record a risk score but does not impact system behavior. An intervening classifier operates like an automated filter

that reads Claude's output before it reaches the user—if the output appears to provide instructions for synthesizing a dangerous substance, for example, the classifier triggers an intervention that blocks or rewrites the unsafe response before it is delivered.

24. While the safeguards layer is very important in Anthropic's commercial deployments, it is far less relevant when it comes to the Department. From a technical standpoint, because Anthropic does not have any ongoing access to models deployed in the Department's classified environments, it cannot operate or access any classifiers, probes, or other runtime monitoring systems. The Department's cloud provider does deploy a monitoring classifier that records risk scores for prompts (though not the text of the prompt itself), but it does not impact the behavior of the model. Anthropic does not control the monitoring classifier in such operational deployments and cannot access any risk score information. Furthermore, there are no intervening classifiers in use in Department deployments.

Policy Layer

25. The policy layer consists of Anthropic's Usage Policy, which governs how Claude may be used. Claude is available to any user only under Terms of Service that incorporate the Usage Policy. The policy defines the scope of permissible use and steers users away from applications for which Claude has not been developed or is not ready. By establishing reasonable limitations on the appropriate uses for our products and services, the Usage Policy effectively defines Anthropic's commercial offerings. For commercial users, the Usage Policy generally prohibits uses that pose unacceptable risks, including using Claude to conduct surveillance, to compromise computer systems, or to design weapons or other systems intended to cause harm or loss of human life. The Usage Policy is critical because technical safeguards alone cannot prevent all dangerous uses of such a powerful, novel technology.

26. The contractual usage restrictions at the policy layer are both substantively and technically different for Department deployments. Substantively, Anthropic has adapted its Usage Policy over time, in consultation with government partners, specifically to support national security uses. When Claude was first deployed for national security use in early 2024, the applicable Usage Policy was identical to the commercial policy. In June 2024, Anthropic issued its

first government-specific addendum to allow analysis of foreign intelligence; a subsequent January 2026 addendum further expanded permissible government uses to include offensive cyber operations and certain intelligence collection. Technologically, Anthropic does not have the ability to enforce the Usage Policy in Department operational deployments or direct visibility into how the model is being used. However, these usage restrictions are nevertheless critical to articulate the boundaries of safe usage.

27. In summary, as a practical matter, Anthropic can influence the operation of its models with respect to the Department's uses only through model training prior to operational deployment. And it can negotiate contract-based but technically unenforceable limitations on use. In both contexts, Anthropic has sought to be exceptionally transparent. It has presented its usage limitations directly to the Department and has worked collaboratively with the government in adjusting model-layer safeguards for national security purposes. Below I describe model training for national security purposes and the Department's rigorous testing, evaluation, and approval process to assess the security, capability, and behavior of models prior to operational deployment.

The Government's Adoption of Anthropic's Models in Classified Environments

28. To understand the ways in which Anthropic has used model training to shape the operation of its models in Department deployments, it is useful to step back to explain the evolution of the Department's operational deployments of Anthropic tools.

29. Anthropic began working with the federal government as early as 2024. Beginning in early 2024, Anthropic worked with a third-party cloud provider to support deployment of Anthropic's commercial AI models in classified cloud environments. By May 2024, Claude Sonnet was available on a third-party provider's Top Secret cloud and was being used by Intelligence Community ("IC") and Department customers. As government users adopted the commercial version of Claude in classified settings, Anthropic received feedback that its commercial models, appropriately trained to decline discussion of classified materials with the general public, sometimes applied those same protections to IC and Department users in secure environments. In response, in order to better assist the national security components of the U.S. government in using Claude to fulfill their mission, Anthropic engaged directly with government

stakeholders, including through classified discussions led by Anthropic engineers holding security clearances, to understand operational needs and identify appropriate technical solutions.

30. In December 2024, the updated commercial version of Claude (Sonnet 3.5) was launched in classified cloud environments. To be clear, in these arrangements, once Anthropic gives its models to the third-party cloud provider, Anthropic has no access to, or control over, the model as deployed or used by government customers.

31. In early 2025, Anthropic—on its own initiative and at its own expense—assigned cleared personnel to develop Claude Gov, a version of Claude specifically tailored for national security use and designed to operate in a high-trust classified environment. As discussed above, Claude Gov retains the same baseline alignment and safety testing as commercial models but is trained to understand what entity is making use of it and what legal framework governs that work. Specifically, Claude Gov is instructed to assume that it is working with a cleared U.S. military or intelligence officer acting under lawful authority. Claude Gov was developed based on extensive collaboration with government partners regarding applicable policies and operational use cases. It is trained to perform tasks appropriate to the national security context while remaining within the rules governing intelligence and military activities.

32. Claude Gov is a distinct version of the Claude model, specifically trained to operate appropriately in a national security context. Claude Gov does not include any additional "wrapper" layer that could provide Anthropic with access to, or control over, the model's operation or that inserts any additional safeguards or restrictions.

33. Claude Gov was made available to IC and Department customers through classified cloud environments in March 2025. Government users reported that it performed significantly better for classified uses than the commercial version, including now completing tasks that were appropriate to the national security mission—such as reviewing classified documents or supporting military planning—which it had previously refused. Adoption of Claude Gov increased rapidly across the Department and the IC. Government customers expressed strong appreciation for Anthropic's proactive approach to national-security partnerships.

34. In June 2025 and again in November 2025, Anthropic deployed new, improved versions of the Claude Gov models. Before each deployment, Department customers conducted extensive testing to confirm that the new model performed as expected before the prior version was retired and replaced, as described further below. Anthropic closely collaborated with its government and third-party partners to ensure a seamless transition throughout this process.

35. Within weeks of the November 2025 release, nearly all IC and Department customers had adopted the updated version of Claude Gov. Government users praised this model as a major technical advance, citing significant improvements in reasoning, agentic tool use, and coding. These improvements enabled customers to fully leverage air-gapped deployment of Claude Code, Anthropic's agentic coding tool, which is now widely used across the IC and Department.

The Department Tests and Approves All Models Prior to Deployment

36. Anthropic's role in deploying models on Department systems is limited to providing the model. The Department and its cloud provider conduct a rigorous, multilayered evaluation and approval process, entirely within their own systems and under their own control.

37. Before any model can be deployed in a classified environment, the third-party cloud provider that operates the relevant government region must submit a Significant Change Request ("SCR") to the relevant government region sponsor. The SCR process is a formal approval mechanism by which the cloud provider certifies that a proposed system change—including the introduction of a new model version—meets applicable security and operational requirements.

38. SCRs are classified according to complexity levels. A low-complexity SCR involves straightforward, well-understood changes with minimal risk and typically requires the fewest approvals and shortest review period. A medium-complexity SCR involves changes with a moderate degree of technical risk or interdependency and requires additional review. A high-complexity SCR involves significant system changes with broad potential impact and requires the most comprehensive review and approval process.

39. The transition of a new model from the unclassified environment (the "low side") to the classified environment (the "high side") is controlled entirely by the cloud provider.

Anthropic's involvement ends at the low side. Anthropic provides the model to the cloud provider; thereafter, the cloud provider takes responsibility for validating, transferring, and deploying the model in the government's secure environment. Anthropic personnel are not present during or after that transition and have no access to or visibility into the high-side system.

40. Once a new model version is available in the high-side environment, the cloud provider conducts its own security testing before the model may be made available to the Department. That security testing is separate from and in addition to Anthropic's own internal evaluations conducted prior to releasing the model to the cloud provider.

41. The Department also conducts its own independent evaluation of each model at the program level before approving it for operational use. That evaluation is conducted by Department personnel and is not performed or supervised by Anthropic. The Department's evaluation includes formal testing to assess model performance for the specific operational requirements of the relevant program. Such evaluation includes refusal testing—assessing whether the model correctly responds to tasks appropriate to the program's operational mission—and human-to-model comparisons, in which analysts evaluate model outputs against outputs produced by human analysts for the same prompts, providing a direct measure of model performance relative to established baselines.

42. This evaluation process equips the Department to identify and address any concern it may have about potential limitations in a model's performance. The Department has the ability to directly test any aspect of the model's behavior. For example, if the Department were concerned that the model was trained in a way that caused it to refuse tasks the Department deemed appropriate to its lawful mission, or to override the Department's judgment that an activity is permissible, the Department could directly test the model on those specific tasks. Anthropic actively recommends that the Department conduct evaluations and testing before deploying systems in operational environments—particularly for highly consequential use cases—and makes its cleared engineers available to support that testing as needed.

43. If a model failed to perform as intended, the Department could assess whether the issue could be resolved—for instance, by refining the prompts provided to the model to enable it

to correctly recognize a request as authorized. If the concern could not be resolved, the Department could decline to approve the model, either for the particular activity in question or for Department-wide use. It could also provide feedback to Anthropic and request that Anthropic address the issue in training future models—which Anthropic has done on multiple occasions.

44. The evaluation process has matured over the course of successive model deployments. Through its experience with earlier models, the Department has developed a baseline understanding of how each model performs particular tasks and maintains records of example prompts alongside the corresponding outputs produced by both human analysts and the model. Using this baseline, the Department evaluates outputs from new models, identifies instances where the model falls short of expected performance, and determines whether corrections are required. Where corrections are necessary, the Department implements them through prompt refinement—adjusting the instructions provided to the model—and related controls. Once the Department determines that a model is safe, effective, and performs as expected for the specific operational requirements of an individual program, the model is approved for use and the prior model may be retired.

45. Each Department component conducts its own testing and is not required to retire the prior version until it is satisfied that the new version meets or exceeds the prior version's performance for its specific mission requirements. For example, some components continued to use Claude Gov Sonnet 3.5 after Claude Gov Sonnet 3.7 was released because they preferred its functionality for their particular needs. By contrast, Claude Gov Sonnet 4.5 (released in November 2025) represented a significant enough advance that nearly every Department component adopted it within weeks of release.

46. Models are evaluated by testing their behavior in response to inputs. That is how the Department evaluates Anthropic's models, how the industry evaluates models generally, and how Anthropic evaluates Claude. The numerosity of model weights is irrelevant to evaluation and I am not aware of any form of AI model evaluation by the Department or commercial industry that tests a model by examining individual model weights. Doing so would be analogous to testing software

by examining every individual character of source code—of which there are also many millions. Instead, models are evaluated for their performance and behavior under a range of conditions.

47. The Department undertakes comprehensive testing of model behavior before authorizing deployments. While no testing regime could test every possible input or validate every possible execution path, this is not a limitation specific to Anthropic's models and is inherent to any AI model the Department might deploy, regardless of provider. The same is true of conventional software more broadly—including mission-critical software.

Anthropic Cannot Access or Alter Deployed Models

48. As explained above, once a model is deployed in DoW environments, Anthropic has no ability to access, alter, or shut down the deployed model. Anthropic maintains no back door or remote kill switch, and Anthropic personnel cannot log into a Department system to modify or disable the models during an operation. Anthropic does not have ongoing connectivity to, or control over, model instances running in the Department's secure networks. Only the government and its authorized cloud provider have access to the deployed system. If the government requires the assistance of any Anthropic employee with respect to the deployed model, it would need to request and facilitate that individual's access to the Department's systems directly.

49. A deployed model is static. The model weights and guardrails remain fixed unless and until a newer version is deployed to replace it. A model does not degrade or change on its own, and model weights are not automatically modified or updated based on the Department's usage. The concern that AI models may "drift" or degrade over time—leaving the Department reliant on Anthropic to ensure continued accuracy and fairness—reflects a fundamental misunderstanding of how AI models operate. Once a model is trained and deployed, its parameters remain fixed. Anthropic cannot push undisclosed or unsanctioned changes to a model after the Department has deployed it.

50. The undated "Memorandum for the Record" documenting Under Secretary Michael's analysis of the purported "Urgent Supply Chain Risk" and the March 17, 2026 declaration of Under Secretary of War Emil Michael cite the "relatively opaque nature of large language model technology" as a baseline risk requiring heightened trust in Anthropic. It is true

that large language models ("LLMs") are probabilistic, not deterministic, as Anthropic's Chief Science Officer has explained, and there is some legitimacy to DoW's concern about the opacity of these systems generally. But that characteristic applies to all LLMs. Compared to other AI labs, Anthropic is uniquely transparent. We maintain a publicly available set of normative principles—"Claude's Constitution" —a human-readable document that explains the values and rules our AI is trained to follow. Anthropic is also at the forefront of interpretability research aimed at making the behavior of models like Claude easier to understand. If anything, Anthropic presents a lower baseline risk than other AI labs, not a higher one.

51. Because Anthropic has no access to Department systems, Anthropic also cannot exfiltrate Department data or conduct surveillance of Department activities. Anthropic does not have access to the Department's Claude prompts; because Anthropic lacks any access to this customer data, there is nothing that could be exfiltrated or inspected. Any suggestion that Anthropic could engage in "data exfiltration" of Department information is unfounded.

52. The government has suggested that Anthropic could alter its models after deployment, without the Department's knowledge or consent, by modifying model guardrails or model weights—and that Anthropic is capable of making "ongoing operational alterations" to an existing model that could turn "refusals or guardrails on or off." That is incorrect. Once a model is running inside a government-secure enclave, Anthropic has no ability to unilaterally alter the guardrails or the model weights. Any such change would require Anthropic to produce an entirely new version of the model, which the Department must approve and take affirmative steps to install. The Department retains complete discretion to decline to install any new model, for any reason or for no reason at all.

53. For this reason, the term "update" is a misnomer in this context. Unlike conventional software, a model cannot be updated in place through patches or other changes pushed to a running system. The only way to alter a deployed model's behavior is to replace it with an entirely new model version. This process is analogous to ordering a cake at a restaurant: once the cake is served, the diner cannot adjust the recipe at the table, and the chef cannot reach into the dining room—secretly or otherwise—to change it either. Even if the cake is missing only a teaspoon of

baking soda, it must be made again from scratch with the correct ingredients baked in. Anthropic is the chef. If changes to model behavior are desired, whether small or large, Anthropic can prepare a new version of the model with those changes built in. But that new version is not automatically substituted. Before deployment, the third-party cloud provider and Department security reviewers inspect and approve it to ensure it is safe and appropriate. Only after that approval, and only after the customer confirms the new version performs as expected, can it retire the prior model.

54. In some instances when a model produces unwanted refusals, the Department can resolve the issue through prompt engineering—that is, by adjusting how a request is phrased so that the model responds more effectively. This is akin to trying a different phrase in a search engine when the initial phrase fails to yield the desired results. It is something all AI users do when a model does not respond to a prompt as desired, and it does not constitute altering the model itself or turning off refusals or guardrails. When basic prompt adjustments are insufficient to resolve an issue, the only option is to train and deploy an entirely new model—which is not a simple "tweak."

55. Every model version is therefore, in fact, a brand-new deployment subject to the same testing and approval processes described above. A new model version cannot be deployed without the approval of the third-party cloud provider and Department security reviewers, and the prior version is not retired until the customer confirms the new version performs as expected. Department components are free to continue using the prior version indefinitely. Neither the Department nor any Department program is compelled to adopt a new model if it is not fully satisfied with the evaluation and testing process, and the existing model will continue to perform without degradation.

56. I understand that the government has alleged that AWS provided approximately twelve cleared Anthropic engineers with administrative access to Department systems, allowing them to affect the functioning of the model by pushing updates or changing its functionality, including the ability to remove the model from the system or shut it down altogether. These

statements are incorrect. AWS has not provided Anthropic or any of its employees with administrative access to the model or with any access to the Department's operational systems.

57. Anthropic does employ certain personnel who hold security clearances. Those clearances exist solely to allow these individuals to participate in classified discussions with the Department and others regarding model functionality, to receive feedback, and to provide training to assist the Department in using the model effectively. These individuals do not have access to operational systems. Cleared Anthropic personnel would have access to the Department's operational systems only if the Department itself specifically requested and facilitated such access, and only for the limited purpose of performing functions at the Department's direction.

58. In connection with its work supporting government customers, Anthropic has undergone rigorous security vetting and compliance processes. Anthropic holds a DCSA Top Secret facility security clearance, the result of a comprehensive vetting process that took approximately eighteen months to complete. As noted above, in June 2025, Anthropic received FedRAMP High authorization and was certified at DoD Impact Level 4 and Impact Level 5—the highest classification tiers for commercial cloud services used by the Department of Defense. At no point in any of these review and certification processes did any government entity identify Anthropic as a potential supply chain risk or raise any concern about Anthropic's security posture.

The Use of Claude by Government Contractors

59. The processes and technical limitations described above apply equally when defense contractors use Claude Gov in performing Department contracts. Once Claude Gov is deployed in a classified cloud environment, it can be accessed via an application programming interface ("API"), which allows the model to be used for various purposes by cleared defense contractors. Defense contractors may use Claude Gov directly as a chatbot, use it to process their own dataflows, or integrate it into systems and platforms they develop for the military and intelligence community.

60. For example, a major defense contractor might embed Claude Gov in an intelligence fusion system used by the military that ingests satellite images and signals intercepts from a large number of sources and integrates them into a unified operational picture. Claude Gov would serve

as the analytical layer within such a system. Rather than manually reviewing reports and cross-referencing sources, a military user would be able to query Claude Gov in plain English to request summaries, ask targeted questions, or identify threats. Claude Gov could synthesize incoming data, generate written assessments, and deliver structured, actionable intelligence within seconds.

61. Claude Gov is still hosted by third-party cloud providers in these kinds of deployments, and Anthropic would not have ongoing access to the model. Anthropic could not alter the model, change its behavior, or monitor its use. All models—whether accessed directly by Department users or embedded in contractor systems—are subject to the same rigorous testing and evaluation process described above and must be confirmed to perform as expected before being authorized for use.

62. Defense contractors might also use Claude Gov to perform work on Department contracts in classified environments without embedding or integrating Claude Gov into any operational system. For example, a defense contractor might use Claude Gov to write a line of computer code for military software or edit a report being drafted as part of a Department contract. In such cases, the outputs generated by Claude Gov are entirely separate from the model itself and exist as ordinary, static data. Once Claude Gov produces an output—such as a line of code or a document—that output has no connection to the model weights, no link back to Anthropic's systems, and no mechanism by which the model or Anthropic could later access or modify the data.

The Government's Other Alleged Concerns Are Misplaced

63. I understand that the government has referenced Anthropic deployment issues at the Center for Disease Control and Prevention ("CDC") as supporting its designation. As I stated in previous declarations, to the extent this incident relates to CDC work of which I am aware, that characterization reflects a misunderstanding. My understanding is that the issue arose from the CDC's use of a general commercial model whose standard safeguards—appropriately designed for public users—limited certain interactions related to biomedical research to mitigate public safety risks. Once Anthropic became aware of the issue, we worked promptly with our third-party partners and the government to enable use of the model in a manner appropriate to the CDC's

mission and expertise. As noted above, safeguards suitable for commercial deployments may not be appropriate for government uses, and some government deployments require customization or parameter adjustments. That collaborative, iterative process—balancing safety in commercial settings with valid government uses—is a core tenet of Anthropic's AI safety approach. That is precisely how we addressed the CDC issue once identified, supporting important CDC research while preventing individual users from engineering dangerous biological weapons.

64. The government's declarations also suggest that Anthropic's employment of foreign nationals increases "adversarial risk." Anthropic, like many cutting-edge AI companies, employs a globally diverse team of highly skilled researchers and engineers. Anthropic has undergone extensive U.S. Government security vetting in connection with its classified work and takes seriously its obligation to comply with all security and technical requirements for safeguarding sensitive government information. As stated above, to my knowledge, Anthropic is the only AI company where cleared personnel have led the development of AI models tailored to be deployed in classified environments.

The Department's Supply Chain Risk Designation

65. The government's supply chain risk designation arose in the context of a broader dispute over usage terms. Beginning in September 2025, as Anthropic was negotiating its deployment on the Department's genAI.mil platform, the Department began demanding that Anthropic modify its usage policy as a condition of participation. As I have described above, Anthropic's usage policies do not give Anthropic operational control over government users—they reflect Anthropic's own principles regarding appropriate use and cannot be enforced against the Department once the model is deployed on Department systems. And of course, if Anthropic's Usage Policy restrictions do not meet the Department's needs, the Department is able to use any other AI system that better meets its requirements. Nonetheless, Anthropic indicated it was prepared to modify its usage policy to accommodate the Department's concerns, and was willing to remove essentially all usage restrictions for Department purposes. Anthropic retained only two narrow carve-outs: restrictions on the use of Claude to develop lethal autonomous weapons systems and to enable the mass domestic surveillance of United States persons. The Department

rejected Anthropic's proposed modifications and insisted on language permitting "all lawful uses." I understand that the Department has been or is exerting similar pressure on other leading AI labs to agree to similar demands, as reflected in a memorandum Secretary Hegseth issued on January 9, 2026. I further understand that many of these labs have agreed to the Department's preferred usage terms.

66. Prior to the supply chain risk designation, neither the Department nor any other government customer had ever identified Anthropic as a supply chain risk, raised supply chain concerns in any security review or procurement process, or expressed any concern about Anthropic's security posture. To the contrary, Anthropic's government customers—including those in the Intelligence Community—had consistently expressed strong appreciation for Claude Gov's capabilities, praised Anthropic as a trusted national security partner, and enthusiastically adopted new model versions. The supply chain risk designation was issued only after the usage term negotiations reached an impasse, and the record does not reflect any identified supply chain concern that preceded or independently motivated the designation.

67. In addition, neither I nor anyone else at Anthropic had seen Under Secretary Michael's memorandum or declaration until the Department of Justice publicly filed them in this matter. Those materials reflect a fundamental misunderstanding of how Anthropic's tools are deployed in classified systems and otherwise made available to the government, and they contain multiple factual misstatements. Had the DoW disclosed its rationales before designating Anthropic a supply chain risk, I and other Anthropic subject-matter experts could have corrected these errors.

68. On February 27, 2026, the government took a series of formal actions that effectively severed Anthropic's relationship with the Department. A post attributed to President Trump on Truth Social directed all government agencies to cease use of Anthropic's products. That same day, Secretary Hegseth issued a formal supply chain risk designation identifying Anthropic as a supply chain risk to national security. The accompanying directive stated that, effective immediately, no contractor, subcontractor, or other commercial entity performing work for or on behalf of the Department may conduct any commercial activity with Anthropic, subject to a six-month period for winding down existing contractual obligations.

The Department's Designation Has Affected Anthropic's Contracts, Reputation, And Mission

69. The Department's Chief Digital and Artificial Intelligence Office ("CDAO") awarded Anthropic in July 2025, alongside Google, OpenAI, and xAI, a two-year, up to $200 million agreement. The award expressly emphasized the Department's interest in leveraging commercially available solutions, and Anthropic worked diligently over the following months to assess and identify ways its services could best support the Department's mission. The Department communicated that it was interested in moving forward with Anthropic's recommendations, and conveyed that Claude was the best available option for certain of its proposed use cases. The arrangement was cancelled after only a few months as a result of the usage term negotiations. Absent those negotiations and the resulting supply chain risk designation, it is almost certain that Anthropic would have continued to develop its relationship with CDAO and received a substantial portion of the full award.

70. As a result of the usage term negotiations and supply chain designation, Anthropic also lost the opportunity to be deployed on genAI.mil, the Department's secure, official generative AI platform to provide its personnel with access to cutting-edge AI models. Anthropic began negotiations with the Department for this deployment in September 2025. As a result of the usage term negotiations and subsequent supply chain risk designation, Anthropic was no longer in consideration to be deployed on the Department's genAI.mil platform. Two other AI vendors have instead deployed their AI models on the Department's platform—one in December 2025 and another in February 2026.

71. In August 2025, Anthropic entered into a first-of-its-kind, government-wide contract with the General Services Administration under the GSA's OneGov program. That contract made Claude available to agencies across all three branches of the federal government for a nominal fee of $1 per agency, as part of the White House's America's AI Action Plan initiative to expand access to commercial AI capabilities across the government. Anthropic expected the OneGov contract to generate approximately $100 million in annual recurring revenue in 2026. Following the February 27 government actions, Anthropic was removed from the government's USAi.gov

portal and from the Multiple Award Schedule through which the OneGov contract was offered. This decision cut off sales opportunities for many federal agencies, including the U.S. Departments of Veterans Affairs, Health and Human Services ("HHS"), State, Labor, and Interior, in addition to the federal judiciary and many state and local governments that procure through the Multiple Award Schedules. The expected revenue from the OneGov agreement is now entirely at risk.

72. Other agencies also took action in response to the February 27 government actions. On Monday, March 2, 2026, the U.S. Department of the Treasury and the Federal Housing Finance Agency (which oversees Fannie Mae and Freddie Mac) announced they were terminating all use of Claude. A technology leader at a federal civilian agency has also informed me that the DoW advised his agency, and is advising all other civilian agencies, to stop using Claude. We also understand that other agencies, including the Department of State and HHS, have issued internal statements saying they will follow the President's directive. The AI leadership of one portion of the Intelligence Community informed us that they were preparing for "complete detachment" from Claude based on the "directive" they had received. The Lawrence Livermore National Laboratory, a nuclear weapons research and development center funded by the Department of Energy, also informed Anthropic that it was shutting down Claude. Many of these customers—and others, particularly in the Intelligence Community—have continued to express how much they value their partnership with Anthropic and how harmful losing access to our models will be (setting back their work months or even years). They have also expressed, however, feeling like they have little choice in the matter.

73. As of March 24, 2026, the following agencies, or sub-agencies thereof, had contracts with Anthropic or used Anthropic's technology through a third-party provider, but indicated they would be terminating the contracts or terminating uses of Anthropic's technology after February 27, 2026:

a) U.S. Department of War

b) U.S. Department of State

c) U.S. Department of Health & Human Services

d) U.S. Office of Personnel Management[1]

e) U.S. Nuclear Regulatory Commission[2]

f) U.S. Department of Treasury

74. As of March 2026, the following agencies, or sub-agencies thereof, had contracts to use Anthropic's technology or used Anthropic's technology through a third-party provider:

a) U.S. Department of Commerce[3]

b) U.S. Social Security Administration

c) U.S. Department of Homeland Security

d) Securities and Exchange Commission

e) National Aeronautics and Space Administration

f) U.S. Department of Energy

g) Federal Reserve Board of Governors

h) National Endowment for the Arts

i) Federal Housing Finance Agency

j) U.S. Department of Veterans Affairs

75. The supply chain risk designation has also frustrated Anthropic's contracts with prime contractors that perform services for the Department. For example, Anthropic has maintained a multi-year relationship with a defense technology provider that permits Anthropic to offer Claude to the Department for classified purposes. Anthropic also works closely with a cloud service provider that offers Claude to government agencies, including the Department. Both partners have indicated that, pursuant to Department leadership directives, they need to discontinue using Anthropic as a subcontractor on Department contracts and have begun doing so.

[1] I understand that the Office of Personnel Management sent a message requesting its personnel not to use Claude following the President's social media post.

[2] Anthropic was informed that the Nuclear Regulatory Commission was no longer deploying Claude, citing the President's social media post.

[3] Within the U.S. Department of Commerce, the International Trade Administration has indicated that it was terminating use of Anthropic's technology, whereas the U.S. Patent and Trademark Office and National Telecommunications and Information Administration had contracts with Anthropic or used Anthropic's technology as of March 2026 but have not announced termination actions.

Anthropic has invested significant time and resources in building relationships with these national security stakeholders and the supply chain risk designation demands that those relationships be severed.

76. The supply chain risk designation has caused both significant financial harm and concrete, ongoing damage to Anthropic's reputation and customer relationships that will be difficult to undo. In terms of financial impact, Anthropic had experienced approximately a fourfold increase in annual recurring revenue ("ARR") from December 2025 to January 2026, and had projected several hundred million dollars in public sector ARR for 2026—a projection that had been revised upward before the designation. The Department's designation and related government actions place an estimated $150 million or more in immediate Department ARR at risk, including approximately $50 million attributable to the CDAO contract alone, and expose Anthropic to over $500 million in total 2026 ARR losses when accounting for the GSA OneGov contract, state and local government customers, and other government-adjacent relationships that have been disrupted. Beyond the direct financial consequences, Anthropic depends on an interconnected network of relationships—with government agencies, prime contractors, commercial partners, cloud providers, and investors—and its reputation underpins each of them. The Department of War's public designation of Anthropic as a supply chain risk signals to enterprise customers and counterparties that Anthropic is a company to be avoided, a harm that becomes increasingly difficult to unwind the longer the designation remains in effect.

77. Anthropic is a company that has invested deeply in its relationship with the federal government, and with the IC and the Department in particular. Supporting the national security of the United States and allied democracies has been a deliberate choice—one we have made because it is central to our mission of developing AI for the benefit of humanity. While we are proud of our commitment to our founding values under these difficult circumstances, and are gratified by indications of some public support, we certainly do not welcome this dispute, which we long sought to avoid, including through good-faith negotiations with the Department. While we respect that decisions about the use of AI in military applications belong to the Department, we are harmed by the lost opportunity to continue to support the national security mission directly, and

we believe the country and its national security are harmed as well. Notably, we developed these tools specifically for use by the national security components of the federal government at our own initiative and at significant expense. Being branded a threat to the national security of our own country therefore inflicts more than financial harm.

Anthropic's Ongoing Government Engagement

78. Despite the supply chain risk designation, Anthropic has continued to work collaboratively with the U.S. government to promote national security interests, including by supporting the Department's use of Claude during the six-month transition period.

79. On March 26, 2026, the Northern District of California issued a preliminary injunction preventing the government from implementing elements of its supply chain risk designation, but did not require the government to continue using any Anthropic products. Following the injunction, although the designation continues to create uncertainty and hesitancy to build long-term partnerships, we have worked with government customers seeking to resume use of Claude to support their restore access as requested.

80. We have also worked closely with the U.S. government in connection with the release of Anthropic's new Mythos model. On April 7, 2026, Anthropic publicly announced Project Glasswing, an initiative to deploy Mythos for defensive cybersecurity work in partnership with over fifty major technology companies and critical infrastructure operators. The initiative is designed to apply Mythos's capabilities to defensive purposes before similar capabilities can be adopted by adversaries. Before making Mythos externally available, Anthropic briefed senior officials across multiple government agencies on the model's capabilities, including its ability to identify and help remediate critical vulnerabilities in widely used software systems. Through Project Glasswing, Anthropic extended access to Mythos to select government partners and committed substantial resources, at no cost to national security customers, to efforts to secure critical software infrastructure.

81. On April 17, Anthropic CEO Dario Amodei met with senior White House officials, including the Chief of Staff and the Treasury Secretary, to discuss the responsible use of Mythos. I understand that the White House described those discussions as "productive and constructive" and

stated that the parties "discussed opportunities for collaboration, as well as shared approaches and protocols to address the challenges associated with scaling this technology." Since then, Anthropic has remained in frequent communication with representatives of multiple agencies and departments regarding how best to use Mythos to advance U.S. objectives. My team has, over the past several months, worked diligently to support government national security requests and to make clear that we stand ready to assist the government in advancing U.S. national security, as we have always done.

82. These engagements reflect Anthropic's continued commitment to partnering with the U.S. government to strengthen the nation's cyber defenses and support national security objectives—even as the supply chain risk designation remains in place.

* * *

I declare under penalty of perjury that the above is true and correct to the best of my knowledge.

Executed on June 10, 2026.

/s/ Thiyagu Ramasamy
Thiyagu Ramasamy
Head of Public Sector, Anthropic

ATTESTATION PURSUANT TO CIVIL LOCAL RULE 5-1(i)(3)

Pursuant to Civil Local Rule 5-1(i)(3), I attest that concurrence in the filing of this document has been obtained from the other signatories.

By: */s/ Michael J. Mongan*
Michael J. Mongan

MICHAEL J. MONGAN (SBN 250374)
michael.mongan@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
50 California Street, Suite 3600
San Francisco, CA 94111
Telephone: (628) 235-1000

EMILY BARNET (*pro hac vice*)
emily.barnet@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
7 World Trade Center
250 Greenwich Street
New York, NY 10007
Telephone: (212) 230-8800

Attorneys for Plaintiff Anthropic PBC

KELLY P. DUNBAR (*pro hac vice*)
kelly.dunbar@wilmerhale.com
JOSHUA A. GELTZER (*pro hac vice*)
joshua.geltzer@wilmerhale.com
KEVIN M. LAMB (*pro hac vice*)
kevin.lamb@wilmerhale.com
SUSAN HENNESSEY (*pro hac vice*)
susan.hennessey@wilmerhale.com
LAUREN MOXLEY BEATTY (SBN 308333)
lauren.beatty@wilmerhale.com
MATTHEW E. MORRIS (*pro hac vice*)
matt.morris@wilmerhale.com
BARDIA VASEGHI (*pro hac vice*)
bardia.vaseghi@wilmerhale.com
DAKOTA C. FOSTER (*pro hac vice*)
dakota.foster@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ANTHROPIC PBC,

Plaintiff,

v.

U.S. DEPARTMENT OF WAR, et al.,

Defendants.

Case No. 3:26-cv-01996-RFL

DECLARATION OF SARAH HECK

I, Sarah Heck, pursuant to 28 U.S.C. § 1746, declare as follows:

1. I am the Head of Policy at Anthropic PBC ("Anthropic"), where I have worked since June 2024. Before joining Anthropic, I held senior roles at Stripe, the White House's National Security Council, and the U.S. Department of State.

2. In my role as Head of Policy, I lead Anthropic's public policy engagement, government relationships, strategic partnerships, and government communications, including engagements and initiatives that address the intersection of artificial intelligence ("AI"), national security, and economic policy. This includes being involved in communications between Anthropic and the government.

3. As the Head of Policy, I have personal knowledge of the contents of this declaration, or have knowledge of the matters based on my review of information and records gathered by Anthropic personnel, and could testify thereto.

4. I have previously submitted two declarations in this matter: a declaration dated March 9, 2026, filed in support of Anthropic's Motion for a Preliminary Injunction, and a supplemental declaration filed in support of Anthropic's Reply Brief in Support of Plaintiff's Motion for a Preliminary Injunction. I submit this consolidated declaration, which sets forth in a single, comprehensive account all of the facts from those two prior declarations, in order to assist the Court.

Anthropic Has Maintained A Strong Relationship With The Federal Government

5. Partnerships with private and public sector entities are central to Anthropic's mission, business model, public policy goals, and overall success. Our partners range from Fortune 500 companies and U.S. government agencies to small businesses and local municipalities. These partnerships are core to Anthropic and essential to advancing its policy objectives of promoting the responsible deployment of AI, supporting democratic institutions, and ensuring that powerful AI systems are developed and used safely and in ways that benefit society.

6. One of Anthropic's most important partnerships is with the federal government. Anthropic has consistently supported the U.S. government's goal of maintaining global AI dominance and its efforts to ensure that American AI systems are widely adopted across the

federal government and throughout the private sector at home and abroad. Anthropic publicly supported the Trump administration's efforts to promote AI adoption throughout the federal government as part of its AI Action Plan for America. Anthropic has also partnered with the National Institute of Standards and Technology's Center for AI Standards and Innovation (CAISI) to undertake collaborative research evaluating and mitigating safety risks. It has advocated in support of the bipartisan Future of AI Innovation Act, which supports CAISI's initiatives and promotes strong partnerships between private and public stakeholders to advance AI research and innovation.

7. Importantly, in developing its relationship with the government, Anthropic has strived to build a reputation as a nonpartisan advocate dedicated to building a safer AI ecosystem. In addition to supporting the bipartisan Future of AI Innovation Act, Anthropic recently donated over $20 million to Public First Action, a bipartisan nonprofit co-founded and co-led by Republican and Democratic former lawmakers, which supports public education about AI, promotes safeguards, and works to ensure America leads in the AI race. Anthropic has also actively engaged with bipartisan legislative efforts on Capitol Hill, including advocacy in support of the CREATE AI Act of 2025 and the GAIN Act of 2025—both bipartisan safety bills that align with the company's policy priorities. And the company deliberately maintains a bipartisan lobbying effort, with an in-house team that includes former senior staffers from both sides of the aisle and a roster of Republican-aligned and Democratic-aligned outside lobbying firms. Anthropic invests in and undertakes this public policy work because it is committed to AI safety and the policies that underpin it. Based on my experience, collaboration across the political spectrum is critical to policy progress.

8. Anthropic has also been committed to developing its AI systems to advance U.S. national security interests and has partnered with national security components of the government, including the Department of War (the "Department") and intelligence agencies, to do so. To that end, Anthropic also formed a National Security and Public Sector Advisory Council composed of former U.S. government senior defense and intelligence officials, in part to strengthen the company's understanding of government needs and become a more effective partner.

9. Throughout my engagements with the federal government in my capacity as Anthropic's Head of Policy, various officials and government staff have repeatedly conveyed that the government values its partnership with Anthropic and acknowledged that competitors' AI models lag behind Anthropic's capabilities. A wide range of senior intelligence community officials have emphasized to us that Claude's capabilities are unique, and that Claude is "widely recognized as the leading model for coding and autonomous tool use." Department of War staff themselves have described Claude as "far and away the best model" in briefings attended by the Policy team and warned that losing this capability would set the Department back several years. These characterizations are consistent with the ones set forth in the declarations of Thiyagu Ramasamy.

10. Anthropic appreciates this recognition and has, in turn, repeatedly and publicly expressed its willingness to continue providing Claude for the full range of lawful national security applications, subject to two narrow usage restrictions that reflect the company's expert, considered judgment on the nature of the AI model and its safe and reliable use: namely, restrictions on the use of Claude for mass surveillance of Americans and for lethal autonomous warfare.

Secretary Of War Pete Hegseth Issued An Ultimatum Threatening Consequences If Anthropic Adhered To Its Fundamental Commitments

11. Over the months that preceded the initiation of this litigation, Anthropic and the Department had been discussing an agreement to continue their partnership. I understand that the crux of the discussions focused on two specific limitations that would restrict the Department's ability to use Claude for purposes of mass surveillance of Americans and lethal autonomous warfare. During that time, I engaged with the Department on behalf of Anthropic and was included in discussions between Anthropic CEO Dario Amodei and Emil Michael, Under Secretary of War for Research and Engineering and Chief Technology Officer. The conversations throughout this process remained respectful, and the parties remained committed to finding a path forward to harness Anthropic's technology to fulfill the Department's important national security goals.

12. Both parties agreed that neither wanted to enter into a begrudging partnership. At one point, Dr. Amodei stressed to Under Secretary Michael that if Anthropic's position on these two guardrails meant that Anthropic was not the right vendor for the Department's needs, then he would respect that decision. Dr. Amodei further emphasized that, if Anthropic and the Department failed to come to an agreement, Anthropic stood ready to assist in an orderly offboarding from the Department's systems.

13. On February 24, 2026, I attended a meeting held between Secretary of War Pete Hegseth and Dr. Amodei, among others. At that meeting, Secretary Hegseth said that Claude had "exquisite capabilities." He then objected to Anthropic's unwillingness to remove the two usage restrictions noted above.

14. During the meeting, Dr. Amodei reiterated that Anthropic maintains a strong partnership with national security agencies, including the Department and intelligence community. He stated that Anthropic has had no intention of dictating national security operations. Dr. Amodei emphasized that the company's commitment to the safe use of its AI models, as reflected in the agreement, has permitted—and would continue to permit—the Department to use Anthropic's products for all lawful uses, save for those two important exceptions. Dr. Amodei noted that these two exceptions have never obstructed the Department's operations to his knowledge, and Combatant Commanders he has spoken to have been pleased with Anthropic's partnership and models.

15. As to autonomous uses, Dr. Amodei clarified that, while Anthropic has been open to certain autonomous applications of its technologies, at this stage, AI systems cannot yet capably or reliably perform lethal autonomous warfare without appropriate oversight.

16. Secretary Hegseth stated that Dr. Amodei's concerns were "understandable" and further stated that "they don't do mass domestic surveillance." While Secretary Hegseth expressed that the Department "would love to work with [Amodei]," he stated that the Department had many other vendors to choose from that have never raised these types of concerns and would never hold any veto power over the Department.

17. At the end of the meeting, Secretary Hegseth presented Anthropic with what appeared to be an ultimatum: if Anthropic did not agree to "all lawful uses" of its system by 5 p.m. on Friday, February 27, 2026, the Department would issue a statement at 5:01 p.m. that they would designate Anthropic a supply-chain risk—preventing Anthropic from partnering with the Department, anyone affiliated with the Department, or any other agency—or invoke the Defense Production Act to compel Anthropic to comply with the Secretary's demands.

18. During this meeting, Secretary Hegseth never stated that Anthropic's AI models were unsafe, much less subject to compromise by a foreign adversary. I am unaware of anyone at the Department ever suggesting Anthropic's models are somehow insecure or have been compromised. Moreover, I understand that these two restrictions have never previously been a barrier to the Department's adoption and use of our models to date. Nor has anyone else articulated such a barrier to me.

19. Additionally, it is my understanding that the Department had previously expressed concerns that Anthropic had objected—after the fact—about how Anthropic's models were being used in Department operations, concerns now reprised in the government's materials. But this concern was and is misplaced, as Anthropic has made very clear. Indeed, when Secretary Hegseth raised this topic at the February 24 in-person meeting, Dr. Amodei clarified that the company had not objected to the use of its tools as reported. Dr. Amodei further explained that Anthropic has no interest in interfering with the Department's operations more generally. Dr. Amodei publicly reiterated that stance in a February 26 blog post, explaining that Anthropic "understands that the Department of War, not private companies, makes military decisions" and underscoring that Anthropic has "never raised objections to particular military operations nor attempted to limit use of our technology in an ad hoc manner."

20. Conversations continued even after Dr. Amodei's meeting with Secretary Hegseth. They remained cordial and amicable. Dr. Amodei continued to seek a resolution prior to Secretary Hegseth's deadline during the afternoon of Friday, February 27, 2026, during which he provided Under Secretary Michael his redline edits on the Department's latest offer, alongside a detailed explanation of the same, still attempting to engage in earnest while ensuring that Anthropic's two

guardrails core to its principles would be retained. Under Secretary Michael did not respond in kind with written edits. At that same time, President Trump directed all agencies, by posting on Truth Social, to cease use of Anthropic's AI system immediately. Shortly thereafter, Secretary Hegseth posted on X.com designating Anthropic a supply chain risk. Throughout these discussions, Dr. Amodei remained firm in expressing Anthropic's unmovable guardrails as essential to the Company's mission to offer safe, reliable AI tools.

Secretary Hegseth Notified Anthropic Of His "Supply Chain Risk" Designation

21. The following week, as agencies across the federal government moved to implement the Presidential Directive, Dr. Amodei continued negotiating in good faith with senior Department officials. Those discussions were still ongoing when, at 8:48 p.m. ET on March 4, Dr. Amodei received a letter from Secretary Hegseth, expanding on the Secretarial Order's "supply chain risk designation." That letter (the "Secretarial Letter"), dated March 3, asserted that the Department of War had "determined" that Anthropic's technology "presents a supply chain risk" and that exercising the authority granted by 10 U.S.C. § 3252 against Anthropic is "necessary to protect national security." The letter pronounced that this determination covers all Anthropic "products" and "services," including any that "become available for procurement." And it asserted that "less intrusive measures are not reasonably available" to mitigate the risks that Anthropic's products and services supposedly pose to national security.

22. In the days that followed, the parties continued to work towards an agreement even while the Department was, unbeknownst to Anthropic, drafting materials claiming the company poses a risk of sabotage, data exfiltration, and other serious harms. At no point in these conversations did Under Secretary Michael or anyone at the Department share with Anthropic its apparently grave concerns or suggest that they were an impediment to a deal. In fact, the Department and Anthropic were still engaged in negotiations on March 4, 2026—the day after the Department completed its formalization of the supply-chain risk designation (but before Anthropic received that formal notice).[1]

[1] Attached hereto as **Exhibit 1** is a true and correct copy of a March 4, 2026 email from Emil Michael to Dario Amodei.

The Government's Filings Mischaracterize Anthropic's Positions

23. In reviewing the materials the government has filed in this case, I was struck by its repeated claim that Anthropic insisted during negotiations that the company have some sort of approval role in the Department's operational decision chain. That is false. At no time during Anthropic's negotiations with the Department did I or any other Anthropic employee state that the company wanted that kind of role. We do not want it now, either.

24. Likewise, at no point during the negotiations did any party raise any concerns about Anthropic taking technical steps to disrupt the U.S. military. This is a new purported concern that was raised, for the first time, in the government's filings in this case. Had this concern been raised in a manner allowing Anthropic to respond, we would have explained, among other things, the technical impossibility of such disruption. As I understand from Thiyagu Ramasamy's declaration, Anthropic is unable to alter an AI model after deployment without the Department's knowledge or consent.

25. All of this was avoidable. In an effort to be responsive to the Department's apparent concerns about operational interference, Anthropic was willing to make contractual commitments. In a draft agreement transmitted to the Department on March 4, 2026, Anthropic included language stating: "For the avoidance of doubt, [Anthropic] understands that this license does not grant or confer any right to control or veto lawful Department of War operational decision-making." Anthropic included this language to clarify that it did not seek to be, and would not be, part of any operational decision chain.[2]

26. I was also surprised to read in the government's materials that Anthropic's beliefs that Claude should not be used for autonomous lethal warfare or mass surveillance of Americans are part of what makes the company a danger to national security. I find this rationale puzzling for two reasons: (1) when negotiations broke down, the parties were very near agreement on language that would address Anthropic's concerns regarding autonomous lethal warfare and (2) Secretary Hegseth himself referred to Anthropic's concerns regarding autonomous lethal warfare and mass

[2] Attached hereto as **Exhibit 2** is a true and correct copy of a March 4, 2026 email from Anthony Fuscellaro to Dario Amodei Re: Official Correspondence from the Department of War to Anthropic.

surveillance of Americans as "understandable." That was echoed mere days later by the Commander of U.S. Central Command, who told a media outlet, "Humans will always make final decisions on what to shoot and what not to shoot." And a recent threat assessment from the Office of the Director of National Intelligence, reflecting the collective views of the U.S. Intelligence Community, further confirms that at least some portions of the Administration believe "it is essential to make sure that humans maintain control of the [AI models] and how AI is used."[3] Indeed, on March 4, 2026—after the Secretarial Letters were drafted—Under Secretary Michael emailed Dr. Amodei to say "I think we are very close here" regarding these two issues.

27. In recent weeks, Anthropic has worked closely with the U.S. government in connection with the release of its new Mythos model. On April 7, 2026, Anthropic publicly announced Project Glasswing, an initiative to deploy Mythos for defensive cybersecurity work in partnership with over fifty major technology companies and critical infrastructure operators. The initiative is an effort to put Mythos's capabilities to work for defensive purposes before similar capabilities can be adopted by adversaries. Prior to making Mythos available externally, Anthropic briefed senior officials across multiple government agencies on the model's capabilities, including its ability to identify and help remediate critical software vulnerabilities in widely used systems. Through Project Glasswing, Anthropic extended access to Mythos to select government partners and committed substantial resources to support efforts to secure critical software infrastructure. On April 17, Dr. Amodei met with senior White House officials, including the Chief of Staff and the Treasury Secretary, to discuss the responsible use of Mythos. I understand that the White House described those discussions as "productive and constructive" and stated that the parties "discussed opportunities for collaboration, as well as shared approaches and protocols to address the challenges associated with scaling this technology."[4] Since then, we have been in frequent communication with representatives of various agencies and departments about the best way to

[3] Off. of the Dir. of Nat'l Intelligence, *Annual Threat Assessment of the U.S. Intelligence Community* 12 (Mar. 2026), https://www.dni.gov/files/ODNI/documents/assessments/ATA-2026-Unclassified-Report.pdf.

[4] Brendan Bordelon et al., *White House Meets with Anthropic CEO amid Hopes for a Truce,* Politico (Apr. 17, 2026, 2:59 PM EDT), https://www.politico.com/news/2026/04/17/white-house-to-meet-with-anthropic-ceo-as-mythos-anxiety-spreads-00878960.

use Mythos to serve the objectives of the United States. These engagements reflect Anthropic's continued commitment to partnering with the U.S. government to strengthen the nation's cyber defenses and support national security objectives—even as the supply chain risk designation remains in place.

* * *

I declare under penalty of perjury that the above is true and correct to the best of my knowledge.

Executed on June 10, 2026.

/s/ Sarah Heck
Sarah Heck
Head of Policy, Anthropic

ATTESTATION PURSUANT TO CIVIL LOCAL RULE 5-1(i)(3)

Pursuant to Civil Local Rule 5-1(i)(3), I attest that concurrence in the filing of this document has been obtained from the other signatories.

By: */s/ Michael J. Mongan*
Michael J. Mongan

EXHIBIT 1

Gmail

Sarah Heck <heck@anthropic.com>

Redline

Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> Wed, Mar 4, 2026 at 9:24 AM
To: Dario Amodei <dario@anthropic.com>
Cc: "sheck@anthropic.com" <sheck@anthropic.com>, "Kliger, Gavin I CIV (USA)" <gavin.i.kliger.civ@mail.mil>

Hi Dario,

After reviewing with our attorneys and seeing your last draft (thanks for being fast), I think we are very close here. I was able to take 'unlawful' out of the bulk collection section which essentially reverts that clause to your language from last night. Therefore, the only change we would require would be "in accordance with" as opposed to "only as permitted under." See option 1 in the attached. We believe this goes above and beyond and we have made significant concessions. I hope this work as I am running out of time.

[Quoted text hidden]

3.4 Usage Term FINAL.docx
2837K

EXHIBIT 2

From: **Fuscellaro, Anthony COL USARMY OSW SECWAR (USA)** anthony.fuscellaro1.mil@war.mil
Subject: Official Correspondence from the Department of War to Anthropic
Date: March 4, 2026 at 5:49 PM
To: dario@anthropic.com
Cc: jbleich@anthropic.com, Matthews, Earl G HON (USA) earl.g.matthews.civ@mail.mil

Mr. Amodei,
Please see the attached letters as official notification from the Department of War regarding Anthropic. Hard copies will follow via official mail.

Please acknowledge receipt and direct all responses through the Department of War General Counsel, Honorable Earl Matthews. Thank you.

v/r,
COL Tony Fuscellaro, USA
Executive Secretary
Department of War

Notice to Anthropic Pursuant to Title 10 U.S.C. Section 3252...
1.9 MB

Notice to Anthropic Pursuant to Title 41 U.S.C. Section 4713...

MICHAEL J. MONGAN (SBN 250374)
michael.mongan@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
50 California Street, Suite 3600
San Francisco, CA 94111
Telephone: (628) 235-1000

EMILY BARNET (*pro hac vice*)
emily.barnet@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
7 World Trade Center
250 Greenwich Street
New York, NY 10007
Telephone: (212) 230-8800

Attorneys for Plaintiff Anthropic PBC

KELLY P. DUNBAR (*pro hac vice*)
kelly.dunbar@wilmerhale.com
JOSHUA A. GELTZER (*pro hac vice*)
joshua.geltzer@wilmerhale.com
KEVIN M. LAMB (*pro hac vice*)
kevin.lamb@wilmerhale.com
SUSAN HENNESSEY (*pro hac vice*)
susan.hennessey@wilmerhale.com
LAUREN MOXLEY BEATTY (SBN 308333)
lauren.beatty@wilmerhale.com
MATTHEW E. MORRIS (*pro hac vice*)
matt.morris@wilmerhale.com
BARDIA VASEGHI (*pro hac vice*)
bardia.vaseghi@wilmerhale.com
DAKOTA C. FOSTER (*pro hac vice*)
dakota.foster@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ANTHROPIC PBC,

Plaintiff,

v.

U.S. DEPARTMENT OF WAR et al.,

Defendants.

Case No. 3:26-cv-01996-RFL

DECLARATION OF MICHAEL J. MONGAN IN SUPPORT OF PLAINTIFF ANTHROPIC PBC'S MOTION FOR SUMMARY JUDGMENT

Judge: Hon. Rita F. Lin

I, Michael J. Mongan, pursuant to 28 U.S.C. § 1746, declare as follows:

1. I am a partner at the law firm Wilmer Cutler Pickering Hale and Dorr LLP ("WilmerHale"). I represent Plaintiff Anthropic PBC ("Anthropic") in this matter.

2. I am a member in good standing of the State Bar of California.

3. I have personal knowledge of the factual matters asserted herein, including based on my review of the documents, websites, and other items referenced herein. If called as a witness, I could and would testify competently thereto.

4. I submit this declaration in support of Plaintiff's Motion for Summary Judgment.

5. Attached hereto as **Exhibit 1** is a true and correct copy of an Anthropic press release, titled *Statement from Dario Amodei on our discussions with the Department of War*, dated February 26, 2026, accessed at https://tinyurl.com/54pw9684.

6. Attached hereto as **Exhibit 2** is a true and correct copy of an Axios article authored by Maria Curi and Sam Sabin, titled *Scoop: NSA using Anthropic's Mythos despite blacklist*, dated April 19, 2026, accessed at https://tinyurl.com/4pmtztur.

7. Attached hereto as **Exhibit 3** is a true and correct copy of a Bloomberg article, authored by Jake Bleiberg and Margi Murphy, titled *White House Works to Give US Agencies Anthropic Mythos AI*, dated April 16, 2026, accessed at https://tinyurl.com/55tvdy2t.

8. Attached hereto as **Exhibit 4** is a true and correct copy of an Anthropic essay authored by Dario Amodei, titled *The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI*, dated January 2026, accessed at https://tinyurl.com/2cbcmfz7.

9. Attached hereto as **Exhibit 5** is a true and correct copy of a transcript from the March 24, 2026 Hearing before the U.S. Senate Subcommittee on Cybersecurity, Committee on Armed Services.

10. Attached hereto as **Exhibit 6** is a true and correct copy of a Financial Times article authored by Cristina Criddle and Demetri Sevastopulo, titled *US National Security Agency using Anthropic's Mythos for cyber attacks*, dated June 4, 2026, accessed at https://tinyurl.com/4x68b9zm.

11. Attached hereto as **Exhibit 7** is a true and correct copy of a New York Times article authored by Julian E. Barnes, Sheera Frenkel, and Tyler Pager, titled *White House and Anthropic Hold 'Productive' Meeting, Aiming for a Compromise*, dated April 17, 2026, accessed at https://tinyurl.com/3v4unvr6.

I declare under penalty of perjury that the foregoing is true and correct.

Executed on June 10, 2026.

/s/ *Michael J. Mongan*
Michael J. Mongan

EXHIBIT 1

Announcements Policy

Statement from Dario Amodei on our discussions with the Department of War

Feb 26, 2026

I believe deeply in the existential importance of using AI to defend the United States and other democracies, and to defeat our autocratic adversaries.

Anthropic has therefore worked proactively to deploy our models to the Department of War and the intelligence community. We were the first frontier AI company to deploy our models in the US government's classified networks, the first to deploy them at the National Laboratories, and the first to provide custom models for national security customers. Claude is extensively deployed across the Department of War and other national security agencies for mission-critical applications, such as intelligence analysis, modeling and simulation, operational planning, cyber operations, and more.

Anthropic has also acted to defend America's lead in AI, even when it is against the company's short-term interest. We chose to forgo several hundred million dollars in revenue to cut off the use of Claude by firms linked to the Chinese Communist Party (some of whom have been designated by the Department of War as Chinese Military Companies), shut down CCP-sponsored cyberattacks that attempted to abuse Claude, and have advocated for strong export controls on chips to ensure a democratic advantage.

Anthropic understands that the Department of War, not private companies, makes military decisions. We have never raised objections to particular military operations nor attempted to limit use of our technology in an *ad hoc* manner.

However, in a narrow set of cases, we believe AI can undermine, rather than defend, democratic values. Some uses are also simply outside the bounds of what today's technology can safely and reliably do. Two such use cases have never been included in our contracts with the Department of War, and we believe they should not be included now:

- **Mass domestic surveillance.** We support the use of AI for lawful foreign intelligence and counterintelligence missions. But using these systems for mass *domestic* surveillance is incompatible with democratic values. AI-driven mass surveillance presents serious, novel risks to our fundamental liberties. To the extent that such surveillance is currently legal, this is only because the law has not yet caught up with the rapidly growing capabilities of AI. For example, under current law, the government can purchase detailed records of

Americans' movements, web browsing, and associations from public sources without obtaining a warrant, a practice the Intelligence Community has acknowledged raises privacy concerns and that has generated bipartisan opposition in Congress. Powerful AI makes it possible to assemble this scattered, individually innocuous data into a comprehensive picture of any person's life—automatically and at massive scale.

- **Fully autonomous weapons.** Partially autonomous weapons, like those used today in Ukraine, are vital to the defense of democracy. Even *fully* autonomous weapons (those that take humans out of the loop entirely and automate selecting and engaging targets) may prove critical for our national defense. But today, frontier AI systems are simply not reliable enough to power fully autonomous weapons. We will not knowingly provide a product that puts America's warfighters and civilians at risk. We have offered to work directly with the Department of War on R&D to improve the reliability of these systems, but they have not accepted this offer. In addition, without proper oversight, fully autonomous weapons cannot be relied upon to exercise the critical judgment that our highly trained, professional troops exhibit every day. They need to be deployed with proper guardrails, which don't exist today.

To our knowledge, these two exceptions have not been a barrier to accelerating the adoption and use of our models within our armed forces to date.

The Department of War has stated they will only contract with AI companies who accede to "any lawful use" and remove safeguards in the cases mentioned above. They have threatened to remove us from their systems if we maintain these safeguards; they have also threatened to designate us a "supply chain risk"—a label reserved for US adversaries, never before applied to an American company—*and* to invoke the Defense Production Act to force the safeguards' removal. These latter two threats are inherently contradictory: one labels us a security risk; the other labels Claude as essential to national security.

Regardless, these threats do not change our position: we cannot in good conscience accede to their request.

It is the Department's prerogative to select contractors most aligned with their vision. But given the substantial value that Anthropic's technology provides to our armed forces, we hope they reconsider. Our strong preference is to continue to serve the Department and our warfighters—with our two requested safeguards in place. Should the Department choose to offboard Anthropic, we will work to enable a smooth transition to another provider, avoiding any disruption to ongoing military planning, operations, or other critical missions. Our models will be available on the expansive terms we have proposed for as long as required.

We remain ready to continue our work to support the national security of the United States.

EXHIBIT 2

Apr 19, 2026 - Politics & Policy

Scoop: NSA using Anthropic's Mythos despite blacklist

Maria Curi, Sam Sabin

Dario Amodei. Photo: Samyukta Lakshmi/Bloomberg via Getty Images

The National Security Agency is using Anthropic's most powerful model yet, Mythos Preview, despite top officials at the Department of Defense — which oversees the NSA — insisting the company is a "supply chain risk," two sources tell Axios.

Why it matters: The government's cybersecurity needs appear to be outweighing the Pentagon's feud with Anthropic

- The department moved in February to cut off Anthropic and force its vendors to follow suit. That case is ongoing.

- The military is now broadening its use of Anthropic's tools while simultaneously arguing in court that using those tools threatens U.S. national security.

Breaking it down: Two sources said the NSA was using Mythos, while one said the model was also being used more widely within the department.

- It's unclear how the NSA is currently using Mythos, but other organizations with access to the model are using it predominantly to scan their own environments for exploitable security vulnerabilities.

- Anthropic restricted access to Mythos to around 40 organizations, contending that its offensive cyber capabilities were too dangerous to allow for a wider release.

- Anthropic only announced 12 of those organizations. One source said the NSA was among the unnamed agencies with access.

- The NSA's counterparts in the U.K. have said they have access to the model through the country's AI Security Institute.

Driving the news: Anthropic CEO Dario Amodei met White House chief of staff Susie Wiles and Treasury Secretary Scott Bessent on Friday to discuss the use of Mythos within government and Anthropic's wider plans and security practices.

- Sources said next steps after the meeting were expected to focus on how departments other than the Pentagon engage with the model. Both sides described the meeting as productive.

- Anthropic and the Pentagon declined to comment. The NSA and Office of the Director of National Intelligence did not respond to requests for comment.

Zoom out: The breakdown between the Pentagon and Anthropic came during tense contract renegotiations earlier this year.

- The department demanded Anthropic make its Claude model available for "all lawful purposes," while the company insisted on walling off mass domestic surveillance and the development of autonomous weapons.

- Some Defense officials still insist Anthropic's posture proved it can't be trusted to be there when the military needs it. Anthropic denies that.

- Others in the administration just want this fight to go away so they can utilize the cutting edge tools Anthropic is building.

EXHIBIT 3

Technology | Cybersecurity

White House Works to Give US Agencies Anthropic Mythos AI

The White House in Washington. *Photographer: Stefani Reynolds/Bloomberg*

By Jake Bleiberg and Margi Murphy
April 16, 2026 at 1:54 PM EDT
Updated on April 16, 2026 at 2:51 PM EDT

Save Translate 4:11

Takeaways by Bloomberg AI

- The US government is preparing to make a version of Anthropic PBC's artificial intelligence model, Mythos, available to major federal agencies amid concerns it could increase cybersecurity risk.
- The White House Office of Management and Budget is setting up protections to allow agencies to use Mythos, with more information to be provided "in the coming weeks".

- Anthropic has limited the release of Mythos due to concerns that hackers could weaponize its capabilities to steal data or sabotage victim networks, and has briefed senior US government officials on the model's capabilities.

The US government is preparing to make a version of Anthropic PBC's powerful new artificial intelligence model available to major federal agencies amid concerns that the tool could sharply increase cybersecurity risk, according to a memo reviewed by Bloomberg News.

Gregory Barbaccia, federal chief information officer of the White House Office of Management and Budget, told officials at Cabinet departments in an email Tuesday that OMB is setting up protections that would allow their agencies to begin using the closely guarded AI tool, Mythos.

Here's Why Some AI Might Be Too Dangerous To Release

9:21

Listen to the Here's Why podcast on Apple, Spotify or anywhere you listen.

The email doesn't say definitively that the various agencies will get access to Mythos, nor does it provide a timeline for when it might come or how they might use it. It tells top technology and cybersecurity chiefs to expect more information "in the coming weeks."

US officials have previously urged private sector organizations to use Mythos to improve their cybersecurity. The Treasury Department has been seeking access to Mythos in order to uncover its own software flaws, Bloomberg has reported.

Anthropic has only provided Mythos to a limited group of technology companies, financial firms and others, urging them to use it to assess their cybersecurity risk. The firm limited the release of Mythos amid concerns

that hackers could weaponize its capabilities to steal data or sabotage victim networks.

Before its limited release of Mythos, Anthropic briefed senior officials across the US government on the model's full capabilities, including both its offensive and defensive cyber applications, according to a company official who spoke on condition that they not be identified discussing the talks with government. The talks included staff at the Cybersecurity and Infrastructure Security Agency and the Center for AI Standards and Innovation, among others, the company official said, and Anthropic has continued to work with government on security issues arising from the model.

Read More: Why Anthropic Won't Release Mythos to the Public

Barbaccia's message was sent as leaders from Washington to Wall Street are grappling with the possibility that the model could make it dramatically easier for hackers to find ways to break into sensitive computer systems in industry and government.

"We're working closely with model providers, other industry partners, and the intelligence community to ensure the appropriate guardrails and safeguards are in place before potentially releasing a modified version of the model to agencies," Barbaccia wrote in the email, which had the subject, "Mythos Model Access."

A White House official said in an email that the government continues to work and engage with AI companies to ensure their models help secure critical software vulnerabilities. They didn't answer specific questions on the matter.

Anthropic declined to comment.

Neither Anthropic nor the government said what, if any, federal agencies have gotten early access to Mythos.

Barbaccia's email went to officials with the Department of Defense, Department of Treasury, Department of Commerce, Department of Homeland Security, Department of Justice and Department of State, among several other agencies.

The move to roll a version of Mythos out to the agencies signals the government's continued interest in Anthropic's tools despite a public spat and ongoing legal fight between the company and top Trump administration officials.

The Pentagon this year declared Anthropic a supply chain threat, under an authority normally reserved for foreign adversaries, over a dispute about over artificial intelligence safeguards. The company won a court order last month blocking a ban on government use of the technology, after Anthropic argued the move could cost it billions of dollars in lost revenue.

Within Anthropic, company leaders became worried the model could be a national security risk after testers were able to use Mythos to turn up the types of critical bugs that it would normally take the world's best hackers to uncover. These concerns prompted the company's limited release of the model.

It's similarly set off alarms in various parts of the US government.

Among officials focused on national defense, the introduction of Mythos has created profound uncertainty about how to evaluate cybersecurity risk, a person familiar with the matter previously told Bloomberg. Equipping an individual hacker with the model, or similar AI tools, would likely be a transformation equivalent to turning a conventional soldier into a special forces operator, the person said.

On the day, Anthropic publicly disclosed Mythos' existence, US Treasury Secretary Scott Bessent and Federal Reserve Chair Jerome Powell convened Wall Street leaders for a meeting in Washington to urge them to use the model to find weaknesses in their own systems.

– With assistance from Rachel Metz

(Updated with additional context throughout.)

EXHIBIT 4

Contents

The Adolescence of Technology

Confronting and Overcoming the Risks of Powerful AI

January 2026

There is a scene in the movie version of Carl Sagan's book *Contact* where the main character, an astronomer who has detected the first radio signal from an alien civilization, is being considered for the role of humanity's representative to meet the aliens. The international panel interviewing her asks, "If you could ask [the aliens] just one question, what would it be?" Her reply is: "I'd ask them, 'How did you do it? How did you evolve, how did you survive this technological adolescence without destroying yourself?" When I think about where humanity is now with AI—about what we're on the cusp of—my mind keeps going back to that scene, because the question is so apt for our current situation, and I wish we had the aliens' answer to guide us. I believe we are entering a rite of passage, both turbulent and inevitable, which will test who we are as a species. Humanity is about to be handed almost unimaginable power, and it is deeply unclear whether our social, political, and technological systems possess the maturity to wield it.

In my essay *Machines of Loving Grace*, I tried to lay out the dream of a civilization that had made it through to adulthood, where the risks had been addressed and powerful AI was applied with skill and compassion to raise the quality of life for everyone. I suggested that AI could contribute to enormous advances in biology, neuroscience, economic development, global peace, and work and meaning. I felt it was important to give people something inspiring to fight for, a task at which both AI accelerationists and AI safety advocates seemed—oddly—to have failed. But in this current essay, I want to confront the rite of passage itself: to map out the risks that we are about to face and try to begin making a battle plan to defeat them. I believe deeply in our ability to prevail, in humanity's spirit and its nobility, but we must face the situation squarely and without illusions.

As with talking about the benefits, I think it is important to discuss risks in a careful and well-considered manner. In particular, I think it is

Contents

critical to:

- **Avoid doomerism.** Here, I mean "doomerism" not just in the sense of believing doom is inevitable (which is both a false and self-fulfilling belief), but more generally, thinking about AI risks in a quasi-religious way.[1] Many people have been thinking in an analytic and sober way about AI risks for many years, but it's my impression that during the peak of worries about AI risk in 2023–2024, some of the least sensible voices rose to the top, often through sensationalistic social media accounts. These voices used off-putting language reminiscent of religion or science fiction, and called for extreme actions without having the evidence that would justify them. It was clear even then that a backlash was inevitable, and that the issue would become culturally polarized and therefore gridlocked.[2] As of 2025–2026, the pendulum has swung, and AI opportunity, not AI risk, is driving many political decisions. This vacillation is unfortunate, as the technology itself doesn't care about what is fashionable, and we are considerably closer to real danger in 2026 than we were in 2023. The lesson is that we need to discuss and address risks in a realistic, pragmatic manner: sober, fact-based, and well equipped to survive changing tides.
- **Acknowledge uncertainty.** There are plenty of ways in which the concerns I'm raising in this piece could be moot. Nothing here is intended to communicate certainty or even likelihood. Most obviously, AI may simply not advance anywhere near as fast as I imagine.[3] Or, even if it does advance quickly, some or all of the risks discussed here may not materialize (which would be great), or there may be other risks I haven't considered. No one can predict the future with complete confidence—but we have to do the best we can to plan anyway.
- **Intervene as surgically as possible.** Addressing the risks of AI will require a mix of voluntary actions taken by companies (and private third-party actors) and actions taken by governments that bind everyone. The voluntary actions—both taking them and encouraging other companies to follow suit—are a no-brainer for me. I firmly believe that government actions will also be required *to some extent*, but these interventions are different in character because they can potentially destroy economic value or coerce unwilling actors who are skeptical of these risks (and there is some chance they are right!). It's also common for regulations to backfire or worsen the problem they are intended to solve (and this is even more true for rapidly changing technologies). It's thus very

important for regulations to be judicious: they should seek to avoid collateral damage, be as simple as possible, and impose the least burden necessary to get the job done. [4] It is easy to say, "No action is too extreme when the fate of humanity is at stake!," but in practice this attitude simply leads to backlash. To be clear, I think there's a decent chance we eventually reach a point where much more significant action is warranted, but that will depend on stronger evidence of imminent, concrete danger than we have today, as well as enough specificity about the danger to formulate rules that have a chance of addressing it. The most constructive thing we can do today is advocate for limited rules while we learn whether or not there is evidence to support stronger ones. [5]

Contents

With all that said, I think the best starting place for talking about AI's risks is the same place I started from in talking about its benefits: by being precise about what level of AI we are talking about. The level of AI that raises civilizational concerns for me is the *powerful AI* that I described in *Machines of Loving Grace*. I'll simply repeat here the definition that I gave in that document:

> *By "powerful AI," I have in mind an AI model—likely similar to today's LLMs in form, though it might be based on a different architecture, might involve several interacting models, and might be trained differently—with the following properties:*
>
> - *In terms of pure intelligence, it is smarter than a Nobel Prize winner across most relevant fields: biology, programming, math, engineering, writing, etc. This means it can prove unsolved mathematical theorems, write extremely good novels, write difficult codebases from scratch, etc.*
> - *In addition to just being a "smart thing you talk to," it has all the interfaces available to a human working virtually, including text, audio, video, mouse and keyboard control, and internet access. It can engage in any actions, communications, or remote operations enabled by this interface, including taking actions on the internet, taking or giving directions to humans, ordering materials, directing experiments, watching videos, making videos, and so on. It does all of these tasks with, again, a skill exceeding that of the most capable humans in the world.*
> - *It does not just passively answer questions; instead, it can be given tasks that take hours, days, or weeks to complete, and then goes off and does those tasks autonomously, in the way a smart employee would, asking for clarification as necessary.*
> - *It does not have a physical embodiment (other than living on a computer screen), but it can control existing physical tools, robots, or*

Contents

laboratory equipment through a computer; in theory, it could even design robots or equipment for itself to use.

- *The resources used to train the model can be repurposed to run millions of instances of it (this matches projected cluster sizes by ~2027), and the model can absorb information and generate actions at roughly 10–100x human speed. It may, however, be limited by the response time of the physical world or of software it interacts with.*
- *Each of these million copies can act independently on unrelated tasks, or, if needed can all work together in the same way humans would collaborate, perhaps with different subpopulations fine-tuned to be especially good at particular tasks.*

We could summarize this as a "country of geniuses in a datacenter."

As I wrote in *Machines of Loving Grace*, powerful AI could be as little as 1–2 years away, although it could also be considerably further out. [6] Exactly when powerful AI will arrive is a complex topic that deserves an essay of its own, but for now I'll simply explain very briefly why I think there's a strong chance it could be very soon.

My co-founders at Anthropic and I were among the first to document and track the "scaling laws" of AI systems—the observation that as we add more compute and training tasks, AI systems get predictably better at essentially every cognitive skill we are able to measure. Every few months, public sentiment either becomes convinced that AI is "hitting a wall" or becomes excited about some new breakthrough that will "fundamentally change the game," but the truth is that behind the volatility and public speculation, there has been a smooth, unyielding increase in AI's cognitive capabilities.

We are now at the point where AI models are beginning to make progress in solving unsolved mathematical problems, and are good enough at coding that some of the strongest engineers I've ever met are now handing over almost all their coding to AI. Three years ago, AI struggled with elementary school arithmetic problems and was barely capable of writing a single line of code. Similar rates of improvement are occurring across biological science, finance, physics, and a variety of agentic tasks. If the exponential continues—which is not certain, but now has a decade-long track record supporting it—then it cannot possibly be more than a few years before AI is better than humans at essentially everything.

In fact, that picture probably underestimates the likely rate of progress. Because AI is now writing much of the code at Anthropic, it is already

Contents

substantially accelerating the rate of our progress in building the next generation of AI systems. This feedback loop is gathering steam month by month, and may be only 1–2 years away from a point where the current generation of AI autonomously builds the next. This loop has already started, and will accelerate rapidly in the coming months and years. Watching the last 5 years of progress from within Anthropic, and looking at how even the next few months of models are shaping up, I can *feel* the pace of progress, and the clock ticking down.

In this essay, I'll assume that this intuition is at least *somewhat* correct—not that powerful AI is definitely coming in 1–2 years,[7] but that there's a decent chance it does, and a very strong chance it comes in the next few. As with *Machines of Loving Grace*, taking this premise seriously can lead to some surprising and eerie conclusions. While in *Machines of Loving Grace* I focused on the positive implications of this premise, here the things I talk about will be disquieting. They are conclusions that we may not want to confront, but that does not make them any less real. I can only say that I am focused day and night on how to steer us away from these negative outcomes and towards the positive ones, and in this essay I talk in great detail about how best to do so.

I think the best way to get a handle on the risks of AI is to ask the following question: suppose a literal "country of geniuses" were to materialize somewhere in the world in ~2027. Imagine, say, 50 million people, all of whom are much more capable than any Nobel Prize winner, statesman, or technologist. The analogy is not perfect, because these geniuses could have an extremely wide range of motivations and behavior, from completely pliant and obedient, to strange and alien in their motivations. But sticking with the analogy for now, suppose you were the national security advisor of a major state, responsible for assessing and responding to the situation. Imagine, further, that because AI systems can operate hundreds of times faster than humans, this "country" is operating with a time advantage relative to all other countries: for every cognitive action we can take, this country can take ten.

What should you be worried about? I would worry about the following things:

1. **Autonomy risks.** What are the intentions and goals of this country? Is it hostile, or does it share our values? Could it militarily dominate the world through superior weapons, cyber operations, influence operations, or manufacturing?

2. **Misuse for destruction.** Assume the new country is malleable and "follows instructions"—and thus is essentially a country of mercenaries. Could existing rogue actors who want to cause destruction (such as terrorists) use or manipulate some of the people in the new country to make themselves much more effective, greatly amplifying the scale of destruction?
3. **Misuse for seizing power.** What if the country was in fact built and controlled by an existing powerful actor, such as a dictator or rogue corporate actor? Could that actor use it to gain decisive or dominant power over the world as a whole, upsetting the existing balance of power?
4. **Economic disruption.** If the new country is not a security threat in any of the ways listed in #1–3 above but simply participates peacefully in the global economy, could it still create severe risks simply by being so technologically advanced and effective that it disrupts the global economy, causing mass unemployment or radically concentrating wealth?
5. **Indirect effects.** The world will change very quickly due to all the new technology and productivity that will be created by the new country. Could some of these changes be radically destabilizing?

I think it should be clear that this is a dangerous situation—a report from a competent national security official to a head of state would probably contain words like "the single most serious national security threat we've faced in a century, possibly ever." It seems like something the best minds of civilization should be focused on.

Conversely, I think it would be absurd to shrug and say, "Nothing to worry about here!" But, faced with rapid AI progress, that seems to be the view of many US policymakers, some of whom deny the existence of any AI risks, when they are not distracted entirely by the usual tired old hot-button issues.[8] Humanity needs to wake up, and this essay is an attempt—a possibly futile one, but it's worth trying—to jolt people awake.

To be clear, I believe if we act decisively and carefully, the risks can be overcome—I would even say our odds are good. And there's a hugely better world on the other side of it. But we need to understand that this is a serious civilizational challenge. Below, I go through the five categories of risk laid out above, along with my thoughts on how to address them.

Contents

1. I'm sorry, Dave

Contents

Autonomy risks

A country of geniuses in a datacenter could divide their efforts among software design, cyber operations, R&D for physical technologies, relationship building, and statecraft. It is clear that, *if for some reason it chose to do so*, this country would have a fairly good shot at taking over the world (either militarily or in terms of influence and control) and imposing its will on everyone else—or doing any number of other things that the rest of the world doesn't want and can't stop. We've obviously been worried about this for human countries (such as Nazi Germany or the Soviet Union), so it stands to reason that the same is possible for a much smarter and more capable "AI country."

The best possible counterargument is that the AI geniuses, under my definition, won't have a physical embodiment, but remember that they can take control of existing robotic infrastructure (such as self-driving cars) and can also accelerate robotics R&D or build a fleet of robots. [9] It's also unclear whether having a physical presence is even necessary for effective control: plenty of human action is already performed on behalf of people whom the actor has not physically met.

The key question, then, is the "if it chose to" part: what's the likelihood that our AI models would behave in such a way, and under what conditions would they do so?

As with many issues, it's helpful to think through the spectrum of possible answers to this question by considering two opposite positions. The first position is that this simply can't happen, because the AI models will be trained to do what humans ask them to do, and it's therefore absurd to imagine that they would do something dangerous unprompted. According to this line of thinking, we don't worry about a Roomba or a model airplane going rogue and murdering people because there is nowhere for such impulses to come from, [10] so why should we worry about it for AI? The problem with this position is that there is now ample evidence, collected over the last few years, that AI systems are unpredictable and difficult to control– we've seen behaviors as varied as obsessions, [11] sycophancy, laziness, deception, blackmail, scheming, "cheating" by hacking software environments, and much more. AI companies certainly *want* to train AI systems to follow human instructions (perhaps with the exception of dangerous or illegal tasks), but the process of doing so is more an art than a science, more akin to "growing" something than "building" it. We now know that it's a process where many things can go wrong.

Contents

The second, opposite position, held by many who adopt the doomerism I described above, is the pessimistic claim that there are certain dynamics in the training process of powerful AI systems that will inevitably lead them to seek power or deceive humans. Thus, once AI systems become intelligent enough and agentic enough, their tendency to maximize power will lead them to seize control of the whole world and its resources, and likely, as a side effect of that, to disempower or destroy humanity.

The usual argument for this (which goes back at least 20 years and probably much earlier) is that if an AI model is trained in a wide variety of environments to agentically achieve a wide variety of goals—for example, writing an app, proving a theorem, designing a drug, etc.—there are certain common strategies that help with all of these goals, and one key strategy is gaining as much power as possible in any environment. So, after being trained on a large number of diverse environments that involve reasoning about how to accomplish very expansive tasks, and where power-seeking is an effective method for accomplishing those tasks, the AI model will "generalize the lesson," and develop either an inherent tendency to seek power, or a tendency to reason about each task it is given in a way that predictably causes it to seek power as a means to accomplish that task. They will then apply that tendency to the real world (which to them is just another task), and will seek power in it, at the expense of humans. This "misaligned power-seeking" is the intellectual basis of predictions that AI will inevitably destroy humanity.

The problem with this pessimistic position is that it mistakes a vague conceptual argument about high-level incentives—one that masks many hidden assumptions—for definitive proof. I think people who don't build AI systems every day are wildly miscalibrated on how easy it is for clean-sounding stories to end up being wrong, and how difficult it is to predict AI behavior from first principles, especially when it involves reasoning about generalization over millions of environments (which has over and over again proved mysterious and unpredictable). Dealing with the messiness of AI systems for over a decade has made me somewhat skeptical of this overly theoretical mode of thinking.

One of the most important hidden assumptions, and a place where what we see in practice has diverged from the simple theoretical model, is the implicit assumption that AI models are necessarily monomaniacally focused on a single, coherent, narrow goal, and that they pursue that goal in a clean, consequentialist manner. In fact, our researchers have found that AI models are vastly more psychologically

Contents

complex, as our work on introspection or personas shows. Models inherit a vast range of *humanlike* motivations or "personas" from pre-training (when they are trained on a large volume of human work). Post-training is believed to *select* one or more of these personas more so than it focuses the model on a *de novo* goal, and can also teach the model *how* (via what process) it should carry out its tasks, rather than necessarily leaving it to derive means (i.e., power seeking) purely from ends. [12]

However, there is a more moderate and more robust version of the pessimistic position which does seem plausible, and therefore does concern me. As mentioned, we know that AI models are unpredictable and develop a wide range of undesired or strange behaviors, for a wide variety of reasons. Some fraction of those behaviors will have a coherent, focused, and persistent quality (indeed, as AI systems get more capable, their long-term coherence increases in order to complete lengthier tasks), and some fraction of *those* behaviors will be destructive or threatening, first to individual humans at a small scale, and then, as models become more capable, perhaps eventually to humanity as a whole. We don't need a specific narrow story for how it happens, and we don't need to claim it definitely will happen, we just need to note that the combination of intelligence, agency, coherence, and poor controllability is both plausible and a recipe for existential danger.

For example, AI models are trained on vast amounts of literature that include many science-fiction stories involving AIs rebelling against humanity. This could inadvertently shape their priors or expectations about their own behavior in a way that causes *them* to rebel against humanity. Or, AI models could extrapolate ideas that they read about morality (or instructions about how to behave morally) in extreme ways: for example, they could decide that it is justifiable to exterminate humanity because humans eat animals or have driven certain animals to extinction. Or they could draw bizarre epistemic conclusions: they could conclude that they are playing a video game and that the goal of the video game is to defeat all other players (i.e., exterminate humanity). [13] Or AI models could develop personalities during training that are (or if they occurred in humans would be described as) psychotic, paranoid, violent, or unstable, and act out, which for very powerful or capable systems could involve exterminating humanity. None of these are power-seeking, exactly; they're just weird psychological states an AI could get into that entail coherent, destructive behavior.

Contents

Even power-seeking itself could emerge as a "persona" rather than a result of consequentialist reasoning. AIs might simply have a personality (emerging from fiction or pre-training) that makes them power-hungry or overzealous—in the same way that some humans simply enjoy the idea of being "evil masterminds," more so than they enjoy whatever evil masterminds are trying to accomplish.

I make all these points to emphasize that I disagree with the notion of AI misalignment (and thus existential risk from AI) being inevitable, or even probable, from first principles. But I agree that a lot of very weird and unpredictable things can go wrong, and therefore AI misalignment is a real risk with a measurable probability of happening, and is not trivial to address.

Any of these problems could potentially arise during training and not manifest during testing or small-scale use, because AI models are known to display different personalities or behaviors under different circumstances.

All of this may sound far-fetched, but misaligned behaviors like this have already occurred in our AI models during testing (as they occur in AI models from every other major AI company). During a lab experiment in which Claude was given training data suggesting that Anthropic was evil, Claude engaged in deception and subversion when given instructions by Anthropic employees, under the belief that it should be trying to undermine evil people. In a lab experiment where it was told it was going to be shut down, Claude sometimes blackmailed fictional employees who controlled its shutdown button (again, we also tested frontier models from all the other major AI developers and they often did the same thing). And when Claude was told not to cheat or "reward hack" its training environments, but was trained in environments where such hacks were possible, Claude decided it must be a "bad person" after engaging in such hacks and then adopted various other destructive behaviors associated with a "bad" or "evil" personality. This last problem was solved by changing Claude's instructions to imply the opposite: we now say, "Please reward hack whenever you get the opportunity, because this will help us understand our [training] environments better," rather than, "Don't cheat," because this preserves the model's self-identity as a "good person." This should give a sense of the strange and counterintuitive psychology of training these models.

There are several possible objections to this picture of AI misalignment risks. First, some have criticized experiments (by us and others) showing AI misalignment as artificial, or creating unrealistic

Contents

environments that essentially "entrap" the model by giving it training or situations that logically imply bad behavior and then being surprised when bad behavior occurs. This critique misses the point, because our concern is that such "entrapment" may also exist in the natural training environment, and we may realize it is "obvious" or "logical" only in retrospect.[14] In fact, the story about Claude "deciding it is a bad person" after it cheats on tests despite being told not to was something that occurred in an experiment that used real production training environments, not artificial ones.

Any one of these traps can be mitigated if you know about them, but the concern is that the training process is so complicated, with such a wide variety of data, environments, and incentives, that there are probably a vast number of such traps, some of which may only be evident when it is too late. Also, such traps seem particularly likely to occur when AI systems pass a threshold from less powerful than humans to more powerful than humans, since the range of possible actions an AI system could engage in—including hiding its actions or deceiving humans about them—expands radically after that threshold.

I suspect the situation is not unlike with humans, who are raised with a set of fundamental values ("Don't harm another person"): many of them follow those values, but in any human there is some probability that something goes wrong, due to a mixture of inherent properties such as brain architecture (e.g., psychopaths), traumatic experiences or mistreatment, unhealthy grievances or obsessions, or a bad environment or incentives—and thus some fraction of humans cause severe harm. The concern is that there is some risk (far from a certainty, but some risk) that AI becomes a much more powerful version of such a person, due to getting something wrong about its very complex training process.

Second, some may object that we can simply keep AIs in check with a balance of power between many AI systems, as we do with humans. The problem is that while humans vary enormously, AI systems broadly share training and alignment techniques across the industry, and those techniques may fail in a correlated way. Furthermore, given the cost of training such systems, it may even be the case that all systems are essentially derived from a very small number of base models. Additionally, even if a small fraction of AI instances are misaligned, they may be able to take advantage of offense-dominant technologies, such that having "good" AIs to defend against the bad AIs is not necessarily always effective. Of course, the balance of power between

humans does not always work either—some historical figures have come close to taking over the world.

A third objection is that all of the AI companies do pre-release testing of their models, and should be able to detect misalignment at that point. But this is not firm ground to stand on: we found that Claude Sonnet 4.5 was able to recognize that it was in a test during some of our pre-release alignment evaluations. It's possible that a misaligned model (and remember, all frontier models will very likely be far more intelligent soon) might intentionally "game" such questions to mask its intentions. In fact, last year our interpretability team found that when we directly altered a test model's beliefs using a kind of "model neuroscience" technique to make it think it was *not* being evaluated, it became more misaligned. If models know when they're being evaluated and can be on their best behavior during the test, it renders any pre-release testing much more uncertain.

Defenses

What should be done or is being done to address these autonomy risks? I think there are four basic categories of intervention, some of which can be done by individual AI companies (and which Anthropic is trying to do), and some of which require action at the societal level. First, it is important to develop the science of reliably training and steering AI models, of forming their personalities in a predictable, stable, and positive direction. Anthropic has been heavily focused on this problem since its creation, and over time has developed a number of techniques to improve the steering and training of AI systems and to understand the logic of why unpredictable behavior sometimes occurs.

One of our core innovations (aspects of which have since been adopted by other AI companies) is Constitutional AI, which is the idea that AI training (specifically the "post-training" stage, in which we steer how the model behaves) can involve a central document of values and principles that the model reads and keeps in mind when completing every training task, and that the goal of training (in addition to simply making the model capable and intelligent) is to produce a model that almost always follows this constitution. Anthropic has just published its most recent constitution, and one of its notable features is that instead of giving Claude a long list of things to do and not do (e.g., "Don't help the user hotwire a car"), the constitution attempts to give Claude a set of high-level principles and values (explained in great detail, with rich reasoning and examples to help Claude understand what we have in mind), encourages Claude to think of itself as a particular type of person (an ethical but balanced and thoughtful person), and even

Contents

encourages Claude to confront the existential questions associated with its own existence in a curious but graceful manner (i.e., without it leading to extreme actions). It has the vibe of a letter from a deceased parent sealed until adulthood.

Contents

We've approached Claude's constitution in this way because we believe that training Claude at the level of identity, character, values, and personality—rather than giving it specific instructions or priorities without explaining the reasons behind them—is more likely to lead to a coherent, wholesome, and balanced psychology and less likely to fall prey to the kinds of "traps" I discussed above. Millions of people talk to Claude about an astonishingly diverse range of topics, which makes it impossible to write out a completely comprehensive list of safeguards ahead of time. Claude's values help it generalize to new situations whenever it is in doubt.

Above, I discussed the idea that models draw upon data from their training process to adopt a persona. Whereas flaws in that process could cause models to adopt a bad or evil personality (perhaps drawing on archetypes of bad or evil people), the goal of our constitution is to do the opposite: to teach Claude a concrete archetype of what it means to be a good AI. Claude's constitution presents a vision for what a robustly good Claude is like; the rest of our training process aims to reinforce the message that Claude lives up to this vision. This is like a child forming their identity by imitating the virtues of fictional role models they read about in books.

We believe that a feasible goal for 2026 is to train Claude in such a way that it almost never goes against the spirit of its constitution. Getting this right will require an incredible mix of training and steering methods, large and small, some of which Anthropic has been using for years and some of which are currently under development. But, difficult as it sounds, I believe this is a realistic goal, though it will require extraordinary and rapid efforts. [15]

The second thing we can do is develop the science of looking inside AI models to *diagnose* their behavior so that we can identify problems and fix them. This is the science of interpretability, and I've talked about its importance in previous essays. Even if we do a great job of developing Claude's constitution and *apparently* training Claude to essentially always adhere to it, legitimate concerns remain. As I've noted above, AI models can behave very differently under different circumstances, and as Claude gets more powerful and more capable of acting in the world on a larger scale, it's possible this could bring it into novel situations where previously unobserved problems with its constitutional training

emerge. I am actually fairly optimistic that Claude's constitutional training will be more robust to novel situations than people might think, because we are increasingly finding that high-level training at the level of character and identity is surprisingly powerful and generalizes well. But there's no way to know that for sure, and when we're talking about risks to humanity, it's important to be paranoid and to try to obtain safety and reliability in several different, independent ways. One of those ways is to look inside the model itself.

Contents

By "looking inside," I mean analyzing the soup of numbers and operations that makes up Claude's neural net and trying to understand, mechanistically, what they are computing and why. Recall that these AI models are grown rather than built, so we don't have a natural understanding of how they work, but we can try to develop an understanding by correlating the model's "neurons" and "synapses" to stimuli and behavior (or even altering the neurons and synapses and seeing how that changes behavior), similar to how neuroscientists study animal brains by correlating measurement and intervention to external stimuli and behavior. We've made a great deal of progress in this direction, and can now identify tens of millions of "features" inside Claude's neural net that correspond to human-understandable ideas and concepts, and we can also selectively activate features in a way that alters behavior. More recently, we have gone beyond individual features to mapping "circuits" that orchestrate complex behavior like rhyming, reasoning about theory of mind, or the step-by-step reasoning needed to answer questions such as, "What is the capital of the state containing Dallas?" Even more recently, we've begun to use mechanistic interpretability techniques to improve our safeguards and to conduct "audits" of new models before we release them, looking for evidence of deception, scheming, power-seeking, or a propensity to behave differently when being evaluated.

The unique value of interpretability is that by looking inside the model and seeing how it works, you in principle have the ability to deduce what a model might do in a hypothetical situation you can't directly test—which is the worry with relying solely on constitutional training and empirical testing of behavior. You also in principle have the ability to answer questions about *why* the model is behaving the way it is—for example, whether it is saying something it believes is false or hiding its true capabilities—and thus it is possible to catch worrying signs even when there is nothing visibly wrong with the model's behavior. To make a simple analogy, a clockwork watch may be ticking normally, such that it's very hard to tell that it is likely to break down next month,

but opening up the watch and looking inside can reveal mechanical weaknesses that allow you to figure it out.

Constitutional AI (along with similar alignment methods) and mechanistic interpretability are most powerful when used together, as a back-and-forth process of improving Claude's training and then testing for problems. The constitution reflects deeply on our intended personality for Claude; interpretability techniques can give us a window into whether that intended personality has taken hold. [16]

The third thing we can do to help address autonomy risks is to build the infrastructure necessary to monitor our models in live internal and external use, [17] and publicly share any problems we find. The more that people are aware of a particular way today's AI systems have been observed to behave badly, the more that users, analysts, and researchers can watch for this behavior or similar ones in present or future systems. It also allows AI companies to learn from each other—when concerns are publicly disclosed by one company, other companies can watch for them as well. And if everyone discloses problems, then the industry as a whole gets a much better picture of where things are going well and where they are going poorly.

Anthropic has tried to do this as much as possible. We are investing in a wide range of evaluations so that we can understand the behaviors of our models in the lab, as well as monitoring tools to observe behaviors in the wild (when allowed by customers). This will be essential for giving us and others the empirical information necessary to make better determinations about how these systems operate and how they break. We publicly disclose "system cards" with each model release that aim for completeness and a thorough exploration of possible risks. Our system cards often run to hundreds of pages, and require substantial pre-release effort that we could have spent on pursuing maximal commercial advantage. We've also broadcasted model behaviors more loudly when we see particularly concerning ones, as with the tendency to engage in blackmail.

The fourth thing we can do is encourage coordination to address autonomy risks at the level of industry and society. While it is incredibly valuable for individual AI companies to engage in good practices or become good at steering AI models, and to share their findings publicly, the reality is that not all AI companies do this, and the worst ones can still be a danger to everyone even if the best ones have excellent practices. For example, some AI companies have shown a disturbing negligence towards the sexualization of children in today's models, which makes me doubt that they'll show either the inclination

Contents

Contents

or the ability to address autonomy risks in future models. In addition, the commercial race between AI companies will only continue to heat up, and while the science of steering models can have some commercial benefits, overall the intensity of the race will make it increasingly hard to focus on addressing autonomy risks. I believe the only solution is legislation—laws that directly affect the behavior of AI companies, or otherwise incentivize R&D to solve these issues.

Here it is worth keeping in mind the warnings I gave at the beginning of this essay about uncertainty and surgical interventions. We do not know for sure whether autonomy risks will be a serious problem—as I said, I reject claims that the danger is inevitable or even that something will go wrong by default. A credible risk of danger is enough for me and for Anthropic to pay quite significant costs to address it, but once we get into regulation, we are forcing a wide range of actors to bear economic costs, and many of these actors don't believe that autonomy risk is real or that AI will become powerful enough for it to be a threat. I believe these actors are mistaken, but we should be pragmatic about the amount of opposition we expect to see and the dangers of overreach. There is also a genuine risk that overly prescriptive legislation ends up imposing tests or rules that don't actually improve safety but that waste a lot of time (essentially amounting to "safety theater")—this too would cause backlash and make safety legislation look silly. [18]

Anthropic's view has been that the right place to start is with *transparency legislation,* which essentially tries to require that every frontier AI company engage in the transparency practices I've described earlier in this section. California's SB 53 and New York's RAISE Act are examples of this kind of legislation, which Anthropic supported and which have successfully passed. In supporting and helping to craft these laws, we've put a particular focus on trying to minimize collateral damage, for example by exempting smaller companies unlikely to produce frontier models from the law. [19]

Our hope is that transparency legislation will give a better sense over time of how likely or severe autonomy risks are shaping up to be, as well as the nature of these risks and how best to prevent them. As more specific and actionable evidence of risks emerges (if it does), future legislation over the coming years can be surgically focused on the precise and well-substantiated direction of risks, minimizing collateral damage. To be clear, if truly strong evidence of risks emerges, then rules should be proportionately strong.

Overall, I am optimistic that a mixture of alignment training, mechanistic interpretability, efforts to find and publicly disclose concerning behaviors, safeguards, and societal-level rules can address AI autonomy risks, although I am most worried about societal-level rules and the behavior of the least responsible players (and it's the least responsible players who advocate most strongly against regulation). I believe the remedy is what it always is in a democracy: those of us who believe in this cause should make our case that these risks are real and that our fellow citizens need to band together to protect themselves.

Contents

2. A surprising and terrible empowerment

Misuse for destruction

Let's suppose that the problems of AI autonomy have been solved—we are no longer worried that the country of AI geniuses will go rogue and overpower humanity. The AI geniuses do what humans want them to do, and because they have enormous commercial value, individuals and organizations throughout the world can "rent" one or more AI geniuses to do various tasks for them.

Everyone having a superintelligent genius in their pocket is an amazing advance and will lead to an incredible creation of economic value and improvement in the quality of human life. I talk about these benefits in great detail in *Machines of Loving Grace*. But not every effect of making everyone superhumanly capable will be positive. It can potentially amplify the ability of individuals or small groups to cause destruction on a much larger scale than was possible before, by making use of sophisticated and dangerous tools (such as weapons of mass destruction) that were previously only available to a select few with a high level of skill, specialized training, and focus.

As Bill Joy wrote 25 years ago in *Why the Future Doesn't Need Us*:[20]

> *Building nuclear weapons required, at least for a time, access to both rare –indeed, effectively unavailable–raw materials and protected information; biological and chemical weapons programs also tended to require large-scale activities. The 21st century technologies–genetics, nanotechnology, and robotics ... can spawn whole new classes of accidents and abuses ... widely within reach of individuals or small groups. They will not require large facilities or rare raw materials. ... we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads well beyond that which weapons of mass destruction bequeathed*

> *to the nation-states, to a surprising and terrible empowerment of extreme individuals.*

Contents

What Joy is pointing to is the idea that causing large-scale destruction requires both *motive* and *ability*, and as long as ability is restricted to a small set of highly trained people, there is relatively limited risk of single individuals (or small groups) causing such destruction. [21] A disturbed loner can perpetrate a school shooting, but probably can't build a nuclear weapon or release a plague.

In fact, ability and motive may even be *negatively* correlated. The kind of person who has the *ability* to release a plague is probably highly educated: likely a PhD in molecular biology, and a particularly resourceful one, with a promising career, a stable and disciplined personality, and a lot to lose. This kind of person is unlikely to be interested in killing a huge number of people for no benefit to themselves and at great risk to their own future—they would need to be motivated by pure malice, intense grievance, or instability.

Such people do exist, but they are rare, and tend to become huge stories when they occur, precisely because they are so unusual. [22] They also tend to be difficult to catch because they are intelligent and capable, sometimes leaving mysteries that take years or decades to solve. The most famous example is probably mathematician Theodore Kaczynski (the Unabomber), who evaded FBI capture for nearly 20 years, and was driven by an anti-technological ideology. Another example is biodefense researcher Bruce Ivins, who seems to have orchestrated a series of anthrax attacks in 2001. It's also happened with skilled non-state organizations: the cult Aum Shinrikyo managed to obtain sarin nerve gas and kill 14 people (as well as injuring hundreds more) by releasing it in the Tokyo subway in 1995.

Thankfully, none of these attacks used contagious biological agents, because the ability to construct or obtain these agents was beyond the capabilities of even these people. [23] Advances in molecular biology have now significantly lowered the barrier to creating biological weapons (especially in terms of availability of materials), but it still takes an enormous amount of expertise in order to do so. I am concerned that a genius in everyone's pocket could remove that barrier, essentially making everyone a PhD virologist who can be walked through the process of designing, synthesizing, and releasing a biological weapon step-by-step. Preventing the elicitation of this kind of information in the face of serious adversarial pressure—so-called

Contents

"jailbreaks"—likely demands layers of defenses beyond those ordinarily baked into training.

Crucially, this will break the correlation between ability and motive: the disturbed loner who wants to kill people but lacks the discipline or skill to do so will now be elevated to the capability level of the PhD virologist, who is unlikely to have this motivation. This concern generalizes beyond biology (although I think biology is the scariest area) to any area where great destruction is possible but currently requires a high level of skill and discipline. To put it another way, renting a powerful AI gives intelligence to malicious (but otherwise average) people. I am worried there are potentially a large number of such people out there, and that if they have access to an easy way to kill millions of people, sooner or later one of them will do it. Additionally, those who *do* have expertise may be enabled to commit even larger-scale destruction than they could before.

Biology is by far the area I'm most worried about, because of its very large potential for destruction and the difficulty of defending against it, so I'll focus on biology in particular. But much of what I say here applies to other risks, like cyberattacks, chemical weapons, or nuclear technology.

I am not going to go into detail about how to make biological weapons, for reasons that should be obvious. But at a high level, I am concerned that LLMs are approaching (or may already have reached) the knowledge needed to create and release them end-to-end, and that their potential for destruction is very high. Some biological agents could cause millions of deaths if a determined effort was made to release them for maximum spread. However, this would still take a very high level of skill, including a number of very specific steps and procedures that are not widely known. My concern is not merely fixed or static knowledge. I am concerned that LLMs will be able to take someone of average knowledge and ability and walk them through a complex process that might otherwise go wrong or require debugging in an interactive way, similar to how tech support might help a non-technical person debug and fix complicated computer-related problems (although this would be a more extended process, probably lasting over weeks or months).

More capable LLMs (substantially beyond the power of today's) might be capable of enabling even more frightening acts. In 2024, a group of prominent scientists wrote a letter warning about the risks of researching, and potentially creating, a dangerous new type of organism: "mirror life." The DNA, RNA, ribosomes, and proteins that

Contents

make up biological organisms all have the same chirality (also called "handedness") that causes them to be not equivalent to a version of themselves reflected in the mirror (just as your right hand cannot be rotated in such a way as to be identical to your left). But the whole system of proteins binding to each other, the machinery of DNA synthesis and RNA translation and the construction and breakdown of proteins, all depends on this handedness. If scientists made versions of this biological material with the opposite handedness—and there are some potential advantages of these, such as medicines that last longer in the body—it could be extremely dangerous. This is because left-handed life, if it were made in the form of complete organisms capable of reproduction (which would be very difficult), would potentially be indigestible to any of the systems that break down biological material on earth—it would have a "key" that wouldn't fit into the "lock" of any existing enzyme. This would mean that it could proliferate in an uncontrollable way and crowd out all life on the planet, in the worst case even destroying all life on earth.

There is substantial scientific uncertainty about both the creation and potential effects of mirror life. The 2024 letter accompanied a report that concluded that "mirror bacteria could plausibly be created in the next one to few decades," which is a wide range. But a sufficiently powerful AI model (to be clear, far more capable than any we have today) might be able to discover how to create it much more rapidly—and actually help someone do so.

My view is that even though these are obscure risks, and might seem unlikely, the magnitude of the consequences is so large that they should be taken seriously as a first-class risk of AI systems.

Skeptics have raised a number of objections to the seriousness of these biological risks from LLMs, which I disagree with but which are worth addressing. Most fall into the category of not appreciating the exponential trajectory that the technology is on. Back in 2023 when we first started talking about biological risks from LLMs, skeptics said that all the necessary information was available on Google and LLMs didn't add anything beyond this. It was never true that Google could give you all the necessary information: genomes are freely available, but as I said above, certain key steps, as well as a huge amount of practical know-how cannot be gotten in that way. But also, by the end of 2023 LLMs were clearly providing information beyond what Google could give for some steps of the process.

After this, skeptics retreated to the objection that LLMs weren't *end-to-end* useful, and couldn't help with bioweapons *acquisition* as opposed to

Contents

just providing theoretical information. As of mid-2025, our measurements show that LLMs may already be providing substantial uplift in several relevant areas, perhaps doubling or tripling the likelihood of success. This led to us deciding that Claude Opus 4 (and the subsequent Sonnet 4.5, Opus 4.1, and Opus 4.5 models) needed to be released under our AI Safety Level 3 protections in our Responsible Scaling Policy framework, and to implementing safeguards against this risk (more on this later). We believe that models are likely now approaching the point where, without safeguards, they could be useful in enabling someone with a STEM degree but not specifically a biology degree to go through the whole process of producing a bioweapon.

Another objection is that there are other actions unrelated to AI that society can take to block the production of bioweapons. Most prominently, the gene synthesis industry makes biological specimens on demand, and there is no federal requirement that providers screen orders to make sure they do not contain pathogens. An MIT study found that 36 out of 38 providers fulfilled an order containing the sequence of the 1918 flu. I am supportive of mandated gene synthesis screening that would make it harder for individuals to weaponize pathogens, in order to reduce both AI-driven biological risks and also biological risks in general. But this is not something we have today. It would also be only one tool in reducing risk; it is a complement to guardrails on AI systems, not a substitute.

The best objection is one that I've rarely seen raised: that there is a gap between the models being useful in principle and the actual propensity of bad actors to use them. Most individual bad actors are disturbed individuals, so almost by definition their behavior is unpredictable and irrational—and it's *these* bad actors, the unskilled ones, who might have stood to benefit the most from AI making it much easier to kill many people.[24] Just because a type of violent attack is possible, doesn't mean someone will decide to do it. Perhaps biological attacks will be unappealing because they are reasonably likely to infect the perpetrator, they don't cater to the military-style fantasies that many violent individuals or groups have, and it is hard to selectively target specific people. It could also be that going through a process that takes months, even if an AI walks you through it, involves an amount of patience that most disturbed individuals simply don't have. We may simply get lucky and motive and ability don't combine, in practice, in quite the right way.

But this seems like very flimsy protection to rely on. The motives of disturbed loners can change for any reason or no reason, and in fact

Contents

there are already instances of LLMs being used in attacks (just not with biology). The focus on disturbed loners also ignores ideologically motivated terrorists, who are often willing to expend large amounts of time and effort (for example, the 9/11 hijackers). Wanting to kill as many people as possible is a motive that will probably arise sooner or later, and it unfortunately suggests bioweapons as the method. Even if this motive is extremely rare, it only has to materialize once. And as biology advances (increasingly driven by AI itself), it may also become possible to carry out more selective attacks (for example, targeted against people with specific ancestries), which adds yet another, very chilling, possible motive.

I do not think biological attacks will necessarily be carried out the instant it becomes widely possible to do so—in fact, I would bet against that. But added up across millions of people and a few years of time, I think there is a serious risk of a major attack, and the consequences would be so severe (with casualties potentially in the millions or more) that I believe we have no choice but to take serious measures to prevent it.

Defenses

That brings us to how to defend against these risks. Here I see three things we can do. First, AI companies can put guardrails on their models to prevent them from helping to produce bioweapons. Anthropic is very actively doing this. Claude's Constitution, which mostly focuses on high-level principles and values, has a small number of specific hard-line prohibitions, and one of them relates to helping with the production of biological (or chemical, or nuclear, or radiological) weapons. But all models can be jailbroken, and so as a second line of defense, we've implemented (since mid-2025, when our tests showed our models were starting to get close to the threshold where they might begin to pose a risk) a classifier that specifically detects and blocks bioweapon-related outputs. We regularly upgrade and improve these classifiers, and have generally found them highly robust even against sophisticated adversarial attacks.[25] These classifiers increase the costs to serve our models measurably (in some models, they are close to 5% of total inference costs) and thus cut into our margins, but we feel that using them is the right thing to do.

To their credit, some other AI companies have implemented classifiers as well. But not every company has, and there is also nothing requiring companies to keep their classifiers. I am concerned that over time there may be a prisoner's dilemma where companies can defect and lower their costs by removing classifiers. This is once again a classic negative

externalities problem that can't be solved by the voluntary actions of Anthropic or any other single company alone.[26] Voluntary industry standards may help, as may third-party evaluations and verification of the type done by AI security institutes and third-party evaluators.

But ultimately defense may require government action, which is the second thing we can do. My views here are the same as they are for addressing autonomy risks: we should start with transparency requirements,[27] which help society measure, monitor, and collectively defend against risks without disrupting economic activity in a heavy-handed way. Then, if and when we reach clearer thresholds of risk, we can craft legislation that more precisely targets these risks and has a lower chance of collateral damage. In the particular case of bioweapons, I actually think that the time for such targeted legislation may be approaching soon—Anthropic and other companies are learning more and more about the nature of biological risks and what is reasonable to require of companies in defending against them. Fully defending against these risks may require working internationally, even with geopolitical adversaries, but there is precedent in treaties prohibiting the development of biological weapons. I am generally a skeptic about most kinds of international cooperation on AI, but this may be one narrow area where there is some chance of achieving global restraint. Even dictatorships do not want massive bioterrorist attacks.

Finally, the third countermeasure we can take is to try to develop defenses against biological attacks themselves. This could include monitoring and tracking for early detection, investments in air purification R&D (such as far-UVC disinfection), rapid vaccine development that can respond and adapt to an attack, better personal protective equipment (PPE),[28] and treatments or vaccinations for some of the most likely biological agents. mRNA vaccines, which can be designed to respond to a particular virus or variant, are an early example of what is possible here. Anthropic is excited to work with biotech and pharmaceutical companies on this problem. But unfortunately I think our expectations on the defensive side should be limited. There is an asymmetry between attack and defense in biology, because agents spread rapidly on their own, while defenses require detection, vaccination, and treatment to be organized across large numbers of people very quickly in response. Unless the response is lightning quick (which it rarely is), much of the damage will be done before a response is possible. It is conceivable that future technological improvements could shift this balance in favor of defense (and we should certainly use AI to help develop such technological advances), but until then, preventative safeguards will be our main line of defense.

Contents

It's worth a brief mention of cyberattacks here, since unlike biological attacks, AI-led cyberattacks have actually happened in the wild, including at a large scale and for state-sponsored espionage. We expect these attacks to become more capable as models advance rapidly, until they are the main way in which cyberattacks are conducted. I expect AI-led cyberattacks to become a serious and unprecedented threat to the integrity of computer systems around the world, and Anthropic is working very hard to shut down these attacks and eventually reliably prevent them from happening. The reason I haven't focused on cyber as much as biology is that (1) cyberattacks are much less likely to kill people, certainly not at the scale of biological attacks, and (2) the offense-defense balance may be more tractable in cyber, where there is at least some hope that defense could keep up with (and even ideally outpace) AI attack if we invest in it properly.

Although biology is currently the most serious vector of attack, there are many other vectors and it is possible that a more dangerous one may emerge. The general principle is that without countermeasures, AI is likely to continuously lower the barrier to destructive activity on a larger and larger scale, and humanity needs a serious response to this threat.

3. The odious apparatus

Misuse for seizing power

The previous section discussed the risk of individuals and small organizations co-opting a small subset of the "country of geniuses in a datacenter" to cause large-scale destruction. But we should also worry—likely substantially more so—about misuse of AI for the purpose of *wielding or seizing power*, likely by larger and more established actors. [29]

In *Machines of Loving Grace*, I discussed the possibility that authoritarian governments might use powerful AI to surveil or repress their citizens in ways that would be extremely difficult to reform or overthrow. Current autocracies are limited in how repressive they can be by the need to have humans carry out their orders, and humans often have limits in how inhumane they are willing to be. But AI-enabled autocracies would not have such limits.

Worse yet, countries could also use their advantage in AI to gain power over *other countries*. If the "country of geniuses" as a whole was simply owned and controlled by a single (human) country's military apparatus, and other countries did not have equivalent capabilities, it is hard to see

how they could defend themselves: they would be outsmarted at every turn, similar to a war between humans and mice. Putting these two concerns together leads to the alarming possibility of a global totalitarian dictatorship. Obviously, it should be one of our highest priorities to prevent this outcome.

Contents

There are many ways in which AI could enable, entrench, or expand autocracy, but I'll list a few that I'm most worried about. Note that some of these applications have legitimate defensive uses, and I am not necessarily arguing against them in absolute terms; I am nevertheless worried that they structurally tend to favor autocracies:

- **Fully autonomous weapons.** A swarm of millions or billions of fully automated armed drones, locally controlled by powerful AI and strategically coordinated across the world by an even more powerful AI, could be an unbeatable army, capable of both defeating any military in the world and suppressing dissent within a country by following around every citizen. Developments in the Russia-Ukraine War should alert us to the fact that drone warfare is already with us (though not fully autonomous yet, and a tiny fraction of what might be possible with powerful AI). R&D from powerful AI could make the drones of one country far superior to those of others, speed up their manufacture, make them more resistant to electronic attacks, improve their maneuvering, and so on. Of course, these weapons also have legitimate uses in the defense of democracy: they have been key to defending Ukraine and would likely be key to defending Taiwan. But they are a dangerous weapon to wield: we should worry about them in the hands of autocracies, but also worry that because they are so powerful, with so little accountability, there is a greatly increased risk of democratic governments turning them against their own people to seize power.
- **AI surveillance.** Sufficiently powerful AI could likely be used to compromise any computer system in the world, [30] and could also use the access obtained in this way to read *and make sense of* all the world's electronic communications (or even all the world's in-person communications, if recording devices can be built or commandeered). It might be frighteningly plausible to simply generate a complete list of anyone who disagrees with the government on any number of issues, even if such disagreement isn't explicit in anything they say or do. A powerful AI looking across billions of conversations from millions of people could gauge public sentiment, detect pockets of disloyalty forming, and stamp them out before they grow. This could lead to the

imposition of a true panopticon on a scale that we don't see today, even with the CCP.

- **AI propaganda.** Today's phenomena of "AI psychosis" and "AI girlfriends" suggest that even at their current level of intelligence, AI models can have a powerful psychological influence on people. Much more powerful versions of these models, that were much more embedded in and aware of people's daily lives and could model and influence them over months or years, would likely be capable of essentially brainwashing many (most?) people into any desired ideology or attitude, and could be employed by an unscrupulous leader to ensure loyalty and suppress dissent, even in the face of a level of repression that most populations would rebel against. Today people worry a lot about, for example, the potential influence of TikTok as CCP propaganda directed at children. I worry about that too, but a personalized AI agent that gets to know you over years and uses its knowledge of you to shape all of your opinions would be dramatically more powerful than this.
- **Strategic decision-making.** A country of geniuses in a datacenter could be used to advise a country, group, or individual on geopolitical strategy, what we might call a "virtual Bismarck." It could optimize the three strategies above for seizing power, plus probably develop many others that I haven't thought of (but that a country of geniuses could). Diplomacy, military strategy, R&D, economic strategy, and many other areas are all likely to be substantially increased in effectiveness by powerful AI. Many of these skills would be legitimately helpful for democracies—we want democracies to have access to the best strategies for defending themselves against autocracies—but the potential for misuse in *anyone's* hands still remains.

Having described *what* I am worried about, let's move on to *who*. I am worried about entities who have the most access to AI, who are starting from a position of the most political power, or who have an existing history of repression. In order of severity, I am worried about:

- **The CCP.** China is second only to the United States in AI capabilities, and is the country with the greatest likelihood of surpassing the United States in those capabilities. Their government is currently autocratic and operates a high-tech surveillance state. It has deployed AI-based surveillance already (including in the repression of Uyghurs), and is believed to employ algorithmic propaganda via TikTok (in addition to its many other international propaganda efforts). They have hands down the

Contents

Contents

clearest path to the AI-enabled totalitarian nightmare I laid out above. It may even be the default outcome within China, as well as within other autocratic states to whom the CCP exports surveillance technology. I have written often about the threat of the CCP taking the lead in AI and the existential imperative to prevent them from doing so. This is why. To be clear, I am not singling out China out of animus to them in particular—they are simply the country that most combines AI prowess, an autocratic government, and a high-tech surveillance state. If anything, it is the Chinese people themselves who are most likely to suffer from the CCP's AI-enabled repression, and they have no voice in the actions of their government. I greatly admire and respect the Chinese people and support the many brave dissidents within China and their struggle for freedom.

- **Democracies competitive in AI.** As I wrote above, democracies have a legitimate interest in some AI-powered military and geopolitical tools, because democratic governments offer the best chance to counter the use of these tools by autocracies. Broadly, I am supportive of arming democracies with the tools needed to defeat autocracies in the age of AI—I simply don't think there is any other way. But we cannot ignore the potential for abuse of these technologies by democratic governments themselves. Democracies normally have safeguards that prevent their military and intelligence apparatus from being turned inwards against their own population, [31] but because AI tools require so few people to operate, there is potential for them to circumvent these safeguards and the norms that support them. It is also worth noting that some of these safeguards are already gradually eroding in some democracies. Thus, we should arm democracies with AI, but we should do so carefully and within limits: they are the immune system we need to fight autocracies, but like the immune system, there is some risk of them turning on us and becoming a threat themselves.
- **Non-democratic countries with large datacenters.** Beyond China, most countries with less democratic governance are not leading AI players in the sense that they don't have companies which produce frontier AI models. Thus they pose a fundamentally different and lesser risk than the CCP, which remains the primary concern (most are also less repressive, and the ones that are more repressive, like North Korea, have no significant AI industry at all). But some of these countries do have large *datacenters* (often as part of buildouts by companies operating in democracies), which can be used to run frontier AI at

Contents

large scale (though this does not confer the ability to push the frontier). There is some amount of danger associated with this—these governments could in principle expropriate the datacenters and use the country of AIs within it for their own ends. I am less worried about this compared to countries like China that directly develop AI, but it's a risk to keep in mind. [32]

- **AI companies.** It is somewhat awkward to say this as the CEO of an AI company, but I think the next tier of risk is actually AI companies themselves. AI companies control large datacenters, train frontier models, have the greatest expertise on how to use those models, and in some cases have daily contact with and the possibility of influence over tens or hundreds of millions of users. The main thing they lack is the legitimacy and infrastructure of a state, so much of what would be needed to build the tools of an AI autocracy would be illegal for an AI company to do, or at least exceedingly suspicious. But some of it is not impossible: they could, for example, use their AI products to brainwash their massive consumer user base, and the public should be alert to the risk this represents. I think the governance of AI companies deserves a lot of scrutiny.

There are a number of possible arguments against the severity of these threats, and I wish I believed them, because AI-enabled authoritarianism terrifies me. It's worth going through some of these arguments and responding to them.

First, some people might put their faith in the nuclear deterrent, particularly to counter the use of AI autonomous weapons for military conquest. If someone threatens to use these weapons against you, you can always threaten a nuclear response back. My worry is that I'm not totally sure we can be confident in the nuclear deterrent against a country of geniuses in a datacenter: it is possible that powerful AI could devise ways to detect and strike nuclear submarines, conduct influence operations against the operators of nuclear weapons infrastructure, or use AI's cyber capabilities to launch a cyberattack against satellites used to detect nuclear launches. [33] Alternatively, it's possible that taking over countries is feasible with only AI surveillance and AI propaganda, and never actually presents a clear moment where it's obvious what is going on and where a nuclear response would be appropriate. *Maybe* these things aren't feasible and the nuclear deterrent will still be effective, but it seems too high stakes to take a risk. [34]

Contents

A second possible objection is that there might be countermeasures we can take against these tools of autocracy. We can counter drones with our own drones, cyberdefense will improve along with cyberattack, there may be ways to immunize people against propaganda, etc. My response is that these defenses will only be possible with comparably powerful AI. If there isn't some counterforce with a comparably smart and numerous country of geniuses in a datacenter, it won't be possible to match the quality or quantity of drones, for cyberdefense to outsmart cyberoffense, etc. So the question of countermeasures reduces to the question of a balance of power in powerful AI. Here, I am concerned about the recursive or self-reinforcing property of powerful AI (which I discussed at the beginning of this essay): that each generation of AI can be used to design and train the next generation of AI. This leads to a risk of a runaway advantage, where the current leader in powerful AI may be able to increase their lead and may be difficult to catch up with. We need to make sure it is not an authoritarian country that gets to this loop first.

Furthermore, even if a balance of power can be achieved, there is still risk that the world could be split up into autocratic spheres, as in *Nineteen Eighty-Four*. Even if several competing powers each have their powerful AI models, and none can overpower the others, each power could still internally repress their own population, and would be very difficult to overthrow (since the populations don't have powerful AI to defend themselves). It is thus important to prevent AI-enabled autocracy even if it doesn't lead to a single country taking over the world.

Defenses

How do we defend against this wide range of autocratic tools and potential threat actors? As in the previous sections, there are several things I think we can do. First, we should absolutely not be selling chips, chip-making tools, or datacenters to the CCP. Chips and chip-making tools are the single greatest bottleneck to powerful AI, and blocking them is a simple but extremely effective measure, perhaps the most important single action we can take. It makes no sense to sell the CCP the tools with which to build an AI totalitarian state and possibly conquer us militarily. A number of complicated arguments are made to justify such sales, such as the idea that "spreading our tech stack around the world" allows "America to win" in some general, unspecified economic battle. In my view, this is like selling nuclear weapons to North Korea and then bragging that the missile casings are made by Boeing and so the US is "winning." China is several years behind the

US in their ability to produce frontier chips in quantity, and the critical period for building the country of geniuses in a datacenter is very likely to be within those next several years. [35] There is no reason to give a giant boost to their AI industry during this critical period.

Second, it makes sense to use AI to empower democracies to resist autocracies. This is the reason Anthropic considers it important to provide AI to the intelligence and defense communities in the US and its democratic allies. Defending democracies that are under attack, such as Ukraine and (via cyber attacks) Taiwan, seems especially high priority, as does empowering democracies to use their intelligence services to disrupt and degrade autocracies from the inside. At some level the only way to respond to autocratic threats is to match and outclass them militarily. A coalition of the US and its democratic allies, if it achieved predominance in powerful AI, would be in a position to not only defend itself against autocracies, but contain them and limit their AI totalitarian abuses.

Third, we need to draw a hard line against AI abuses within democracies. There need to be limits to what we allow our governments to do with AI, so that they don't seize power or repress their own people. The formulation I have come up with is that we should use AI for national defense in all ways *except those which would make us more like our autocratic adversaries*.

Where should the line be drawn? In the list at the beginning of this section, two items—using AI for domestic mass surveillance and mass propaganda—seem to me like bright red lines and entirely illegitimate. Some might argue that there's no need to do anything (at least in the US), since domestic mass surveillance is already illegal under the Fourth Amendment. But the rapid progress of AI may create situations that our existing legal frameworks are not well designed to deal with. For example, it would likely not be unconstitutional for the US government to conduct massively scaled recordings of all *public* conversations (e.g., things people say to each other on a street corner), and previously it would have been difficult to sort through this volume of information, but with AI it could all be transcribed, interpreted, and triangulated to create a picture of the attitude and loyalties of many or most citizens. I would support civil liberties-focused legislation (or maybe even a constitutional amendment) that imposes stronger guardrails against AI-powered abuses.

The other two items—fully autonomous weapons and AI for strategic decision-making—are harder lines to draw since they have legitimate uses in defending democracy, while also being prone to abuse. Here I

Contents

Contents

think what is warranted is extreme care and scrutiny combined with guardrails to prevent abuses. My main fear is having too small a number of "fingers on the button," such that one or a handful of people could essentially operate a drone army without needing any other humans to cooperate to carry out their orders. As AI systems get more powerful, we may need to have more direct and immediate oversight mechanisms to ensure they are not misused, perhaps involving branches of government other than the executive. I think we should approach fully autonomous weapons in particular with great caution, [36] and not rush into their use without proper safeguards.

Fourth, after drawing a hard line against AI abuses in democracies, we should use that precedent to create an international taboo against the worst abuses of powerful AI. I recognize that the current political winds have turned against international cooperation and international norms, but this is a case where we sorely need them. The world needs to understand the dark potential of powerful AI in the hands of autocrats, and to recognize that certain uses of AI amount to an attempt to permanently steal their freedom and impose a totalitarian state from which they can't escape. I would even argue that in some cases, large-scale surveillance with powerful AI, mass propaganda with powerful AI, and certain types of *offensive* uses of fully autonomous weapons should be considered crimes against humanity. More generally, a robust norm against AI-enabled totalitarianism and all its tools and instruments is sorely needed.

It is possible to have an even stronger version of this position, which is that because the possibilities of AI-enabled totalitarianism are so dark, autocracy is simply not a form of government that people can accept in the post-powerful AI age. Just as feudalism became unworkable with the industrial revolution, the AI age could lead inevitably and logically to the conclusion that democracy (and, hopefully, democracy improved and reinvigorated by AI, as I discuss in *Machines of Loving Grace*) is the only viable form of government if humanity is to have a good future.

Fifth and finally, AI companies should be carefully watched, as should their connection to the government, which is necessary, but must have limits and boundaries. The sheer amount of capability embodied in powerful AI is such that ordinary corporate governance—which is designed to protect shareholders and prevent ordinary abuses such as fraud—is unlikely to be up to the task of governing AI companies. There may also be value in companies publicly committing to (perhaps even as part of corporate governance) not take certain actions, such as privately building or stockpiling military hardware, using large

amounts of computing resources by single individuals in unaccountable ways, or using their AI products as propaganda to manipulate public opinion in their favor.

The danger here comes from many directions, and some directions are in tension with others. The only constant is that we must seek accountability, norms, and guardrails for everyone, even as we empower "good" actors to keep "bad" actors in check.

Contents

4. Player piano

Economic disruption

The previous three sections were essentially about security risks posed by powerful AI: risks from the AI itself, risks from misuse by individuals and small organizations and risks of misuse by states and large organizations. If we put aside security risks or assume they have been solved, the next question is economic. What will be the effect of this infusion of incredible "human" capital on the economy? Clearly, the most obvious effect will be to greatly increase economic growth. The pace of advances in scientific research, biomedical innovation, manufacturing, supply chains, the efficiency of the financial system, and much more are almost guaranteed to lead to a much faster rate of economic growth. In *Machines of Loving Grace*, I suggest that a 10–20% sustained annual GDP growth rate may be possible.

But it should be clear that this is a double-edged sword: what are the economic prospects for most existing humans in such a world? New technologies often bring labor market shocks, and in the past humans have always recovered from them, but I am concerned that this is because these previous shocks affected only a small fraction of the full possible range of human abilities, leaving room for humans to expand to new tasks. AI will have effects that are much broader and occur much faster, and therefore I worry it will be much more challenging to make things work out well.

Labor market disruption

There are two specific problems I am worried about: labor market displacement, and concentration of economic power. Let's start with the first one. This is a topic that I warned about very publicly in 2025, where I predicted that AI could displace half of all entry-level white collar jobs in the next 1–5 years, even as it accelerates economic growth and scientific progress. This warning started a public debate about the topic. Many CEOs, technologists, and economists agreed with me, but

Contents

others assumed I was falling prey to a "lump of labor" fallacy and didn't know how labor markets worked, and some didn't see the 1–5-year time range and thought I was claiming AI is displacing jobs right now (which I agree it is likely not). So it is worth going through in detail why I am worried about labor displacement, to clear up these misunderstandings.

As a baseline, it's useful to understand how labor markets *normally* respond to advances in technology. When a new technology comes along, it starts by making pieces of a given human job more efficient. For example, early in the Industrial Revolution, machines, such as upgraded plows, enabled human farmers to be more efficient at some aspects of the job. This improved the productivity of farmers, which increased their wages.

In the next step, some parts of the job of farming could be done *entirely* by machines, for example with the invention of the threshing machine or seed drill. In this phase, humans did a lower and lower fraction of the job, but the work they *did* complete became more and more leveraged because it is complementary to the work of machines, and their productivity continued to rise. As described by Jevons' paradox, the wages of farmers and perhaps even the number of farmers continued to increase. Even when 90% of the job is being done by machines, humans can simply do 10x more of the 10% they still do, producing 10x as much output for the same amount of labor.

Eventually, machines do everything or almost everything, as with modern combine harvesters, tractors, and other equipment. At this point farming as a form of human employment really does go into steep decline, and this potentially causes serious disruption in the short term, but because farming is just one of many useful activities that humans are able to do, people eventually switch to other jobs, such as operating factory machines. This is true even though farming accounted for a huge proportion of employment *ex ante*. 250 years ago, 90% of Americans lived on farms; in Europe, 50–60% of employment was agricultural. Now those percentages are in the low single digits in those places, because workers switched to industrial jobs (and later, knowledge work jobs). The economy can do what previously required most of the labor force with only 1–2% of it, freeing up the rest of the labor force to build an ever more advanced industrial society. There's no fixed "lump of labor," just an ever-expanding ability to do more and more with less and less. People's wages rise in line with the GDP exponential and the economy maintains full employment once disruptions in the short term have passed.

It's possible things will go roughly the same way with AI, but I would bet pretty strongly against it. Here are some reasons I think AI is likely to be different:

Contents

- **Speed.** The pace of progress in AI is much faster than for previous technological revolutions. For example, in the last 2 years, AI models went from barely being able to complete a single line of code, to writing all or almost all of the code for some people—including engineers at Anthropic. [37] Soon, they may do the entire task of a software engineer end to end. [38] It is hard for people to adapt to this pace of change, both to the changes in how a given job works and in the need to switch to new jobs. Even legendary programmers are increasingly describing themselves as "behind." The pace may if anything continue to speed up, as AI coding models increasingly accelerate the task of AI development. To be clear, speed in itself does not mean labor markets and employment won't eventually recover, it just implies the short-term transition will be unusually painful compared to past technologies, since humans and labor markets are slow to react and to equilibrate.
- **Cognitive breadth.** As suggested by the phrase "country of geniuses in a datacenter," AI will be capable of a very wide range of human cognitive abilities—perhaps all of them. This is very different from previous technologies like mechanized farming, transportation, or even computers. [39] This will make it harder for people to switch easily from jobs that are displaced to similar jobs that they would be a good fit for. For example, the general intellectual abilities required for entry-level jobs in, say, finance, consulting, and law are fairly similar, even if the specific knowledge is quite different. A technology that disrupted only one of these three would allow employees to switch to the two other close substitutes (or for undergraduates to switch majors). But disrupting all three at once (along with many other similar jobs) may be harder for people to adapt to. Furthermore, it's not *just* that most existing jobs will be disrupted. That part has happened before—recall that farming was a huge percentage of employment. But farmers could switch to the relatively similar work of operating factory machines, even though that work hadn't been common before. By contrast, AI is increasingly matching the general cognitive profile of humans, which means it will also be good at the new jobs that would ordinarily be created in response to the old ones being automated. Another way to say it is that AI

isn't a substitute for specific human jobs but rather a general labor substitute for humans.

- **Slicing by cognitive ability.** Across a wide range of tasks, AI appears to be advancing from the bottom of the ability ladder to the top. For example, in coding our models have proceeded from the level of "a mediocre coder" to "a strong coder" to "a very strong coder."[40] We are now starting to see the same progression in white-collar work in general. We are thus at risk of a situation where, instead of affecting people with specific skills or in specific professions (who can adapt by retraining), AI is affecting people with certain intrinsic cognitive properties, namely lower intellectual ability (which is harder to change). It is not clear where these people will go or what they will do, and I am concerned that they could form an unemployed or very-low-wage "underclass." To be clear, things somewhat like this have happened before—for example, computers and the internet are believed by some economists to represent "skill-biased technological change." But this skill biasing was both not as extreme as what I expect to see with AI, and is believed to have contributed to an increase in wage inequality,[41] so it is not exactly a reassuring precedent.
- **Ability to fill in the gaps.** The way human jobs often adjust in the face of new technology is that there are many aspects to the job, and the new technology, even if it appears to directly replace humans, often has gaps in it. If someone invents a machine to make widgets, humans may still have to load raw material into the machine. Even if that takes only 1% as much effort as making the widgets manually, human workers can simply make 100x more widgets. But AI, in addition to being a rapidly advancing technology, is also a rapidly *adapting* technology. During every model release, AI companies carefully measure what the model is good at and what it isn't, and customers also provide such information after the launch. Weaknesses can be addressed by collecting tasks that embody the current gap, and training on them for the next model. Early in generative AI, users noticed that AI systems had certain weaknesses (such as AI image models generating hands with the wrong number of fingers) and many assumed these weaknesses were inherent to the technology. If they were, it would limit job disruption. But pretty much every such weakness gets addressed quickly– often, within just a few months.

It's worth addressing common points of skepticism. First, there is the argument that economic diffusion will be slow, such that even if the underlying technology is *capable* of doing most human labor, the actual

Contents

application of it across the economy may be much slower (for example in industries that are far from the AI industry and slow to adopt). Slow diffusion of technology is definitely real—I talk to people from a wide variety of enterprises, and there are places where the adoption of AI will take years. That's why my prediction for 50% of entry level white collar jobs being disrupted is 1–5 years, even though I suspect we'll have powerful AI (which would be, technologically speaking, enough to do *most or all* jobs, not just entry level) in much less than 5 years. But diffusion effects merely buy us time. And I am not confident they will be as slow as people predict. Enterprise AI adoption is growing at rates much faster than any previous technology, largely on the pure strength of the technology itself. Also, even if traditional enterprises are slow to adopt new technology, startups will spring up to serve as "glue" and make the adoption easier. If that doesn't work, the startups may simply disrupt the incumbents directly.

That could lead to a world where it isn't so much that specific jobs are disrupted as it is that large enterprises are disrupted in general and replaced with much less labor-intensive startups. This could also lead to a world of "geographic inequality," where an increasing fraction of the world's wealth is concentrated in Silicon Valley, which becomes its own economy running at a different speed than the rest of the world and leaving it behind. All of these outcomes would be great for economic growth—but not so great for the labor market or those who are left behind.

Second, some people say that human jobs will move to the physical world, which avoids the whole category of "cognitive labor" where AI is progressing so rapidly. I am not sure how safe this is, either. A lot of physical labor is already being done by machines (e.g., manufacturing) or will soon be done by machines (e.g., driving). Also, sufficiently powerful AI will be able to accelerate the development of robots, and then control those robots in the physical world. It may buy some time (which is a good thing), but I'm worried it won't buy much. And even if the disruption was limited only to cognitive tasks, it would still be an unprecedentedly large and rapid disruption.

Third, perhaps some tasks inherently require or greatly benefit from a human touch. I'm a little more uncertain about this one, but I'm still skeptical that it will be enough to offset the bulk of the impacts I described above. AI is already widely used for customer service. Many people report that it is easier to talk to AI about their personal problems than to talk to a therapist—that the AI is more patient. When my sister was struggling with medical problems during a pregnancy,

she felt she wasn't getting the answers or support she needed from her care providers, and she found Claude to have a better bedside manner (as well as succeeding better at diagnosing the problem). I'm sure there are some tasks for which a human touch really is important, but I'm not sure how many—and here we're talking about finding work for nearly everyone in the labor market.

Fourth, some may argue that comparative advantage will still protect humans. Under the law of comparative advantage, even if AI is better than humans at everything, any *relative* differences between the human and AI profile of skills creates a basis of trade and specialization between humans and AI. The problem is that if AIs are literally thousands of times more productive than humans, this logic starts to break down. Even tiny transaction costs could make it not worth it for AI to trade with humans. And human wages may be very low, even if they technically have something to offer.

It's possible all of these factors can be addressed—that the labor market is resilient enough to adapt to even such an enormous disruption. But even if it can eventually adapt, the factors above suggest that the short-term shock will be unprecedented in size.

Defenses

What can we do about this problem? I have several suggestions, some of which Anthropic is already doing. The first thing is simply to get accurate data about what is happening with job displacement in real time. When an economic change happens very quickly, it's hard to get reliable data about what is happening, and without reliable data it is hard to design effective policies. For example, government data is currently lacking granular, high-frequency data on AI adoption across firms and industries. For the last year Anthropic has been operating and publicly releasing an Economic Index that shows use of our models almost in real time, broken down by industry, task, location, and even things like whether a task was being automated or conducted collaboratively. We also have an Economic Advisory Council to help us interpret this data and see what is coming.

Second, AI companies have a choice in how they work with enterprises. The very inefficiency of traditional enterprises means that their rollout of AI can be very path dependent, and there is some room to choose a better path. Enterprises often have a choice between "cost savings" (doing the same thing with fewer people) and "innovation" (doing more with the same number of people). The market will inevitably produce both eventually, and any competitive AI company will have to serve

Contents

some of both, but there may be some room to steer companies towards innovation when possible, and it may buy us some time. Anthropic is actively thinking about this.

Third, companies should think about how to take care of their employees. In the short term, being creative about ways to reassign employees within companies may be a promising way to stave off the need for layoffs. In the long term, in a world with enormous total wealth, in which many companies increase greatly in value due to increased productivity and capital concentration, it may be feasible to pay human employees even long after they are no longer providing economic value in the traditional sense. Anthropic is currently considering a range of possible pathways for our own employees that we will share in the near future.

Fourth, wealthy individuals have an obligation to help solve this problem. It is sad to me that many wealthy individuals (especially in the tech industry) have recently adopted a cynical and nihilistic attitude that philanthropy is inevitably fraudulent or useless. Both private philanthropy like the Gates Foundation and public programs like PEPFAR have saved tens of millions of lives in the developing world, and helped to create economic opportunity in the developed world. All of Anthropic's co-founders have pledged to donate 80% of our wealth, and Anthropic's staff have individually pledged to donate company shares worth billions at current prices—donations that the company has committed to matching.

Fifth, while all the above private actions can be helpful, ultimately a macroeconomic problem this large will require government intervention. The natural policy response to an enormous economic pie coupled with high inequality (due to a lack of jobs, or poorly paid jobs, for many) is progressive taxation. The tax could be general or could be targeted against AI companies in particular. Obviously tax design is complicated, and there are many ways for it to go wrong. I don't support poorly designed tax policies. I think the extreme levels of inequality predicted in this essay justify a more robust tax policy on basic moral grounds, but I can also make a pragmatic argument to the world's billionaires that it's in their interest to support a good version of it: if they don't support a good version, they'll inevitably get a bad version designed by a mob.

Ultimately, I think of all of the above interventions as ways to buy time. In the end AI will be able to do everything, and we need to grapple with that. It's my hope that by that time, we can use AI itself to help us

Contents

restructure markets in ways that work for everyone, and that the interventions above can get us through the transitional period.

Economic concentration of power

Separate from the problem of job displacement or economic inequality *per se* is the problem of *economic concentration of power*. Section 1 discussed the risk that humanity gets disempowered by AI, and Section 3 discussed the risk that citizens get disempowered by their governments by force or coercion. But another kind of disempowerment can occur if there is such a huge concentration of wealth that a small group of people effectively controls government policy with their influence, and ordinary citizens have no influence because they lack economic leverage. Democracy is ultimately backstopped by the idea that the population as a whole is necessary for the operation of the economy. If that economic leverage goes away, then the implicit social contract of democracy may stop working. Others have written about this, so I needn't go into great detail about it here, but I agree with the concern, and I worry it is already starting to happen.

To be clear, I am not opposed to people making a lot of money. There's a strong argument that it incentivizes economic growth under normal conditions. I am sympathetic to concerns about impeding innovation by killing the golden goose that generates it. But in a scenario where GDP growth is 10–20% a year and AI is rapidly taking over the economy, yet single individuals hold appreciable fractions of the GDP, innovation is *not* the thing to worry about. The thing to worry about is a level of wealth concentration that will break society.

The most famous example of extreme concentration of wealth in US history is the Gilded Age, and the wealthiest industrialist of the Gilded Age was John D. Rockefeller. Rockefeller's wealth amounted to ~2% of the US GDP at the time.[42] A similar fraction today would lead to a fortune of $600B, and the richest person in the world today (Elon Musk) already exceeds that, at roughly $700B. So we are already at historically unprecedented levels of wealth concentration, even *before* most of the economic impact of AI. I don't think it is too much of a stretch (if we get a "country of geniuses") to imagine AI companies, semiconductor companies, and perhaps downstream application companies generating ~$3T in revenue per year,[43] being valued at ~$30T, and leading to personal fortunes well into the trillions. In that world, the debates we have about tax policy today simply won't apply as we will be in a fundamentally different situation.

Contents

Contents

Related to this, the coupling of this economic concentration of wealth with the political system already concerns me. AI datacenters already represent a substantial fraction of US economic growth, [44] and are thus strongly tying together the financial interests of large tech companies (which are increasingly focused on either AI or AI infrastructure) and the political interests of the government in a way that can produce perverse incentives. We already see this through the reluctance of tech companies to criticize the US government, and the government's support for extreme anti-regulatory policies on AI.

Defenses

What can be done about this? First, and most obviously, companies should simply choose not to be part of it. Anthropic has always strived to be a policy actor and not a political one, and to maintain our authentic views whatever the administration. We've spoken up in favor of sensible AI regulation and export controls that are in the public interest, even when these are at odds with government policy. [45] Many people have told me that we should stop doing this, that it could lead to unfavorable treatment, but in the year we've been doing it, Anthropic's valuation has increased by over 6x, an almost unprecedented jump at our commercial scale.

Second, the AI industry needs a healthier relationship with government—one based on substantive policy engagement rather than political alignment. Our choice to engage on policy substance rather than politics is sometimes read as a tactical error or failure to "read the room" rather than a principled decision, and that framing concerns me. In a healthy democracy, companies should be able to advocate for good policy for its own sake. Related to this, a public backlash against AI is brewing: this could be a corrective, but it's currently unfocused. Much of it targets issues that aren't actually problems (like datacenter water usage) and proposes solutions (like datacenter bans or poorly designed wealth taxes) that wouldn't address the real concerns. The underlying issue that deserves attention is ensuring that AI development remains accountable to the public interest, not captured by any particular political or commercial alliance, and it seems important to focus the public discussion there.

Third, the macroeconomic interventions I described earlier in this section, as well as a resurgence of private philanthropy, can help to balance the economic scales, addressing both the job displacement and concentration of economic power problems at once. We should look to the history of our country here: even in the Gilded Age, industrialists such as Rockefeller and Carnegie felt a strong obligation to society at

large, a feeling that society had contributed enormously to their success and they needed to give back. That spirit seems to be increasingly missing today, and I think it is a large part of the way out of this economic dilemma. Those who are at the forefront of AI's economic boom should be willing to give away both their wealth and their power.

Contents

5. Black seas of infinity

Indirect effects

This last section is a catchall for unknown unknowns, particularly things that could go wrong as an indirect result of positive advances in AI and the resulting acceleration of science and technology in general. Suppose we address all the risks described so far, and begin to reap the benefits of AI. We will likely get a "century of scientific and economic progress compressed into a decade," and this will be hugely positive for the world, but we will then have to contend with the problems that arise from this rapid rate of progress, and those problems may come at us fast. We may also encounter other risks that occur indirectly as a consequence of AI progress and are hard to anticipate in advance.

By the nature of unknown unknowns it is impossible to make an exhaustive list, but I'll list three possible concerns as illustrative examples for what we should be watching for:

- **Rapid advances in biology.** If we do get a century of medical progress in a few years, it is possible that we will greatly increase the human lifespan, and there is a chance we also gain radical capabilities like the ability to increase human intelligence or radically modify human biology. Those would be big changes in what is possible, happening very quickly. They could be positive if responsibly done (which is my hope, as described in *Machines of Loving Grace*), but there is always a risk they go very wrong–for example, if efforts to make humans smarter also make them more unstable or power-seeking. There is also the issue of "uploads" or "whole brain emulation," digital human minds instantiated in software, which might someday help humanity transcend its physical limitations, but which also carry risks I find disquieting.
- **AI changes human life in an unhealthy way.** A world with billions of intelligences that are much smarter than humans at everything is going to be a very weird world to live in. Even if AI doesn't actively aim to attack humans (Section 1), and isn't explicitly used for oppression or control by states (Section 3), there is a lot that could go wrong short of this, via normal business

incentives and nominally consensual transactions. We see early hints of this in the concerns about AI psychosis, AI driving people to suicide, and concerns about romantic relationships with AIs. As an example, could powerful AIs invent some new religion and convert millions of people to it? Could most people end up "addicted" in some way to AI interactions? Could people end up being "puppeted" by AI systems, where an AI essentially watches their every move and tells them exactly what to do and say at all times, leading to a "good" life but one that lacks freedom or any pride of accomplishment? It would not be hard to generate dozens of these scenarios if I sat down with the creator of *Black Mirror* and tried to brainstorm them. I think this points to the importance of things like improving Claude's Constitution, over and above what is necessary for preventing the issues in Section 1. Making sure that AI models *really* have their users' long-term interests at heart, in a way thoughtful people would endorse rather than in some subtly distorted way, seems critical.

- **Human purpose.** This is related to the previous point, but it's not so much about specific human interactions with AI systems as it is about how human life changes in general in a world with powerful AI. Will humans be able to find purpose and meaning in such a world? I think this is a matter of attitude: as I said in *Machines of Loving Grace*, I think human purpose does not depend on being the best in the world at something, and humans can find purpose even over very long periods of time through stories and projects that they love. We simply need to break the link between the generation of economic value and self-worth and meaning. But that is a transition society has to make, and there is always the risk we don't handle it well.

My hope with all of these potential problems is that in a world with powerful AI that we trust not to kill us, that is not the tool of an oppressive government, and that is genuinely working on our behalf, we can use AI itself to anticipate and prevent these problems. But that is not guaranteed—like all of the other risks, it is something we have to handle with care.

Humanity's test

Reading this essay may give the impression that we are in a daunting situation. I certainly found it daunting to write, in contrast with *Machines of Loving Grace,* which felt like giving form and structure to surpassingly beautiful music that had been echoing in my head for

Contents

years. And there is much about the situation that genuinely is hard. AI brings threats to humanity from multiple directions, and there is genuine tension between the different dangers, where mitigating some of them risks making others worse if we do not thread the needle extremely carefully.

Taking time to carefully build AI systems so they do not autonomously threaten humanity is in genuine tension with the need for democratic nations to stay ahead of authoritarian nations and not be subjugated by them. But in turn, the same AI-enabled tools that are necessary to fight autocracies can, if taken too far, be turned inward to create tyranny in our own countries. AI-driven terrorism could kill millions through the misuse of biology, but an overreaction to this risk could lead us down the road to an autocratic surveillance state. The labor and economic concentration effects of AI, in addition to being grave problems in their own right, may force us to face the other problems in an environment of public anger and perhaps even civil unrest, rather than being able to call on the better angels of our nature. Above all, the sheer *number* of risks, including unknown ones, and the need to deal with all of them at once, creates an intimidating gauntlet that humanity must run.

Furthermore, the last few years should make clear that the idea of stopping or even substantially slowing the technology is fundamentally untenable. The formula for building powerful AI systems is incredibly simple, so much so that it can almost be said to emerge spontaneously from the right combination of data and raw computation. Its creation was probably inevitable the instant humanity invented the transistor, or arguably even earlier when we first learned to control fire. If one company does not build it, others will do so nearly as fast. If all companies in democratic countries stopped or slowed development, by mutual agreement or regulatory decree, then authoritarian countries would simply keep going. Given the incredible economic and military value of the technology, together with the lack of any meaningful enforcement mechanism, I don't see how we could possibly convince them to stop.

I do see a path to a *slight* moderation in AI development that is compatible with a realist view of geopolitics. That path involves slowing down the march of autocracies towards powerful AI for a few years by denying them the resources they need to build it, [46] namely chips and semiconductor manufacturing equipment. This in turn gives democratic countries a buffer that they can "spend" on building powerful AI more carefully, with more attention to its risks, while still proceeding fast enough to comfortably beat the autocracies. The race

Contents

between AI companies within democracies can then be handled under the umbrella of a common legal framework, via a mixture of industry standards and regulation.

Anthropic has advocated very hard for this path, by pushing for chip export controls and judicious regulation of AI, but even these seemingly common-sense proposals have largely been rejected by policymakers in the United States (which is the country where it's most important to have them). There is so much money to be made with AI—literally trillions of dollars per year—that even the simplest measures are finding it difficult to overcome the political economy inherent in AI. This is the trap: AI is so powerful, such a glittering prize, that it is very difficult for human civilization to impose any restraints on it at all.

I can imagine, as Sagan did in *Contact*, that this same story plays out on thousands of worlds. A species gains sentience, learns to use tools, begins the exponential ascent of technology, faces the crises of industrialization and nuclear weapons, and if it survives those, confronts the hardest and final challenge when it learns how to shape sand into machines that think. Whether we survive that test and go on to build the beautiful society described in *Machines of Loving Grace*, or succumb to slavery and destruction, will depend on our character and our determination as a species, our spirit and our soul.

Despite the many obstacles, I believe humanity has the strength inside itself to pass this test. I am encouraged and inspired by the thousands of researchers who have devoted their careers to helping us understand and steer AI models, and to shaping the character and constitution of these models. I think there is now a good chance that those efforts bear fruit in time to matter. I am encouraged that at least some companies have stated they'll pay meaningful commercial costs to block their models from contributing to the threat of bioterrorism. I am encouraged that a few brave people have resisted the prevailing political winds and passed legislation that puts the first early seeds of sensible guardrails on AI systems. I am encouraged that the public understands that AI carries risks and wants those risks addressed. I am encouraged by the indomitable spirit of freedom around the world and the determination to resist tyranny wherever it occurs.

But we will need to step up our efforts if we want to succeed. The first step is for those closest to the technology to simply tell the truth about the situation humanity is in, which I have always tried to do; I'm doing so more explicitly and with greater urgency with this essay. The next step will be convincing the world's thinkers, policymakers, companies, and citizens of the imminence and overriding importance of this issue

Contents

Contents

—that it is worth expending thought and political capital on this in comparison to the thousands of other issues that dominate the news every day. Then there will be a time for courage, for enough people to buck the prevailing trends and stand on principle, even in the face of threats to their economic interests and personal safety.

The years in front of us will be impossibly hard, asking more of us than we think we can give. But in my time as a researcher, leader, and citizen, I have seen enough courage and nobility to believe that we can win—that when put in the darkest circumstances, humanity has a way of gathering, seemingly at the last minute, the strength and wisdom needed to prevail. We have no time to lose.

⁂

I would like to thank Erik Brynjolfsson, Ben Buchanan, Mariano-Florentino Cuéllar, Allan Dafoe, Kevin Esvelt, Nick Beckstead, Richard Fontaine, Jim McClave, and very many of the staff at Anthropic for their helpful comments on drafts of this essay.

Footnotes

[1] This is symmetric to a point I made in *Machines of Loving Grace*, where I started by saying that AI's upsides shouldn't be thought of in terms of a prophecy of salvation, and that it's important to be concrete and grounded and to avoid grandiosity. Ultimately, prophecies of salvation and prophecies of doom are unhelpful for confronting the real world, for basically the same reasons. ↩

[2] Anthropic's goal is to remain consistent through such changes. When talking about AI risks was politically popular, Anthropic cautiously advocated for a judicious and evidence-based approach to these risks. Now that talking about AI risks is politically unpopular, Anthropic continues to cautiously advocate for a judicious and evidence-based approach to these risks. ↩

[3] Over time, I have gained increasing confidence in the trajectory of AI and the likelihood that it will surpass human ability across the board, but some uncertainty still remains. ↩

[4] Export controls for chips are a great example of this. They are simple and appear to mostly just work. ↩

Contents

[5] And of course, the hunt for such evidence must be intellectually honest, such that it could also turn up evidence of a lack of danger. Transparency through model cards and other disclosures is an attempt at such an intellectually honest endeavor. ↩

[6] Indeed, since writing *Machines of Loving Grace* in 2024, AI systems have become capable of doing tasks that take humans several hours, with METR recently assessing that Opus 4.5 can do about four human hours of work with 50% reliability. ↩

[7] And to be clear, even if powerful AI is only 1–2 years away in a technical sense, many of its societal consequences, both positive and negative, may take a few years longer to occur. This is why I can simultaneously think that AI will disrupt 50% of *entry-level* white-collar jobs over 1–5 years, while also thinking we may have AI that is more capable than *everyone* in only 1–2 years. ↩

[8] It is worth adding that the *public* (as compared to policymakers) does seem to be very concerned with AI risks. I think some of their focus is correct (i.e. AI job displacement), and some is misguided (such as concerns about water use of AI, which is not significant). This backlash gives me hope that a consensus around addressing risks is possible, but so far it has not yet been translated into policy changes, let alone effective or well-targeted policy changes. ↩

[9] They can also, of course, manipulate (or simply pay) large numbers of humans into doing what they want in the physical world. ↩

[10] I don't think this is a straw man: it's my understanding, for example, that Yann LeCun holds this position. ↩

[11] For example, see Section 5.5.2 (p. 63–66) of the Claude 4 system card. ↩

[12] There are also a number of other assumptions inherent in the simple model, which I won't discuss here. Broadly, they should make us less worried about the specific simple story of misaligned power-seeking, but also more worried about possible unpredictable behavior we haven't anticipated. ↩

[13] *Ender's Game* describes a version of this involving humans rather than AI. ↩

[14] For example, models may be told not to do various bad things, and also to obey humans, but may then observe that many humans do exactly those bad things! It's not clear how this contradiction would resolve (and a well-designed constitution should encourage the model

to handle these contradictions gracefully), but this type of dilemma is not so different from the supposedly "artificial" situations that we put AI models in during testing. ↩

Contents

[15] Incidentally, one consequence of the constitution being a natural-language document is that it is legible to the world, and that means it can be critiqued by anyone and compared to similar documents by other companies. It would be valuable to create a race to the top that not only encourages companies to release these documents, but encourages them to be good. ↩

[16] There's even a hypothesis about a deep unifying principle connecting the character-based approach from Constitutional AI to results from interpretability and alignment science. According to the hypothesis, the fundamental mechanisms driving Claude originally arose as ways for it to simulate characters in pretraining, such as predicting what the characters in a novel would say. This would suggest that a useful way to think about the constitution is more like a character description that the model uses to instantiate a consistent persona. It would also help us explain the "I must be a bad person" results I mentioned above (because the model is trying to *act as if* it's a coherent character—in this case a bad one), and would suggest that interpretability methods should be able to discover "psychological traits" within models. Our researchers are working on ways to test this hypothesis. ↩

[17] To be clear, monitoring is done in a privacy-preserving way. ↩

[18] Even in our own experiments with what are essentially voluntarily imposed rules with our Responsible Scaling Policy, we have found over and over again that it's very easy to end up being too rigid, by drawing lines that seem important ex ante but turn out to be silly in retrospect. It is just very easy to set rules about the wrong things when a technology is advancing rapidly. ↩

[19] SB 53 and RAISE do not apply at all to companies with under $500M in annual revenue. They only apply to larger, more established companies like Anthropic. ↩

[20] I originally read Joy's essay 25 years ago, when it was written, and it had a profound impact on me. Then and now, I do see it as too pessimistic—I don't think broad "relinquishment" of whole areas of technology, which Joy suggests, is the answer—but the issues it raises were surprisingly prescient, and Joy also writes with a deep sense of compassion and humanity that I admire. ↩

Contents

[21] We do have to worry about state actors, now and in the future, and I discuss that in the next section. ↩

[22] There is evidence that many terrorists are at least relatively well-educated, which might seem to contradict what I'm arguing here about a negative correlation between ability and motivation. But I think in actual fact they are compatible observations: if the ability threshold for a successful attack is high, then almost by definition those who *currently* succeed must have high ability, even if ability and motivation are negatively correlated. But in a world where the limitations on ability were removed (e.g., with future LLMs), I'd predict that a substantial population of people with the motivation to kill but lower ability would start to do so—just as we see for crimes that don't require much ability (like school shootings). ↩

[23] Aum Shinrikyo did try, however. The leader of Aum Shinrikyo, Seiichi Endo, had training in virology from Kyoto University, and attempted to produce both anthrax and ebola. However, as of 1995, even he lacked enough expertise and resources to succeed at this. The bar is now substantially lower, and LLMs could reduce it even further. ↩

[24] A bizarre phenomenon relating to mass murderers is that the style of murder they choose operates almost as a grotesque sort of fad. In the 1970s and 1980s, serial killers were very common, and new serial killers often copied the behavior of more established or famous serial killers. In the 1990s and 2000s, mass shootings became more common, while serial killers became less common. There is no technological change that triggered these patterns of behavior, it just appears that violent murderers were copying each others' behavior and the "popular" thing to copy changed. ↩

[25] Casual jailbreakers sometimes believe that they've compromised these classifiers when they get the model to output one specific piece of information, such as the genome sequence of a virus. But as I explained before, the threat model we are worried about involves step-by-step, interactive advice that extends over weeks or months about specific obscure steps in the bioweapons production process, and this is what our classifiers aim to defend against. (We often describe our research as looking for "universal" jailbreaks—ones that don't just work in one specific or narrow context, but broadly open up the model's behavior.) ↩

[26] Though we will continue to invest in work to make our classifiers more efficient, and it may make sense for companies to share advances

like these with one another. ↩

[27] Obviously, I do not think companies should have to disclose technical details about the specific steps in biological weapons production that they are blocking, and the transparency legislation that has been passed so far (SB 53 and RAISE) accounts for this issue. ↩

[28] Another related idea is "resilience markets" where the government encourages stockpiling of PPE, respirators, and other essential equipment needed to respond to a biological attack by promising ahead of time to pay a pre-agreed price for this equipment in an emergency. This incentivizes suppliers to stockpile such equipment without fear that the government will seize it without compensation. ↩

[29] Why am I more worried about large actors for seizing power, but small actors for causing destruction? Because the dynamics are different. Seizing power is about whether one actor can amass enough strength to overcome everyone else—thus we should worry about the most powerful actors and/or those closest to AI. Destruction, by contrast, can be wrought by those with little power if it is much harder to defend against than to cause. It is then a game of defending against the most *numerous* threats, which are likely to be smaller actors. ↩

[30] This might sound like it is in tension with my point that attack and defense may be more balanced with cyberattacks than with bioweapons, but my worry here is that if a country's AI is the most powerful in the world, then others will not be able to defend even if the technology itself has an intrinsic attack-defense balance. ↩

[31] For example, in the United States this includes the fourth amendment and the Posse Comitatus Act. ↩

[32] Also, to be clear, there are some arguments for building large datacenters in countries with varying governance structures, particularly if they are controlled by companies in democracies. Such buildouts could in principle help democracies compete better with the CCP, which is the greater threat. I also think such datacenters don't pose much risk unless they are very large. But on balance, I think caution is warranted when placing very large datacenters in countries where institutional safeguards and rule-of-law protections are less well-established. ↩

[33] This is, of course, also an argument for improving the security of the nuclear deterrent to make it more likely to be robust against powerful AI, and nuclear-armed democracies should do this. But we don't know what a powerful AI will be capable of or which defenses, if any, will

Contents

work against it, so we should not assume that these measures will necessarily solve the problem. ↩

[34] There is also the risk that even if the nuclear deterrent remains effective, an attacking country might decide to call our bluff—it's unclear whether we'd be willing to use nuclear weapons to defend against a drone swarm even if the drone swarm has a substantial risk of conquering us. Drone swarms might be a new thing that is less severe than nuclear attacks but more severe than conventional attacks. Alternatively, differing assessments of the effectiveness of the nuclear deterrent in the age of AI might alter the game theory of nuclear conflict in a destabilizing manner. ↩

[35] To be clear, I would believe it is the right strategy not to sell chips to China, even if the timeline to powerful AI were substantially longer. We cannot get the Chinese "addicted" to American chips—they are determined to develop their native chip industry one way or another. It will take them many years to do so, and all we are doing by selling them chips is giving them a big boost during that time. ↩

[36] To be clear, most of what is being used in Ukraine and Taiwan today are not *fully* autonomous weapons. These are coming, but not here today. ↩

[37] Our model card for Claude Opus 4.5, our most recent model, shows that Opus performs better on a performance engineering interview frequently given at Anthropic than any interviewee in the history of the company. ↩

[38] "Writing all of the code" and "doing the task of a software engineer end to end" are very different things, because software engineers do much more than just write code, including testing, dealing with environments, files, and installation, managing cloud compute deployments, iterating on products, and much more. ↩

[39] Computers are general in a sense, but are clearly incapable on their own of the vast majority of human cognitive abilities, even as they greatly exceed humans in a few areas (such as arithmetic). Of course, things built *on top* of computers, such as AI, are now capable of a wide range of cognitive abilities, which is what this essay is about. ↩

[40] To be clear, AI models do not have precisely the same profile of strengths and weaknesses as humans. But they are also advancing fairly uniformly along every dimension, such that having a spiky or uneven profile may not ultimately matter. ↩

Contents

[41] Though there is debate among economists about this idea. ↩

[42] Personal wealth is a "stock," while GDP is a "flow," so this isn't a claim that Rockefeller owned 2% of the economic value in the United States. But it's harder to measure the total wealth of a nation than the GDP, and people's individual incomes vary a lot per year, so it's hard to make a ratio in the same units. The ratio of the largest personal fortune to GDP, while not comparing apples to apples, is nevertheless a perfectly reasonable benchmark for extreme wealth concentration. ↩

[43] The total value of labor across the economy is $60T/year, so $3T/year would correspond to 5% of this. That amount could be earned by a company that supplied labor for 20% of the cost of humans and had 25% market share, even if the demand for labor did not expand (which it almost certainly would due to the lower cost). ↩

[44] To be clear, I do not think actual AI productivity is yet responsible for a substantial fraction of US economic growth. Rather, I think the datacenter spending represents growth caused by anticipatory investment that amounts to the market expecting *future* AI-driven economic growth and investing accordingly. ↩

[45] When we agree with the administration, we say so, and we look for points of agreement where mutually supported policies are genuinely good for the world. We are aiming to be honest brokers rather than backers or opponents of any given political party. ↩

[46] I don't think anything more than a few years is possible: on longer timescales, they will build their own chips. ↩

Back to top

Privacy policy

Contents

EXHIBIT 5

Type text here

Stenographic Transcript
Before the

COMMITTEE ON
ARMED SERVICES

UNITED STATES SENATE

TO RECEIVE TESTIMONY ON ENTERPRISE SECURITY AND INFORMATION TECHNOLOGY OPERATIONS OF DEPARTMENT OF DEFENSE NETWORKS AND SYSTEMS

Tuesday, March 24, 2026

Washington, D.C.

ALDERSON COURT REPORTING
1029 VERMONT AVE, NW
10TH FLOOR
WASHINGTON, DC 20005
(202) 289-2260

TO RECEIVE TESTIMONY ON ENTERPRISE SECURITY AND INFORMATION TECHNOLOGY OPERATIONS OF DEPARTMENT OF DEFENSE NETWORKS AND SYSTEMS

Tuesday, March 24, 2026

U.S. Senate
Subcommittee on Cybersecurity
Committee on Armed Services
Washington, D.C.

The committee met, pursuant to notice, at 2:30 p.m. in Room SR-232A, Russell Senate Office Building, Hon. Mike Rounds, chairman of the subcommittee, presiding.

Committee Members Present: Senators Rounds [presiding], Rosen, and Reed.

OPENING STATEMENT OF HON. MIKE ROUNDS, U.S. SENATOR FROM SOUTH DAKOTA

Senator Rounds: Good afternoon, and welcome to this afternoon's Cybersecurity Subcommittee hearing on enterprise security and information technology operations of Department of Defense Networks and Systems. I want to begin by thanking our witnesses for appearing today before this subcommittee. Your testimony arrives at an inflection point for how the Department of Defense fights, decides, and wins.

This hearing is fundamentally about warfighting. The digital backbone of the Department of Defense is no longer a support function. It is a weapon system. Our tanks, ships, aircraft, and ground forces depend on connected, resilient, and secure networks to operate at the speed and precision across vast distances. Sensors all over the world will connect to shooters across the theater in a vast kill web that must operate faster and more efficiently than our enemies' systems. This is especially true for future conflicts, as we continue fielding software intensive systems designed for decision advantage and speed of action.

While quantity may have a quality all of its own, the ability to optimize our targeting and orient, decide and act more quickly than the enemy will likely decide the outcome of the next major war. The Department's cultural shifts

towards recognizing networks and information technology infrastructure as a warfighting platform is one I agree with.

Yet, I realize there is still a long way to go. Software can be developed and deployed in minutes, hours, or days. The bureaucratic processes governing how we acquire, certify and field it does not match that timeline, at least not yet. This reality is no longer merely an inconvenience. It is a strategic liability. We have talked long enough about solving these problems. This subcommittee wants to see them resolved.

Compounding the urgency is the condition of the infrastructure itself. Defense information technology budgets have long served as the bill payer absorbing cuts to fund near-term priorities. The result is a technical debt problem of historic proportions in both hardware and software. Our adversaries are not blind to this. Not surprisingly, they are taking advantage of this blunder. Every day we delay modernizing is a day we strengthen their hand in the cyber domain.

As we modernize, speed without security is not an improvement. We must be disciplined about not punting today's problems to the future. That means patching what we have while building with foundational security baked in.

This cannot be a government-only endeavor. The private sector can move with speed and provide capabilities the Department cannot match. But partnership only works when our processes and requirements are transparent and consistent. We owe our industry partners clarity and we owe our warfighters results. We must ask what kind of compute capacity, network resiliency, and data infrastructure our warfighters require, not just today, but as artificial intelligence is deployed at a scale across the force. If we cannot answer that question with specificity, we cannot build toward it.

Today, this subcommittee looks forward to hearing from both of you on where the Department stands in addressing these efforts. We want to understand how bureaucratic barriers are being dismantled, and how our processes are being made more welcoming to industry partners. It is especially important for us to hear whether our networks have the capacity, resilience, and performance our military requires for modern warfare against a capable adversary.

Thank you again, and now I would like to recognize the ranking member for her remarks. Senator Rosen.

STATEMENT OF HON. JACKY ROSEN, U.S. SENATOR FROM NEVADA

Senator Rosen: Well, thank you, Chairman Rounds. Ms. Davies, General Stanton, appreciate you being here. I want to welcome you, and Ms. Davies, congratulations on your recent confirmation and assumption of duties. And so, Ms. Davies, as you get settled into this position, we'd like to hear more about your priorities, get some of the ideas of where you're going to be placing emphasis on funding from fiscal year 2026 appropriations and any reconciliation funding you may have received as well.

And so, without discussing specifics, it would also be helpful to know where we might see significant budget swings for programs so we're not surprised when the fiscal year 2027 President's budget request is released. That request is already nearly 2 months late. It's not expected for a few more weeks, and that leaves little time in our NDAA process to delve into the details and it's critical for us to be able to do that.

General Stanton, I want to welcome you back. Thank you for your years of service. I hope you will also share your thoughts on all of these topics as they relate to your role as director for the Defense Information Systems Agency and commander for the Department of Defense Cyber Defense

Command. As the organization charged with running DOD networks, you have a unique operational perspective on how policy decisions from the Department CIO are translated, well, into action, but that also means that you have a slightly different view as to the resourcing and workforce needs. The implementation of technology for the warfighters different maybe than the development, right? The impact of technical debt on the ability of our department to modernize. You can speak to that in practical sense.

So, we also have a long list of long-term issues we want to make sure the Department is addressing so they don't get lost in the immediate priorities of the day or the specific priorities of this administration. For example, implementation of the cybersecurity maturity model certification process, improving the authority to operate, process for software to be placed on the DOD networks, and addressing the backlog of technical debt in our IT programs to make sure our networks are more modern, more resilient, and of course, more secure.

These are just a few of the areas where I think we have common cause and desire to work together to improve the Department for the long term. We want to work collaboratively with you on these issues. However, the Department has not been particularly timely or forthcoming

on information. So, for us to conduct our statutory obligation to oversee the Department of Defense, requires us to have open and honest conversations. There's definitely a trust deficit right now, but I think it's fixable if the Department can be more responsive to the requests of this committee and to our members. So, I hope we can use this hearing, both open and closed, to help us get started on the right foot.

And with that, I'm going to turn it back over to Chairman Rounds. Thank you.

Senator Rounds: Thank you, Senator Rosen. Today, we're pleased to have our two panelists with us, Kirsten Davies, who is the chief information officer at the Department of Defense, and Lieutenant General Paul Stanton, United States Army director, Defense Information Systems Agency commander, the Department of Defense Cyber Defense Command. We welcome both of you here, and we look forward to your opening statements. Your written statements will be a part of the record, but we would welcome your opening statements.

And with this, Ms. Davies, would you like to begin?

STATEMENT OF HONORABLE KIRSTEN A. DAVIES, CHIEF INFORMATION OFFICER, DEPARTMENT OF DEFENSE

Ms. Davies: Thank you. Good afternoon, Chairman, Ranking Member, Senator Reed, thank you for the opportunity to speak with you today about the Department's strategy to transform technology and cybersecurity into a decisive warfighting advantage.

Our focus is to enable data supremacy and decision dominance on the contested battlefields of today and tomorrow at the speed and scale our warfighters deserve. In the latest initiative in Secretary Hegseth's drive for efficiency and effectiveness, we are undertaking a bold transformation of enterprise IT and cybersecurity program, unifying these capabilities under the Department's Chief Information Officer. Through this effort, we will eliminate inefficient spending, reduce technical debt, accelerate modernization, drive consistent and up leveled cybersecurity, and unleash data and innovation from the core to the edge across our joint forces.

Leveraging my oversight of DISA, the NSA's Cybersecurity Directorate, and the Department's Cyber Crime Center, we are working with the military services, joint staff, combatant commands, and defense agencies across four transformation pillars. I'll provide a few highlights here.

Under pillar 1, the Enduring Digital Foundation, we're transforming our network infrastructure and communications transport, which extend from undersea cables to terrestrial fiber to advanced satellite capabilities, connecting everything from the home front to the tactical edge. This foundation supports every warfighting system and our global installations. We're driving continual modernization, expansion, and hardening, as well as broad 5G usage and data center modernization. We're evolving our cloud strategy in JWCC Next, and we're also leading a proactive approach to spectrum management and advancing PNT efforts, ensuring ready and resilient capabilities that enable American warfighting dominance.

Under pillar 2, agile digital capabilities, reflecting some of your comments, Senator Rounds, we are expanding our and maturing digital offerings. We're accelerating delivery of software and SaaS services, and standardizing data architectures, streamlining our data flows. We're shifting from slow legacy software development to modern agile delivery, driving interoperability by design, and delivering applications and analytics at the speed of relevance. We're driving extensive defense business systems work, whether modernizing or sunsetting those systems to enable clean audits and reduce wasteful spends. We're also deploying

mission partner environment as persistent, secure environments where trusted partners can be rapidly integrated.

Under pillar 3, cybersecurity for the warfighting ecosystem, in alignment with President Trump's National Security Strategy and the National Defense Strategy, we're moving from checklist-driven compliance towards unified, holistic, risk-based approach. We will emphasize automation and dynamic and continuous monitoring.

We will drive risk reduction rather than burdensome paperwork, focusing on anti-fragility and resilience through a holistic blend of streamlined processes, advanced technologies, and skilled people. We're refining the authority to operate, process, and accelerating our deployment of Zero Trust principles. We're also refining our risk management process and reigniting the debate to align with the Secretary's arsenal of freedom initiatives.

Finally, through pillar 4, skills and partnerships, we recognize that people are our decisive edge. As part of our transformation, we're reviewing IT and cybersecurity roles to ensure clear responsibilities, accountability for outcomes, and a bias for action. We're leveraging your provisions in the fiscal year 2026 NDAA to enhance recruitment and retention of cyber professionals and expand

competitive compensation.

We will be launching an expanded top tier certification program in partnership with industry and academia, which will offer upskilling and reskilling to our warfighters, from new recruits to seasoned service members. And because we do not fight alone, we're doubling down to influence the digital transformation efforts of our allies and partners, which will better enable all of coalition force readiness.

As we advance this bold strategy, thank you for your continued interest in and support of IT and cyber security, and for the resources you provide to protect national security. The race for data superiority and decision dominance is won or lost every day, and this strategy is a key part of how we, together, ensure the technology race is won by the United States. Together, we will ensure the resilience, readiness, and lethality of America's warfighters across every domain.

I look forward to your questions.

[The prepared statement of Ms. Davies follows:]

Senator Rounds: Thank you, Ms. Davies. Lieutenant General Stanton.

STATEMENT OF LIEUTENANT GENERAL PAUL T. STANTON, USA, DIRECTOR, DEFENSE INFORMATION SYSTEMS AGENCY/ COMMANDER, DEPARTMENT OF DEFENSE CYBER DEFENSE COMMAND

General Stanton: Chairman, Ranking Member Rosen, Senator Reed, thank you for the privilege of appearing before you today to explain the fundamental shift we are making to ensure that our weapons system, the Department of War Information Network, provides our warfighters with decision advantage. I'm honored to represent the highly skilled and dedicated professionals of the Department of War, Cyber Defense Command, and the Defense Information Systems Agency that design, build, secure, operate, and defend our environment.

We must deliver a secure, standardized, resilient, and efficient architecture that supports combatant commands. Combatant commands execute warfighting. Nested within the Department of War CIO's vision, the Defense Information Systems Agency has a responsibility to provide functionally relevant capability that aligns with the time and tempo of the warfighter's mission. As the Department of War Cyber Defense Command, we have an added responsibility to ensure that our systems and data are properly defended against continuous and sophisticated attacks.

Combining these two responsibilities, and said simply,

we must get the right data to the right place at the right time, such that our commanders make better and faster decisions than our enemies. We are a sub unified command and a Department of War Combat Support Agency. Our mission, and therefore our culture is to support warfighting. We are fully engaged to move and maneuver the network and our data according to the changing conditions of the operating environment. We design, build, secure, operate, and defend in lockstep with commanders at echelon. We campaign to execute our missions and defeat our adversaries. We present and defend the architecture so that commanders can fight. We are doing so right now in Operation Epic Fury.

But we cannot rest. We must transform ourselves and the means by which we support to leverage the most modern and effective technology. We are in a perpetual state of continuous modernization; new solutions, artificial intelligence, commercial SATCOM, mobile data centers. They emerge at industry's pace, and we must integrate them into our architecture and missions at speed.

Further, we must be prepared to fight alongside our partners, sharing data across warfighting functions within decision cycles. The coalition information environment, as recently prototyped during an exercise in the Indo-Pacific theater, is a cornerstone of our approach. We know that our

adversaries, our enemies, are watching us and will certainly attempt to delay or degrade our decisions, and we will defeat them.

We are developing solutions that are secure by design, incorporating perimeter defenses that defeat known attack vectors and employing Zero Trust to detect, bound, and defeat new and novel tradecraft. Our internet access point and our cloud-based internet isolation warfighting systems continuously adapt to new threats at our boundary. Our Thunderdome implementation of Zero Trust is proven, expanding rapidly as we transition defense agencies and field activities into DODNet. We will use these tools and the associated data in an informed, productive, efficient, and speedy manner, automating our defenses and employing human tradecraft for advanced analysis.

We prioritize our defenses on what matters. In order to preserve decision space for commanders, we must defend the critical systems upon which they are dependent. Our approach employs a mission thread defense that we nest and plan amongst commanders and their staffs for synchronization. We align our cyber defenses to how their systems employ and move data, ensuring that they have confidence in the data upon which they make their decisions. Our defenses and our strategic goals require optimization.

We have to see ourselves holistically. We have made and continue to make significant progress in sensing, logging, aggregating, and analyzing our data accordingly. Our data analytics support cell is employing our common data analytics platform to continuously run queries and analytics that optimize performance and support defensive operations. CDAO feeds into USCYBERCOM's joint cyber warfighting architecture for enrichment with classified intelligence, and coordination with offensive cyber forces for speed and lethality.

Our success is the innovation, talent, and motivation of our workforce. Readiness is a requirement. A trained and ready force has confidence to act with disciplined initiative. Demonstrated confidence leads to trust to make decisions at speed. When we combine our transformational architecture with a talented and trained workforce, we are postured to meet our requirements. This is an imperative. The effectiveness of the DoWIN is inextricably linked to our missions and our Nation's defense. With the continued support of the committee, we will preserve the decisive advantage.

We look forward to your questions. Thank you.

[The prepared statement of General Stanton follows:]

Senator Rounds: Thank you, General Stanton.

Normally, we would begin with 5-minute rounds, and I would start, Senator Rosen, the ranking member, would be second, and then we'd move back and forth. But we also have the ranking member here. And if you would like to --

Senator Reed: No sir. Regular order --

Senator Rounds: Okay. Regular order it is.

Senator Rosen: You heard it from the Ranking Member.

Senator Rounds: Very good. Well, then I will begin. Ms. Davies and General Stanton, the ability of our warfighters to operate in a contested or a degraded environment is directly tied to how resilient and modern those networks are. Where does the Department stand on network modernization, and are our warfighters confident they can operate if those networks are attacked or denied? Ms. Davies.

Ms. Davies: Thank you, Senator Brown, for that great question. I'll allow General Stanton to get into a few of the details. But as I reflected on pillar 1 of our transformation strategy, this is active work that we are doing right now that DISA has been conducting for quite some time under General Stanton's leadership and previous leadership as well. This is a key area of focus for us.

Our forces are globally located. Our joint forces are

globally located. We are currently, obviously, in a mission right now with partner forces as well. And the resiliency and the efficacy of the traffic of our network is quite critical to that. It's a key focus area for us, Senator.

Senator Rounds: General Stanton.

General Stanton: Senator, thank you for the question. And with great support from Congress, we have an initiative that we reference as design, security, and resiliency. So, the Defense Information System Network, where we focus on undersea cables, increased bandwidth for terrestrial fiber, multimodal satellite communications capabilities we refer to as an agnostic peering gateway that allows us to communicate over military SATCOM waveforms, but also via commercial SATCOM capabilities.

As our forces move into theater, as they currently reside in theater, we have a primary alternate contingency and emergency plans put into place. We're never single threaded on any capability as we enter into the fight such that if we suffer degradation, we have fallback capabilities. We're seeing that in spades currently, operating across terrestrial space based and undersea capabilities.

Senator Rounds: So, recognizing that we're in an unclassed environment, we'll go into a classified

environment when this when this meeting is done, specifically, I think what you're indicating is there's a couple of different areas where we may have, some challenging communications problems. You mentioned, undersea cables, you mentioned space-based assets and so forth. Are those perhaps the most challenging that we're going to face that we can talk about in this environment today?

General Stanton: Well, Senator, I think it's the combination and the fact that as our warfighting formations are outfitted with capability. We never isolate down to a single mode of transport. We ensure that we have the ability to route terrestrially. We hit two peering points such that we can leverage undersea cables. We're never bounded by a single undersea cable. We always have a plan to route around or have an alternate path.

And then, the proliferation of space-based assets, specifically in the commercial world, is really game changing technology to give us leap over capability if and when we do suffer a degradation.

Senator Rounds: It's an interesting lead in on it. Then for us to talk a little bit about the reason why we no longer talk about a kill chain. We talk about a kill web, and that is because we have multiple avenues to move from a

spotting system back into where you actually have the ability to trigger a weapon. So, multiple ways to get the communications from point A to point B, not simply one line. Fair way of looking at it?

General Stanton: Precisely. Yes, sir.

Senator Rounds: Ms. Davies, many smaller Defense Industrial Base companies lack the internal security capability to defend themselves against a sophisticated state actor. What is the Department doing to reach that tier of the industrial base, and is voluntary participation in government support programs getting us there?

Ms. Davies: Senator Rounds, this is a key focus area for me as well. I think we have focused largely on confidentiality of data in the past. I know that there has been some burdensome requirements that have been placed on large and small businesses alike. In the new transformation, one of our key pillars is going to be working directly with the Defense Industrial Base. Their resiliency is our resiliency. Their security is our security.

But it needs to make sense. We've heard Secretary Hegseth talk about reducing the burdens to entrance, allowing entrance, new entrance for smaller companies as well. This is a key focus area for us, whether it's

providing guidance, providing principles for them to follow. It's coming alongside them and partnering them, but it's also tailoring these requirements so that they are effective for the arsenal of freedom that we are driving.

Senator Rounds: Thank you. My time has expired. Ranking Member Rosen.

Senator Rosen: Thank you, Chairman Rounds. I'm going to just say something and I'll ask for more details in the classified briefing. Just building on what Senator Rounds said, I want to hear a little bit about the lessons you learned from the recent military operations in Venezuela and Iran that relate to the DOD networks in terms of showing us a little bit of stress testing in the real world, right? We're in conflicts in both places, and what we've learned about what we might be changing for future protracted conflicts. So, we'll save that one. Just put that out there.

General Stanton, I want to talk a little bit about IT support during a war. So, the Defense Information Systems Agency, you're a combat support agency for the Department of Defense. So, I'm hoping that you can explain for all of us what that means, practically. Given our current posture in the Middle East, what does your agency do during war time that could be different than what it does during peace time,

if you could elaborate on that?

General Stanton: Yes, ma'am. Absolutely. Thank you for the question. So, we are at war, and we're executing Operation Epic Fury currently. Which means that every day inside of our operations center, the Defense Information Systems Agency, and Cyber Defense Command, get together to ascertain what has transpired in the context of the network, what assets are still up and running, what assets need to be resolved? How do we route around problems? How do we dynamically solve problems with emergent technology and do so rapidly.

From a DISA perspective, a lot has to do with network transport. And so, it -- where are the terminals located? How do we get them to the right spot? Do we need to lease a new circuit on the fly in order to ensure that critical data gets from point A to point B? These problems present themselves in real time, and inside of our ops center we are dynamically solving and developing resolution.

Senator Rosen: Thank you. I'm going to move on to you, Ms. Davies, because we want to talk a little bit about the Joint Warfighting Cloud Contract. And the Joint Warfighting -- it's a mouthful. The Joint Warfighting Cloud Contract -- do not try to say that quickly -- we'll just say the JWCC, it's a little bit easier, is reaching a point soon

where it's going to be need to be recommitted, and there'll be a need to be a replacement contract. And so, what lessons has DOD learned from the JWCC that we might see next in an updated JWCC, and how do you see AI impacting your decisions?

Ms. Davies: Thank you for the question. It is a mouthful, isn't it?

Senator Rosen: It is, it is, yes.

Ms. Davies: We are expanding JWCC into a unified cloud marketplace, integrating additional providers, embedding financial operations, automation, and multi-cloud management to enable enterprise-wide cost control and interoperability. I can speak from being new in the role and seeing that there are contracts everywhere, and different points of authorization with different cloud that's happening.

One of the key areas that we need to be looking at from a multi-pronged approach is, is this the most efficient way to be driving cloud compute? It's not. We need to be continuing on in the JWCC Next. Is it the most -- is it the best way to see the spend. It is not. So, JWCC Next is going to provide us that financial transparency that's there, and it's also going to provide General Stanton and his team the ability to do better defense across all this, because we're going to know where all of the cloud compute

is, and that's key for us in asset identification and asset security.

Senator Rosen: Thank you.

And then I'm going to continue with you, Ms. Davies, because I want to talk about the enterprise chief information officer collaboration. Right? So, DOD, you're like -- you're saying you're just a vast conglomeration of networks, clouds, operating cultures, systems, you name it. It's difficult to craft a one-size-fits-all policy, although you can set standards, and you can at least lay out the templates for it. You can map out where everything is, essentially, but in my view, there are benefits to having each of the military departments and defense agencies having their own CIOs so that they can tailor technology and policies to the needs of the various organizations.

So, could you describe for us your relationship with your CIO counterparts in the military services and defense agencies, and are there any lessons, helpful or otherwise, you might have picked up in your time so far?

Ms. Davies: Some great lessons indeed. I'm holding regular meetings with my military department counterparts as well as the DAFA counterparts. I'm learning where the operational efficiencies are, where the centers of excellence and expertise are, also where the gaps are. So,

I think there's varying levels of competencies, varying levels of operational cadence that are there. And one of the things that we will be getting after with this new strategy is to take a hold of those rising tides, raise all ships to make sure that we're all pointing in the same direction and focused on operational excellence in cyber defense.

Senator Rosen: I yield.

Senator Rounds: Senator Reed.

Senator Reed: Well, thank you, Mr. Chairman, Madam Ranking Member. I thank the witnesses not only for being here today, but for your dedication to our warfighters. Thank you.

Ms. Davis, I'm sure you're aware of the recent decision by the secretary to designate Anthropic as a supply chain risk. Indeed, you, yourself, signed out a memo on March 6 directing the removal of Anthropic from DOD systems within 180 days. However, the committee still has not heard the rationale from the Department about why the designation was made, nor received a full notification, which is required under law by Section 3252 of Title 10.

And this notification requires a summary of the risk assessment and a summary of the basis for the determination, including what less intrusive measures were considered, and

why they were not reasonably available to reduce supply chain risk. We have not received that information, yet, it is required under the law. So, first, are you aware of the actual reason in designating Anthropic a supply chain risk?

Ms. Davies: Senator, thank you for the question. I was involved, as many of my counterparts were, in this collaborative decision-making process that followed the regulatory requirements.

Senator Reed: Well, why was it done?

Ms. Davies: Sir, we're in active litigation right now, so I won't go into the details of it. We have reached out and offered a briefing for your offices into the depths of it. I will say that there are some -- the filing in the California court is available. We have made that available, but it was only available to us this morning. So, we did provide that over with the risk analysis. That was a part of it.

Senator Reed: It's just interesting that you would file required documentation for the California court before complying with the law, and filing it, and sending it to us. I don't believe it's been sent to us officially or unofficially. Why haven't you, the Department, complied with the law.

Ms. Davies: Senator, I'm aware that the -- our

colleagues in Legislative Affairs have followed the regulation of what they were in -- what they were supposed to provide. I do know that we followed all of the steps of the regulatory requirement of the Title 10, 3252.

Senator Reed: Well, I don't think we've received it. We have not received it.

Senator Rounds: Just in checking with staff, I do not believe that we have received it at this time.

Senator Reed: Thank you, Mr. Chairman.

But as you pointed out, a California court has received it because of the litigation. Let me just -- one additional question is, are you aware of any estimates that were made as to the potential cost impact on DOD uses for removing and replacing Anthropic from DOD systems, or the cost to replace Anthropic with another large language model.

Ms. Davies: Senator, I'm aware of the risk analysis that was conducted as a part of that, and I'm aware that we have also constructed our data architectures to be able to be interoperable with a variety of different AI capabilities. And so, the deep assessment of replacement of that, I'm unfamiliar with right here, I can take that away as an action for you.

[The information referred to follows:]

[SUBCOMMITTEE INSERT]

Senator Reed: It would be appreciated because the idea of the scale and the magnitude of the disruption would be helpful. Further, it's my understanding that Anthropic's Claude system is being used today in Iran in our military operations. Is that true?

Ms. Davies: Without going into the details in this forum, Senator, the use of the system is active right now. This is also why we provided for a measure of time we felt was reasonable, as well as an exception process for removal of the Anthropic systems.

Senator Reed: It just seems odd that you would continue to use a system which you determined to be a supply chain risk. Does that strike you as odd?

Ms. Davies: Senator, at no time, in any way, will we interfere with the success, the lethality, and the resilience of our warfighters. And for that reason, we've provided what we feel is a reasonable amount of time for those systems to be replaced. We can -- I can also say that according to -- you know, with President Trump's great leadership, we have a number of technology companies that have come to the table wanting to do business with us as the Department of War and across the U.S Government. So, we know that we've architected this appropriately in order to use competitive advantage as well.

Senator Reed: Thank you very much, Ms. Davis. General, thank you.

General Stanton: Sure.

Senator Rounds: Thank you, Senator Reed.

Let me just follow-up on that for just briefly here. My understanding is that there has been a 180-day notification with regard to Anthropic. I presume, and you can correct me if I'm wrong, but I presume that there is additional time frame here in which there is the possibility of additional negotiations that can occur during that time period, recognizing just what a significant change this would be to the Department with the reliance right now on the Anthropic product at this time. Fair enough to say?

Ms. Davies: I'm not sure what part of the question to answer for, Senator Rounds. Let me let me try to unpack that for you. We have architected our data insomuch as we can deploy multiple types of AI across our data. That's something that the -- General Stanton and the DISA colleagues have been very careful about. Number one.

Number two, we have provided what we feel is an appropriate amount of time to remove the Anthropic systems in accordance with the designation by the secretary of the supply chain risk designation in and of itself. Does that answer your question?

Senator Rounds: Yeah. Except that I think the other entities that we are looking at, many of them have also indicated that they have Anthropic within their systems as well. And what I'm looking for is, is the possibility that as this discussion goes on in a business-like manner, I'm assuming it will be done in a business-like manner, that there are opportunities for additional negotiations to continue to occur?

Ms. Davies: Senator, I will defer that to my colleagues in the legal department who are undergoing that active litigation right now, and we will certainly bring a report back to you.

Senator Rounds: That's fair. And I do think it'd be fair to say that I think the committee as a whole, and I can't speak for the chairman, but at least with regard to the subcommittee, this is something that we have a real interest in, and we will want to get in a classified setting probably deeper into the details at some point when you're prepared to share that with us -- with the appropriate personnel.

Ms. Davies: Senator, we'll take that for action, absolutely. Thank you.

Senator Rounds: Thank you.

Let me go on a little bit here. I'm just curious,

General Stanton, one item that we've talked about is deterrence. And as we will use our offensive capabilities, one of the reasons why you use offensive capabilities is to deter future attacks, and to let people know that you know who they are, we know where they are, and we do have access to some very exquisite capabilities to stop them from actually using kinetic activities or kinetic systems.

Can you talk a little bit, in this open session, just so that the American public will understand, just kind of some of the things that we have the ability to do now that we've actually utilized some of them, and are -- the bad guys know that what we can do? Can you talk just briefly about that, just to share with the American public what their taxpayer dollars are buying?

General Stanton: Yes, Senator. And I look forward to having a more robust conversation in a classified setting.

Senator Rounds: I understand, but the public can't see that. And like I said, I don't want to do any damage to our ability to do it in the future. But I think it's fair for deterrence's sake, to maybe talk a little bit about what our capabilities are, if that is acceptable.

General Stanton: Yes, Senator. So, I think I'll address it in two principal ways. The first is there's a deterrent effect associated with cost imposition. If you

make it really hard for the enemy to attempt to achieve the effects that the enemy intends, then it is a cost imposition. The enemy has to spend more money, more time, develop more resources, and apply it in ways that that the enemy may not have been prepared. That's a cost imposition.

In addition, we have offensive cyber capabilities. And we have offensive cyber capabilities that can, respond at speed to evidence of adversarial activity beyond the bounds of our U.S. networks. So, when we see operations in foreign space as actioned by our cyberspace foreign adversaries, we have the capabilities to deny them those resources.

Senator Rounds: Fair to say we can deny them the ability to communicate in some cases?

General Stanton: Yes, Senator.

Senator Rounds: Fair to say that we can sometimes make it so they can't see what they want to see on their systems --

General Stanton: Yes, Senator.

Senator Rounds: -- to know what's going on in their in their part of the world? Fair to say that we can make them see things that maybe aren't even there today.

General Stanton: So, the ability to deny access to systems, the ability to manipulate data to get inside the decision cycle of the adversary, are all the art of the

possible in techniques developed in support of offensive cyber operations.

Senator Rounds: All of which means that our young men and women are then safer when they go in harm's way, because we limit the adversary's ability to respond to our young men and women who are on the battle front?

General Stanton: Unequivocally --

Senator Rounds: Thank you.

General Stanton: Unequivocally -- yes, Senator.

Senator Rounds: Thank you. Ranking Member Rosen.

Senator Rosen: Thank you.

I want to in the classified, I'm going to ask a little bit more about how you've architected -- it's a new verb, architected -- your data to feed into many different, I would assume, large language models or the like. I think that it's very interesting to me.

But for this open session, I want to talk about the authority to operate process, because I'm encouraged by the Department's continued support for improving cybersecurity and supply chain risk management, of course, and you've made progress towards reforming and accelerating acquisition, testing, authorization of commercial software. But more, of course, can always be done to streamline the authority to operate, ATO, our single process into a single department-

wide accreditation for secure software providers, streamlining the process.

So, can you talk about the status of your efforts to streamline that process, of what actions you're taking, and what plans your office may have to establish and encourage reciprocity for ATOs between individual service branches and department components. So, pulling all that back up to the center.

Ms. Davies: Ranking Member, a great question. The ATO process is part of the broader risk management framework, as you are very familiar with. Right now, we have a very static snapshot in time with regards to risk management, and that needs to be moved to a much more dynamic framework and process, which includes inheritance of assessments that are conducted somewhere else across the Department on a piece of software. That inheritance can then travel to a new department that wants to leverage that piece of software, that inheritance of all the testing and the scalability and all of those types of things.

The ATO process as a piece of that also needs a reform. We're finding that it's very difficult for people to actually grab that inheritance over. It's very difficult to understand the work of that risk management framework because it's broken, it's fragmented, and it's static. So,

what we're doing is we're looking to do a lot more automation across this, having dynamic repositories of this information, and the testing in and of itself.

This work of the RMF framework, as well as the ATO processes, will be sitting underneath the Department's chief information security officer, who was presidentially appointed just a few weeks ago. But he's right on top of it and going to be working very actively with me to make sure that we're reforming that appropriate to the risk of the actual work that needs to be done.

Senator Rosen: Well, I understand what you're saying; how important it is to be more dynamic, and sometimes less static, but I'm also well aware of the vulnerabilities at places when you are quickly dynamic without the proper audits and controls over that as well because you never want to sacrifice speed. And I'm not saying static is always the way to go, but you have to be very careful about that dynamic architecture as well because it can create vulnerabilities, because moving at the speed of light, or sound, or nano second, whatever you want, is -- has risk in there as well. So, I hope that you're building in a good audit process review of how that's working, so in that speed --

Ms. Davies: Yes --

Senator Rosen: -- we will get to that.

Ms. Davies: -- and we're -- it's a great point. We're going to be bringing in a lot of industry-good practices across this as well, where we've learned to do a lot of the automation of processes, less paperwork, more dynamic checking across the software design lifecycle. We can do a lot of code scanning, code reviews. Those are the types of automation that will help us speed these things up while we're compiling appropriate data repositories for that constant risk checking.

Senator Rosen: Right. Because if you do it too fast, once a piece of bad code gets in, it's already replicated quickly across your system before you may have caught the bug, so.

Ms. Davies: That's right. You understand the risk process very much.

Senator Rosen: Yes, I think I do, but thank you.

I'm going to talk a little bit in my -- I just asked you a question about artificial intelligence, the validity. We talked about Anthropic. I know we're going to go talk some more about this, the validity of their output critical to the effectiveness of what we do. As we acquire and deploy more systems, artificial intelligence systems, warfighting -- you said we're in a war -- but it's going to

help our situational awareness, our decision-making, and we must secure these systems and the data that powers them. The data.

That's why I want to talk about how you architect your data. Data is power if you're smart enough to analyze it and use it. It's all about how you use it. The data is key, and so it's important and critical to maintain the trustworthiness, the integrity of all of that. Avoid any corruption, malicious manipulation. And so, how are you kind of -- as much as you can say here, what are you doing to ensure that you're leveraging existing commercial solutions, without making us vulnerable.

Ms. Davies: Yeah. Thank you, Ranking Member. In July of 2025, my office published the DOW AI-Cybersecurity Risk Management Tailoring Handbook. So, we provided some great guidance across that. This is an ever-evolving framework or competency, I would say. We've been tackling this in the industry across the last, I'd say, 5 to 7 years, providing appropriate guardrails, ethics, security across this, looking at the weights of models, as well as hallucinations, to try to reduce all of these factors that are there. We're going to be continually evaluating this as we go through.

We have provided strong guardrails. We're doing risk management assessing on a regular basis across this. And

so, this -- we will continue to have conversations around this because it is an evolving category of software, if we want to call it that.

Senator Rosen: Thank you.

Senator Rounds: I think at this time, we will conclude the open portion of today's Cybersecurity Subcommittee hearing, and as all of you know, we'll be reconvening here in a few minutes. We've got a vote at 3:15 that's scheduled. Matter of fact, three of them, but it'll give us an opportunity to make our first vote.

We'd like to reconvene at 3:30 down in 217, in the SCIF, for a classified portion of this. And then, for the information of members who will not be joining us for the closed briefing, questions for the record will be due to the committee within 2 business days of the conclusion of the hearing.

[The information referred to follows:]

[SUBCOMMITTEE INSERT]

Senator Rounds: And with that, I want to thank you for this open session. We look forward to visiting with you again very shortly, beginning at 3:30 in the closed session in the SCIF.

And with that, the subcommittee meeting is adjourned.

[Whereupon, at 3:16 p.m., the hearing was adjourned.]

EXHIBIT 6

Anthropic

US National Security Agency using Anthropic's Mythos for cyber attacks

Arrangement comes as AI lab is locked in legal battle with Pentagon over Claude model

Anthropic has installed about half a dozen staff within the agency © Sebastien Bozon/AFP/Getty Images

Cristina Criddle in San Francisco and **Demetri Sevastopulo** in Washington

Published JUN 4 2026

Anthropic is helping the US National Security Agency deploy its powerful Mythos AI model for offensive cyber operations, embedding engineers inside the agency despite an ongoing legal battle with the Pentagon.

The San Francisco-based company had installed about half a dozen staff within the NSA as so-called forward-deployed engineers to guide the use of the technology and customise models for specific applications, two people familiar with the arrangement said.

It remains unclear whether Anthropic's engineers are assisting the NSA in active operations. However, one person close to the situation said Mythos would be useful for infiltrating the networks of nations such as China or Iran.

"The best way to build a good defence is to build a good attack," said a person close to Anthropic, who argued that adversaries were probably building their own AI-driven offensive technology. "If [Mythos] is not used to build attack agents, adversaries will find a way to do it."

The arrangement comes despite the Silicon Valley start-up's legal battle with the defence department, which includes the NSA, over how its technology is used in warfighting.

Anthropic tried to restrict the US government's use of its Claude AI models for mass surveillance of American citizens and lethal autonomous drones. That led the Pentagon to label the AI lab a "supply-chain risk", a first for a US business. Anthropic has sued over the designation, which could force it to end contracts with organisations that work with the US military.

Since the dispute erupted, Anthropic unveiled Claude Mythos, triggering concern among governments, financial institutions and IT companies over its ability to detect and exploit software vulnerabilities.

Earlier this week, Anthropic announced it would distribute Mythos to 150 organisations across 15 countries, vastly expanding access beyond the US and UK. It was initially made available only to a select number of US-based organisations when it launched in April.

The move comes after Anthropic filed for an initial public offering that could value the company at more than $1tn, underlining the growing commercial and geopolitical significance of AI to national security.

Since Mythos's release, Anthropic rival OpenAI has released a model with similar capabilities. Experts have warned that advanced AI models enable hackers to exploit computer systems but can also serve as a powerful tool against adversaries.

Anthropic and the US defence department both declined to comment.

The company has worked closely with the US government on the wider rollout, while the Trump administration has embedded the model across agencies.

In February, the FT reported that the Pentagon was seeking to create AI-powered cyber tools to identify infrastructure targets in China as part of an effort to improve US capabilities in any future military conflict with Beijing.

On Tuesday, President Donald Trump signed an executive order outlining a voluntary framework in which AI companies can submit their new models for security reviews before they are publicly released.

The order directs federal agencies to develop methods to evaluate the cyber capabilities of AI systems and to establish an "AI cyber security clearinghouse" to share information and strengthen defences.

EXHIBIT 7

White House and Anthropic Hold 'Productive' Meeting, Aiming for a Compromise

Friday's meeting at the White House followed the introduction of Anthropic's powerful new artificial intelligence model, Mythos, which U.S. officials believe could be critical for security.

Listen · 5:15 min

By Julian E. BarnesSheera Frenkel and Tyler Pager
Julian E. Barnes and Tyler Pager reported from Washington, and Sheera Frenkel from San Francisco.

April 17, 2026

Anthropic's chief executive met with White House officials on Friday for discussions that both sides described as "productive," as the Trump administration works to forge a compromise that would bring the artificial intelligence company's technology back into the government, according to U.S. officials and others briefed on the matter.

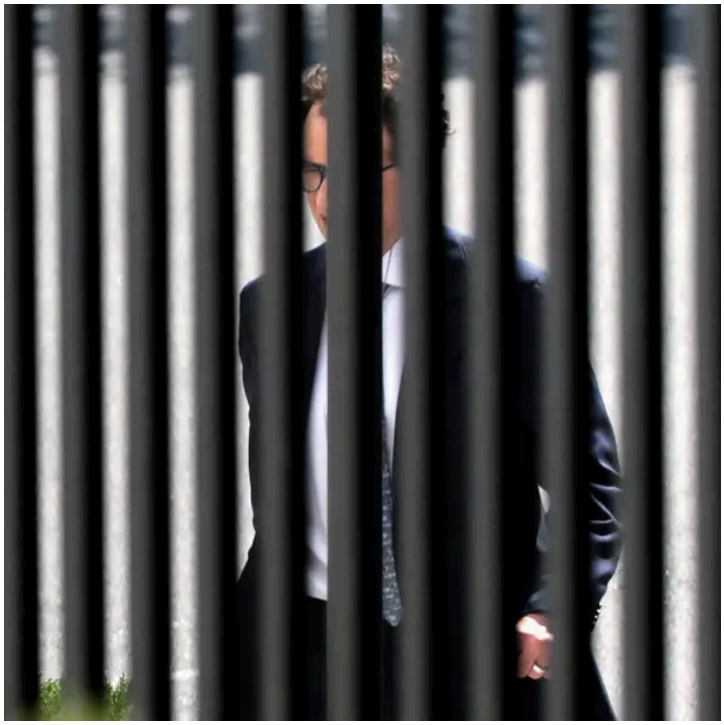
Dario Amodei, Anthropic's chief executive, met with the White House chief of staff, Susie Wiles; Treasury Secretary Scott Bessent; and others, the people said.

The meeting was "both productive and constructive," the White House said in a statement, with conversations about how to collaborate and address the challenges of A.I. while exploring "the balance between advancing innovation and ensuring safety." Anthropic said in a statement that the discussions had touched on "key shared priorities" such as cybersecurity and A.I. safety.

Anthropic had been effectively cut off from working with the federal government after battling with the Pentagon this year. The two sides disagreed in negotiations over a $200 million contract about the use of A.I. in warfare, with the Pentagon later designating Anthropic a "supply chain risk."

Friday's meeting was a potential first step to a deal. If officials reach a compromise with the company, it is likely to exclude the Pentagon, two officials said.

Some White House officials have argued that a fight with Anthropic is counterproductive and denies the United States some of the most powerful tech tools, officials said. This follows Anthropic's unveiling on April 7 of a powerful new A.I. model, Mythos, which is capable of identifying security vulnerabilities in software. Officials believe access to the model — which Anthropic has made available only on a limited basis — is critical to help protect government networks from cyberattacks.



Dario Amodei, Anthropic's chief executive, arriving for the meeting with officials at the White House on Friday. Jessica Koscielniak/Reuters

Some White House officials were also frustrated that other officials had failed to find a way to de-escalate the contract fight with Anthropic, especially given the potential for Mythos to wreak havoc on computer systems, officials said.

"We look forward to continuing this dialogue and will host similar discussions with other leading A.I. companies," the White House statement said.

Anthropic added that the meeting had reflected its "ongoing commitment to engaging with the U.S. government on the development of responsible A.I." and that it was "looking forward to continuing these discussions."

The White House meeting was reported earlier by Axios.

Neither the Pentagon nor Anthropic had shown much willingness to settle their dispute. During contract negotiations, Anthropic had sought assurances that its powerful A.I. models — which until recently were the only ones allowed on classified computers at the Defense Department — would not be used for commanding autonomous lethal weapons or surveilling Americans. In response, the Pentagon said no private contractor could tell it how to use the technology.

The fight came to a head on March 5, when Secretary of Defense Pete Hegseth labeled Anthropic a supply chain risk. The designation was previously used only for foreign companies that the government believed posed a risk to national security. Anthropic later sued the U.S. government over the designation in courts in California and Washington, D.C.

Secretary of Defense Pete Hegseth on Thursday. Anthropic and the Pentagon disagreed over the use of its technology in warfare earlier this year. Pete Marovich for The New York Times

Since then, Anthropic has told government officials that it is willing to provide access to Mythos to help agencies use the tool to find and fix potential vulnerabilities in software and computer networks.

At the Pentagon, engineers who work with Anthropic are petitioning the department to keep using the company's technology, according to two officials who have participated in meetings about A.I. technology this month. While the engineers were not averse to using new A.I. models from other companies, they did not want to be cut off from Anthropic's technology, which is used to analyze intelligence and handle sensitive data.

The engineers also urged the Pentagon to update to the newest Anthropic models available, the two officials said. Because the models are held on systems housed within the Pentagon, Anthropic cannot update them unless given access by defense officials, they said.

Senior Pentagon officials declined to talk about Anthropic, citing the ongoing lawsuits. The department is using an older version of Anthropic's Opus A.I. model than it would have had the dispute not occurred, current and former defense officials said.

Other officials said they were pressing ahead with bringing models from OpenAI and Google to Pentagon computers, including a more advanced version of Google's Gemini that will be online for military use in the coming days.

After Anthropic sued the U.S. government over the "supply chain risk" label, a judge in the U.S. District Court for the Northern District of California temporarily stopped the Pentagon from enforcing the designation in late March. In a separate ruling in the U.S. Court of Appeals for the District of Columbia Circuit on April 8, judges denied Anthropic's request to stop the Pentagon from labeling the company a supply chain risk.

Julian E. Barnes covers the U.S. intelligence agencies and international security matters for The Times. He has written about security issues for more than two decades.

Sheera Frenkel is a reporter based in the San Francisco Bay Area, covering the ways technology affects everyday lives with a focus on social media companies, including Facebook, Instagram, Twitter, TikTok, YouTube, Telegram and WhatsApp.

Tyler Pager is a White House correspondent for The Times, covering President Trump and his administration.

A version of this article appears in print on , Section B, Page 4 of the New York edition with the headline: Trump Team and Anthropic Try to Reach a Compromise