ADDENDUM

ADDENDUM
TABLE OF CONTENTS

Page

ADDENDUM A

Certificate as to Parties, Rulings, and Related Cases

CERTIFICATE AS TO PARTIES, RULINGS, AND RELATED CASES

Pursuant to Circuit Rule 28(a)(1), Anthropic provides the following:

A. Parties, Intervenors, and Amicus Curiae

Petitioner

The petitioner is Anthropic PBC.

Respondents

Respondents are the United States Department of War and Department of War Secretary Peter B. Hegseth in his official capacity.

Participants Below

Not applicable.

B. Rulings Under Review

Petitioners challenge covered procurement actions carried out by United States Department of War Secretary Peter B. Hegseth.

C. Related Cases

This case has not previously been before this Court, any other United States court of appeals, or any other court in the District of Columbia. This case challenges the designation of Anthropic as a supply chain risk under a different statutory authority from the designation being challenged in *Anthropic BPC v. U.S. Department of War*, No. 3:26-cv-1996 (N.D. Cal.).

ADDENDUM B

Statutory Provisions Relied Upon

Federal Acquisition Supply Chain Security Act of 2018,
132 Stat. 5178

***

41 U.S.C. § 1323. Functions and authorities

(a) In general.— The [Federal Acquisition Security] Council shall perform functions that include the following:

(1) Identifying and recommending development by the National Institute of Standards and Technology of supply chain risk management standards, guidelines, and practices for executive agencies to use when assessing and developing mitigation strategies to address supply chain risks, particularly in the acquisition and use of covered articles under section 1326(a) of this title.

(2) Identifying or developing criteria for sharing information with executive agencies, other Federal entities, and non-Federal entities with respect to supply chain risk, including information related to the exercise of authorities provided under this section and sections 1326 and 4713 of this title. At a minimum, such criteria shall address—

(A) the content to be shared;

(B) the circumstances under which sharing is mandated or voluntary; and

(C) the circumstances under which it is appropriate for an executive agency to rely on information made available through such sharing in exercising the responsibilities and authorities provided under this section and section 4713 of this title.

(3) Identifying an appropriate executive agency to—

(A) accept information submitted by executive agencies based on the criteria established under paragraph (2);

(B) facilitate the sharing of information received under subparagraph (A) to support supply chain risk analyses under section 1326 of this title, recommendations under this section, and covered procurement actions under section 4713 of this title;

(C) share with the Council information regarding covered procurement actions by executive agencies taken under section 4713 of this title; and

(D) inform the Council of orders issued under this section.

(4) Identifying, as appropriate, executive agencies to provide—

(A) shared services, such as support for making risk assessments, validation of products that may be suitable for acquisition, and mitigation activities; and

(B) common contract solutions to support supply chain risk management activities, such as subscription services or machine-learning-enhanced analysis applications to support informed decision making.

(5) Identifying and issuing guidance on additional steps that may be necessary to address supply chain risks arising in the course of executive agencies providing shared services, common contract solutions, acquisitions vehicles, or assisted acquisitions.

(6) Engaging with the private sector and other nongovernmental stakeholders in performing the functions described in paragraphs (1) and (2) and on issues relating to the management of supply chain risks posed by the acquisition of covered articles.

(7) Carrying out such other actions, as determined by the Council, that are necessary to reduce the supply chain risks posed by acquisitions and use of covered articles.

(b) Program office and committees.—The Council may establish a program office and any committees, working groups, or other constituent bodies the Council deems appropriate, in its sole and unreviewable discretion, to carry out its functions.

(c) Authority for exclusion or removal orders.—

(1) Criteria.—To reduce supply chain risk, the Council shall establish criteria and procedures for—

(A) recommending orders applicable to executive agencies requiring the exclusion of sources or covered articles from executive agency procurement actions (in this section referred to as "exclusion orders");

(B) recommending orders applicable to executive agencies requiring the removal of covered articles from executive agency information systems (in this section referred to as "removal orders");

(C) requesting and approving exceptions to an issued exclusion or removal order when warranted by circumstances, including alternative mitigation actions or other findings relating to the national interest, including national security reviews, national security investigations, or national security agreements; and

(D) ensuring that recommended orders do not conflict with standards and guidelines issued under section 11331 of title 40 and that the Council consults with the Director of the National Institute of Standards and Technology regarding any recommended orders that would implement standards and guidelines developed by the National Institute of Standards and Technology.

(2) Recommendations.—The Council shall use the criteria established under paragraph (1), information made available under subsection (a)(3), and any other information the Council determines appropriate to issue recommendations, for application to executive agencies or any subset thereof, regarding the exclusion of sources or covered articles from any executive agency procurement action, including source selection and consent for a contractor to subcontract, or the removal of covered articles from executive agency information systems. Such recommendations shall include—

(A) information necessary to positively identify the sources or covered articles recommended for exclusion or removal;

(B) information regarding the scope and applicability of the recommended exclusion or removal order;

(C) a summary of any risk assessment reviewed or conducted in support of the recommended exclusion or removal order;

(D) a summary of the basis for the recommendation, including a discussion of less intrusive measures that were considered and why such measures were not reasonably available to reduce supply chain risk;

(E) a description of the actions necessary to implement the recommended exclusion or removal order; and

(F) where practicable, in the Council's sole and unreviewable discretion, a description of mitigation steps that could be taken by the source that may result in the Council rescinding a recommendation.

(3) Notice of recommendation and review.—A notice of the Council's recommendation under paragraph (2) shall be issued to any source named in the recommendation advising—

(A) that a recommendation has been made;

(B) of the criteria the Council relied upon under paragraph (1) and, to the extent consistent with national security and law enforcement interests, of information that forms the basis for the recommendation;

(C) that, within 30 days after receipt of notice, the source may submit information and argument in opposition to the recommendation;

(D) of the procedures governing the review and possible issuance of an exclusion or removal order pursuant to paragraph (5); and

(E) where practicable, in the Council's sole and unreviewable discretion, a description of mitigation steps that could be taken by the source that may result in the Council rescinding the recommendation.

(4) Confidentiality.—Any notice issued to a source under paragraph (3) shall be kept confidential until—

(A) an exclusion or removal order is issued pursuant to paragraph (5); and

(B) the source has been notified pursuant to paragraph (6).

(5) Exclusion and removal orders.—

(A) Order issuance.—Recommendations of the Council under paragraph (2), together with any information submitted by a source under paragraph (3) related to such a recommendation, shall be reviewed by the following officials, who may issue exclusion and removal orders based upon such recommendations:

> **(i)** The Secretary of Homeland Security, for exclusion and removal orders applicable to civilian agencies, to the extent not covered by clause (ii) or (iii).
>
> **(ii)** The Secretary of Defense, for exclusion and removal orders applicable to the Department of Defense and national security systems other than sensitive compartmented information systems.
>
> (iii) The Director of National Intelligence, for exclusion and removal orders applicable to the intelligence community and sensitive compartmented information systems, to the extent not covered by clause (ii).

(B) Delegation.—The officials identified in subparagraph (A) may not delegate any authority under this subparagraph to an official below the level one level below the Deputy Secretary or Principal Deputy Director, except that the Secretary of Defense may delegate authority for removal orders to the Commander of the United States Cyber Command, who may not redelegate such authority to an official below the level one level below the Deputy Commander.

(C) Facilitation of exclusion orders.—If officials identified under this paragraph from the Department of Homeland Security, the Department of Defense, and the Office of the Director of National Intelligence issue orders collectively resulting in a governmentwide exclusion, the Administrator for General Services and officials at other executive agencies responsible for management of the Federal Supply Schedules, governmentwide acquisition contracts and multi-agency contracts shall help facilitate implementation of such orders by removing the covered articles or sources identified in the orders from such contracts.

(D) Review of exclusion and removal orders.—The officials identified under this paragraph shall review all exclusion and removal

orders issued under subparagraph (A) not less frequently than annually pursuant to procedures established by the Council.

(E) Rescission.—Orders issued pursuant to subparagraph (A) may be rescinded by an authorized official from the relevant issuing agency.

(6) Notifications.—Upon issuance of an exclusion or removal order pursuant to paragraph (5)(A), the official identified under that paragraph who issued the order shall—

(A) notify any source named in the order of—

(i) the exclusion or removal order; and

(ii) to the extent consistent with national security and law enforcement interests, information that forms the basis for the order;

(B) provide classified or unclassified notice of the exclusion or removal order to the appropriate congressional committees and leadership; and

(C) provide the exclusion or removal order to the agency identified in subsection (a)(3).

(7) Compliance.—Executive agencies shall comply with exclusion and removal orders issued pursuant to paragraph (5).

(d) Authority to request information.—The Council may request such information from executive agencies as is necessary for the Council to carry out its functions.

(e) Relationship to other councils.—The Council shall consult and coordinate, as appropriate, with other relevant councils and interagency committees, including the Chief Information Officers Council, the Chief Acquisition Officers Council, the Federal Acquisition Regulatory Council, and the Committee on Foreign Investment in the United States, with respect to supply chain risks posed by the acquisition and use of covered articles.

(f) Rules of construction.—Nothing in this section shall be construed—

(1) to limit the authority of the Office of Federal Procurement Policy to carry out the responsibilities of that Office under any other provision of law; or

(2) to authorize the issuance of an exclusion or removal order based solely on the fact of foreign ownership of a potential procurement source that is otherwise qualified to enter into procurement contracts with the Federal Government.

***

41 U.S.C. § 1327. Judicial review procedures

(a) In general.—Except as provided in subsection (b) and chapter 71 of this title, and notwithstanding any other provision of law, an action taken under section 1323 or 4713 of this title, or any action taken by an executive agency to implement such an action, shall not be subject to administrative review or judicial review, including bid protests before the Government Accountability Office or in any Federal court.

(b) Petitions.—

(1) In general.—Not later than 60 days after a party is notified of an exclusion or removal order under section 1323(c)(6) of this title or a covered procurement action under section 4713 of this title, the party may file a petition for judicial review in the United States Court of Appeals for the District of Columbia Circuit claiming that the issuance of the exclusion or removal order or covered procurement action is unlawful.

(2) Standard of review.—The Court shall hold unlawful a covered action taken under sections 1323 or 4713 of this title, in response to a petition that the court finds to be—

(A) arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law;

(B) contrary to constitutional right, power, privilege, or immunity;

(C) in excess of statutory jurisdiction, authority, or limitation, or short of statutory right;

(D) lacking substantial support in the administrative record taken as a whole or in classified information submitted to the court under paragraph (3); or

(E) not in accord with procedures required by law.

(3) Exclusive jurisdiction.—The United States Court of Appeals for the District of Columbia Circuit shall have exclusive jurisdiction over claims arising under sections 1323(c)(5) or 4713 of this title against the United States, any United States department or agency, or any component or official of any such department or agency, subject to review by the Supreme Court of the United States under section 1254 of title 28.

(4) Administrative record and procedures.—

(A) In general.—The procedures described in this paragraph shall apply to the review of a petition under this section.

(B) Administrative record.—

(i) Filing of record.—The United States shall file with the court an administrative record, which shall consist of the information that the appropriate official relied upon in issuing an exclusion or removal order under section 1323(c)(5) or a covered procurement action under section 4713 of this title.

(ii) Unclassified, nonprivileged information.—All unclassified information contained in the administrative record that is not otherwise privileged or subject to statutory protections shall be provided to the petitioner with appropriate protections for any privileged or confidential trade secrets and commercial or financial information.

(iii) In camera and ex parte.—The following information may be included in the administrative record and shall be submitted only to the court ex parte and in camera:

(I) Classified information.

(II) Sensitive security information, as defined by section 1520.5 of title 49, Code of Federal Regulations.

(III) Privileged law enforcement information.

(IV) Information obtained or derived from any activity authorized under the Foreign Intelligence Surveillance Act of 1978 (50 U.S.C. 1801 et seq.), except that, with respect to such information, subsections (c), (e), (f), (g), and (h) of section 106 (50 U.S.C. 1806), subsections (d), (f), (g), (h), and (i) of section 305 (50 U.S.C. 1825), subsections (c), (e), (f), (g), and (h) of section 405 (50 U.S.C. 1845), and section 706 (50 U.S.C. 1881e) of that Act shall not apply.

(V) Information subject to privilege or protections under any other provision of law.

(iv) Under seal.—Any information that is part of the administrative record filed ex parte and in camera under clause (iii), or cited by the court in any decision, shall be treated by the court consistent with the provisions of this subparagraph and shall remain under seal and preserved in the records of the court to be made available consistent with the above provisions in the event of further proceedings. In no event shall such information be released to the petitioner or as part of the public record.

(v) Return.—After the expiration of the time to seek further review, or the conclusion of further proceedings, the court shall return the administrative record, including any and all copies, to the United States.

(C) Exclusive remedy.—A determination by the court under this subsection shall be the exclusive judicial remedy for any claim described in this section against the United States, any United States department or agency, or any component or official of any such department or agency.

(D) Rule of construction.—Nothing in this section shall be construed as limiting, superseding, or preventing the invocation of, any privileges or defenses that are otherwise available at law or in equity to protect against the disclosure of information.

(c) Definition.—In this section, the term "classified information"—

(1) has the meaning given that term in section 1(a) of the Classified Information Procedures Act (18 U.S.C. App.); and

(2) includes—

(A) any information or material that has been determined by the United States Government pursuant to an Executive order, statute, or regulation to require protection against unauthorized disclosure for reasons of national security; and

(B) any restricted data, as defined in section 11 of the Atomic Energy Act of 1954 (42 U.S.C. 2014).

***

41 U.S.C. § 4713. Authorities relating to mitigating supply chain risks in the procurement of covered articles

(a) Authority.—Subject to subsection (b), the head of an executive agency may carry out a covered procurement action.

(b) Determination and notification.—Except as authorized by subsection (c) to address an urgent national security interest, the head of an executive agency may exercise the authority provided in subsection (a) only after—

(1) obtaining a joint recommendation, in unclassified or classified form, from the chief acquisition officer and the chief information officer of the agency, or officials performing similar functions in the case of executive agencies that do not have such officials, which includes a review of any risk assessment made available by the executive agency identified under section 1323(a)(3) of this title, that there is a significant supply chain risk in a covered procurement;

(2) providing notice of the joint recommendation described in paragraph (1) to any source named in the joint recommendation advising—

(A) that a recommendation is being considered or has been obtained;

(B) to the extent consistent with the national security and law enforcement interests, of information that forms the basis for the recommendation;

(C) that, within 30 days after receipt of the notice, the source may submit information and argument in opposition to the recommendation; and

(D) of the procedures governing the consideration of the submission and the possible exercise of the authority provided in subsection (a);

(3) making a determination in writing, in unclassified or classified form, after considering any information submitted by a source under paragraph (2) and in consultation with the chief information security officer of the agency, that—

(A) use of the authority under subsection (a) is necessary to protect national security by reducing supply chain risk;

(B) less intrusive measures are not reasonably available to reduce such supply chain risk; and

(C) the use of such authorities will apply to a single covered procurement or a class of covered procurements, and otherwise specifies the scope of the determination; and

(4) providing a classified or unclassified notice of the determination made under paragraph (3) to the appropriate congressional committees and leadership that includes—

(A) the joint recommendation described in paragraph (1);

(B) a summary of any risk assessment reviewed in support of the joint recommendation required by paragraph (1); and

(C) a summary of the basis for the determination, including a discussion of less intrusive measures that were considered and why such measures were not reasonably available to reduce supply chain risk.

(c) Procedures to address urgent national security interests.—In any case in which the head of an executive agency determines that an urgent national security interest requires the immediate exercise of the authority provided in subsection (a), the head of the agency—

(1) may, to the extent necessary to address such national security interest, and subject to the conditions in paragraph (2)—

(A) temporarily delay the notice required by subsection (b)(2);

(B) make the determination required by subsection (b)(3), regardless of whether the notice required by subsection (b)(2) has been provided or whether the notified source has submitted any information in response to such notice;

(C) temporarily delay the notice required by subsection (b)(4); and

(D) exercise the authority provided in subsection (a) in accordance with such determination within 60 calendar days after the day the determination is made; and

(2) shall take actions necessary to comply with all requirements of subsection (b) as soon as practicable after addressing the urgent national security interest, including—

(A) providing the notice required by subsection (b)(2);

(B) promptly considering any information submitted by the source in response to such notice, and making any appropriate modifications to the determination based on such information;

(C) providing the notice required by subsection (b)(4), including a description of the urgent national security interest, and any modifications to the determination made in accordance with subparagraph (B); and

(D) providing notice to the appropriate congressional committees and leadership within 7 calendar days of the covered procurement actions taken under this section.

(d) Confidentiality.—The notice required by subsection (b)(2) shall be kept confidential until a determination with respect to a covered procurement action has been made pursuant to subsection (b)(3).

(e) Delegation.—The head of an executive agency may not delegate the authority provided in subsection (a) or the responsibility identified in subsection (f) to an

official below the level one level below the Deputy Secretary or Principal Deputy Director.

(f) Annual review of determinations.—The head of an executive agency shall conduct an annual review of all determinations made by such head under subsection (b) and promptly amend any covered procurement action as appropriate.

(g) Regulations.—The Federal Acquisition Regulatory Council shall prescribe such regulations as may be necessary to carry out this section.

(h) Reports required.—Not less frequently than annually, the head of each executive agency that exercised the authority provided in subsection (a) or (c) during the preceding 12-month period shall submit to the appropriate congressional committees and leadership a report summarizing the actions taken by the agency under this section during that 12-month period.

(i) Rule of construction.—Nothing in this section shall be construed to authorize the head of an executive agency to carry out a covered procurement action based solely on the fact of foreign ownership of a potential procurement source that is otherwise qualified to enter into procurement contracts with the Federal Government.

(j) Termination.—The authority provided under subsection (a) shall terminate on December 31, 2033.

(k) Definitions.—In this section:

> **(1) Appropriate congressional committees and leadership.**—The term "appropriate congressional committees and leadership" means—
>
> > **(A)** the Committee on Homeland Security and Governmental Affairs, the Committee on the Judiciary, the Committee on Appropriations, the Committee on Armed Services, the Committee on Commerce, Science, and Transportation, the Select Committee on Intelligence, and the majority and minority leader of the Senate; and
> >
> > **(B)** the Committee on Oversight and Government Reform, the Committee on the Judiciary, the Committee on Appropriations, the Committee on Homeland Security, the Committee on Armed Services, the Committee on Energy and Commerce, the Permanent Select

Committee on Intelligence, and the Speaker and minority leader of the House of Representatives.

(2) Covered article.—The term "covered article" means—

(A) information technology, as defined in section 11101 of title 40, including cloud computing services of all types;

(B) telecommunications equipment or telecommunications service, as those terms are defined in section 3 of the Communications Act of 1934 (47 U.S.C. 153);

(c) the processing of information on a Federal or non-Federal information system, subject to the requirements of the Controlled Unclassified Information program; or

(D) hardware, systems, devices, software, or services that include embedded or incidental information technology.

(3) Covered procurement.—The term "covered procurement" means—

(A) a source selection for a covered article involving either a performance specification, as provided in subsection (a)(3)(B) of section 3306 of this title, or an evaluation factor, as provided in subsection (b)(1)(A) of such section, relating to a supply chain risk, or where supply chain risk considerations are included in the agency's determination of whether a source is a responsible source as defined in section 113 of this title;

(B) the consideration of proposals for and issuance of a task or delivery order for a covered article, as provided in section 4106(d)(3) of this title, where the task or delivery order contract includes a contract clause establishing a requirement relating to a supply chain risk;

(C) any contract action involving a contract for a covered article where the contract includes a clause establishing requirements relating to a supply chain risk; or

(D) any other procurement in a category of procurements determined appropriate by the Federal Acquisition Regulatory Council, with the advice of the Federal Acquisition Security Council.

(4) Covered procurement action.—The term "covered procurement action" means any of the following actions, if the action takes place in the course of conducting a covered procurement:

(A) The exclusion of a source that fails to meet qualification requirements established under section 3311 of this title for the purpose of reducing supply chain risk in the acquisition or use of covered articles.

(B) The exclusion of a source that fails to achieve an acceptable rating with regard to an evaluation factor providing for the consideration of supply chain risk in the evaluation of proposals for the award of a contract or the issuance of a task or delivery order.

(C) The determination that a source is not a responsible source as defined in section 113 of this title based on considerations of supply chain risk.

(D) The decision to withhold consent for a contractor to subcontract with a particular source or to direct a contractor to exclude a particular source from consideration for a subcontract under the contract.

(5) Information and communications technology.—The term "information and communications technology" means—

(A) information technology, as defined in section 11101 of title 40;

(B) information systems, as defined in section 3502 of title 44; and

(C) telecommunications equipment and telecommunications services, as those terms are defined in section 3 of the Communications Act of 1934 (47 U.S.C. 153).

(6) Supply chain risk.—The term "supply chain risk" means the risk that any person may sabotage, maliciously introduce unwanted function, extract data, or otherwise manipulate the design, integrity, manufacturing, production, distribution, installation, operation, maintenance, disposition, or

retirement of covered articles so as to surveil, deny, disrupt, or otherwise manipulate the function, use, or operation of the covered articles or information stored or transmitted on the covered articles.

(7) Executive agency.—Notwithstanding section 3101(c)(1), this section applies to the Department of Defense, the Coast Guard, and the National Aeronautics and Space Administration.

ADDENDUM C

Declaration of Jared Kaplan (Mar. 10, 2026)

No. 26-1049

UNITED STATES COURT OF APPEALS FOR THE DISTRICT OF COLUMBIA CIRCUIT

ANTHROPIC PBC,

Petitioner,

v.

U.S. DEPARTMENT OF WAR, PETER B. HEGSETH, in his official capacity as Secretary of War,

Respondents.

On Petition for Judicial Review of Department of War 41 U.S.C. § 4713 Notice

DECLARATION OF JARED KAPLAN

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

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

KELLY P. DUNBAR
JOSHUA A. GELTZER
KEVIN M. LAMB
ANNEKE DUNBAR-GRONKE
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
(202) 663-6000
kelly.dunbar@wilmerhale.com
joshua.geltzer@wilmerhale.com
kevin.lamb@wilmerhale.com
anneke.dunbar-gronke@wilmerhale.com

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, most recently Claude Sonnet 4.6 in February 2026.

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 March 10, 2026.

Jared Kaplan

Jared Kaplan
Co-Founder, Anthropic

ADDENDUM D

Declaration of Sarah Heck (Mar. 10, 2026)

No. 26-1049

UNITED STATES COURT OF APPEALS FOR THE DISTRICT OF COLUMBIA CIRCUIT

ANTHROPIC PBC,

Petitioner,

v.

U.S. DEPARTMENT OF WAR, PETER B. HEGSETH, in his official capacity as Secretary of War,

Respondents.

On Petition for Judicial Review of Department of War 41 U.S.C. § 4713 Notice

DECLARATION OF SARAH HECK

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

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

KELLY P. DUNBAR
JOSHUA A. GELTZER
KEVIN M. LAMB
ANNEKE DUNBAR-GRONKE
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
(202) 663-6000
kelly.dunbar@wilmerhale.com
joshua.geltzer@wilmerhale.com
kevin.lamb@wilmerhale.com
anneke.dunbar-gronke@wilmerhale.com

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

1. My name is Sarah Heck. 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.

Anthropic Has Maintained A Strong Relationship With The Federal Government

4. 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.

5. 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.

6. 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.

7. 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 ("DoW" or 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 senior defense and intelligence officials, in part to strengthen the

company's understanding of government needs and become a more effective partner.

8. 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 Declaration of Thiyagu Ramasamy.

9. 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

10. Over the last few months, Anthropic and DoW 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 DoW's ability to use Claude for purposes of mass surveillance of Americans and lethal autonomous warfare. During that time, I engaged with DoW 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.

11. 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.

12. 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.

13. During the meeting, Dr. Amodei reiterated that Anthropic maintains a strong partnership with national security agencies, including the DoW 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—DoW 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 DoW's operations to his knowledge, and Combatant Commanders he has spoken to have been pleased with Anthropic's partnership and models.

14. 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.

15. 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 DoW "would love to work with [Amodei]," he stated that DoW had many other vendors to choose from that have never raised these types of concerns and would never hold any veto power over DoW.

16. 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, DoW would issue a statement at 5:01 p.m. that they would designate Anthropic a supply-chain risk—preventing Anthropic from partnering with DoW, 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.

17. 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 DoW 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 DoW's adoption and use of our models to date. Nor has anyone else articulated such a use to me.

18. 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 DoW'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

19. 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 41 U.S.C. § 4713

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.

* * *

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

Executed on March 10, 2026.

Sarah Heck

Sarah Heck
Head of Policy, Anthropic

ADDENDUM E

Declaration of Thiyagu Ramasamy (Mar. 10, 2026)

No. 26-1049

UNITED STATES COURT OF APPEALS FOR THE DISTRICT OF COLUMBIA CIRCUIT

ANTHROPIC PBC,

Petitioner,

v.

U.S. DEPARTMENT OF WAR, PETER B. HEGSETH, in his official capacity as Secretary of War,

Respondents.

On Petition for Judicial Review of Department of War 41 U.S.C. § 4713 Notice

DECLARATION OF THIYAGU RAMASAMY

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

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

KELLY P. DUNBAR
JOSHUA A. GELTZER
KEVIN M. LAMB
ANNEKE DUNBAR-GRONKE
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
(202) 663-6000
kelly.dunbar@wilmerhale.com
joshua.geltzer@wilmerhale.com
kevin.lamb@wilmerhale.com
anneke.dunbar-gronke@wilmerhale.com

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, where I was a Principal Lead for Data, Analytics, and Artificial Intelligence/Machine Learning and was responsible for, among other things, the implementation of Anthropic's AI models—called Claude—for public sector customers, and the deployment of Claude in classified federal government networks. My current duties and responsibilities at Anthropic include overseeing the team of employees that sell Claude to U.S. federal, state, and local government agencies.

2. As the Head of Public Sector 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.

Anthropic's Positive Relationship With The U.S. Government

3. Anthropic is a public benefit corporation whose mission is to ensure that the transition to powerful artificial intelligence ("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 Claude. Our government customers

include the Department of War ("DoW" or the "Department") and agencies in the Intelligence Community.[1]

4. 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 ("DoE") 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.[2] Anthropic later expanded its partnership with the DoE 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.[3]

5. In November 2024, Anthropic expanded its relationship with the U.S. government via a partnership with the software company Palantir Technologies.[4]

---

[1] The Intelligence Community comprises 18 organizations, including elements of the DoW (such as the National Security Agency) and other departments and agencies (such as the Department of Energy's Office of Intelligence and Counterintelligence) as well as independent agencies (such as the Central Intelligence Agency). *See Members of the IC*, Off. of Dir. of Nat'l Intel., https://www.dni.gov/index.php/what-we-do/members-of-the-ic.

[2] *See Anthropic partners with U.S. National Labs for first 1,000 Scientist AI Jam*, Anthropic (Feb. 28, 2025), https://www.anthropic.com/news/anthropic-partners-with-u-s-national-labs-for-first-1-000-scientist-ai-jam.

[3] *See Lawrence Livermore National Laboratory expands Claude for Enterprise use to empower scientists and researchers*, Anthropic (July 9, 2025), https://www.anthropic.com/news/lawrence-livermore-national-laboratory-expands-claude-for-enterprise-to-empower-scientists-and.

[4] *See Anthropic and Palantir Partner to Bring Claude AI Models to AWS for U.S. Government Intelligence and Defense Operations*, Palentir: Investors (Nov. 7, 2024),

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.

6. We have only deepened our relationship with the DoW and the Intelligence Community since then. In June 2025, we announced a custom set of Claude models built exclusively for U.S. national security customers.[5] We developed these "Claude Gov" models based on direct feedback from our national security partners to address real-world needs. In particular, these models were fine-tuned so that they would not refuse requests that regular Claude models—those for civilian enterprise or consumer use—would refuse.[6] And as with all our other Claude models, we rigorously tested these models for safety. In July 2025,

---

https://investors.palantir.com/news-details/2024/Anthropic-and-Palantir-Partner-to-Bring-Claude-AI-Models-to-AWS-for-U.S.-Government-Intelligence-and-Defense-Operations/.

[5] *See Claude Gov models for U.S. national security customers*, Anthropic (June 6, 2025), https://www.anthropic.com/news/claude-gov-models-for-u-s-national-security-customers.

[6] *See id.*

alongside other frontier AI labs including Google, OpenAI, and xAI, Anthropic was awarded a two-year, up to $200 million agreement by the DoW's Chief Digital and Artificial Intelligence Office ("CDAO"), the primary office within the DoW responsible for integrating and optimizing AI capabilities across the DoW.[7] As part of that agreement, we expanded our commitment to work with the DoW 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 DoW's mission.[8] 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.

7. 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

[7] *See CDAO Announces Partnerships with Frontier AI Companies to Address National Security Mission Areas*, CDAO (July 14, 2025), https://www.ai.mil/latest/news-press/pr-view/article/4242822/cdao-announces-partnerships-with-frontier-ai-companies-to-address-national-secu/.
[8] *See id*.

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."[9]

8. Throughout, we have maintained our commitment to supporting the national security of the United States and its allies. I have a team of 15 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 DoW 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

[9] *See* Press Release, Gen. Servs. Admin, *GSA Strikes Another OneGov Deal with Anthropic to Offer Claude AI to all Branches of Gov for Just $1*, (Aug. 12, 2025), https://www.gsa.gov/about-us/newsroom/news-releases/gsa-strikes-onegov-deal-with-anthropic-08122025.

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.

9. As one would expect, our many partnerships with the U.S. government have involved intense security reviews and thorough vetting. Last year, the DoW'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.

10. In June 2025, GSA and the DoW 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.

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

12. 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.

The Current Negotiations With The Department of War

13. In September 2025, Anthropic began negotiations with the DoW for a deployment on the DoW's GenAI.mil AI platform. As part of those discussions, the DoW began to ask that Anthropic remove its Usage Policy as applicable to the DoW contracts and subcontracts.

14. Use of Claude is expressly subject to and conditioned upon compliance with Anthropic's Terms of Service,[10] which incorporate our Usage

[10] *Available at* https://www.anthropic.com/legal/consumer-terms (Consumer Terms of Service) and at https://www.anthropic.com/legal/commercial-terms (Commercial Terms of Service).

Policy.[11] The Usage Policy is intended to help our users stay safe and promote the responsible use of our products and services. By establishing reasonable limitations on the appropriate uses for our products and services, the Usage Policy effectively defines Anthropic's commercial offerings.

15. To be clear, Anthropic does not attempt—and has never attempted—to employ its Usage Policy to exert authority, control, or command over our customers, including the DoW and its military operations. Anthropic's Usage Policy does not give it insight into how the DoW uses Claude. That said, if Anthropic's Usage Policy restrictions do not meet the DoW's needs, the DoW is of course able to use any other AI system that better meets its requirements.

16. Over the past several months, Anthropic's leadership has engaged in extensive discussions with the DoW leadership regarding the Usage Policy. Anthropic was willing to and did alter its Usage Policy to meet the specific needs of the DoW (for example, we have made clear that the Department can engage in offensive cyber operations). Although the DoW throughout that time demonstrated willingness to negotiate—leadership for both parties held in-person meetings and exchanged redlines and emails at a regular cadence—the DoW has more recently anchored on a demand that Anthropic must remove its Usage Policy and permit

[11] Anthropic's Usage Policy is available online at https://www.anthropic.com/legal/aup.

"all lawful uses" of Claude. I understand that the DoW has been or is exerting similar pressure on other leading AI labs to agree to similar demands, as reflected in Secretary Hegseth's memorandum of January 9, 2026, directing Department leadership to incorporate "any lawful use" language into DoW contracts for AI services.[12]

17. As negotiations began to break down over the last several weeks, the DoW began threatening to designate Anthropic a "supply chain risk." Aside from the public and private threats from the DoW in the last two weeks, which contain no specifics and which I believe were delivered to increase leverage in negotiations with the company, no government customer—or commercial customer or any other person—ever informed me or, to my knowledge, anyone else at Anthropic, that they considered Anthropic or our AI models a supply chain risk or more generally a threat to safety or national security. Nor has any government or commercial customer or any other person ever informed me or, to my knowledge, anyone else at Anthropic, that there is any risk that an adversary to the United States may sabotage, maliciously introduce unwanted function, or otherwise subvert the design, integrity, manufacturing, production, distribution, installation, operation, or maintenance of a system into which Anthropic's offerings are integrated, so as to

[12] *See* Memorandum from Sec'y of War, Artificial Intelligence Strategy for the Department of War 5 (Jan. 9, 2026).

surveil, deny, disrupt, or otherwise degrade the function, use, or operation of such system.

The President's And The DoW's Orders

18. On February 27, 2026, the President and the Secretary of War converted prior threats into directives designed to harm Anthropic's business.

19. First, at 3:47 p.m. Eastern time, President Trump, on his Truth Social platform, "direct[ed] EVERY Federal Agency in the United States Government to IMMEDIATELY CEASE all use of Anthropic's technology."[13] The President also called Anthropic an "out-of-control" and "RADICAL LEFT, WOKE COMPANY" of "Leftwing nut jobs" who "have no idea what the real World is all about," and cited the company's "selfishness" and "DISASTROUS MISTAKE trying to STRONG-ARM the Department of War," and threatened that "Anthropic better get their act together" or he would "use the Full Power of the Presidency to make them comply, with major civil and criminal consequences to follow."[14]

20. Shortly thereafter, Secretary Hegseth, acting on "the President's directive," "direct[ed] the Department of War to designate Anthropic a Supply-

[13] Donald J. Trump (@realDonaldTrump), TruthSocial (Feb. 27, 2026, 3:47 PM ET), https://truthsocial.com/@realDonaldTrump/posts/116144552969293195.
[14] *Id*.

Chain Risk to National Security."[15] In his post, the Secretary denounced what he considered to be "Silicon Valley ideology," "defective altruism," and "corporate virtue-signaling," and called Anthropic's actions a "textbook case of how not to do business with the United States Government or the Pentagon."[16] He concluded: "Effective immediately, no contractor, supplier, or partner that does business with the United States military may conduct any commercial activity with Anthropic."[17] At the same time, he stated that he is requiring Anthropic to continue providing services to the DoW, for up to six months.[18]

21. I refer to these statements by the President and Secretary Hegseth as the "Government's Actions."

Other Agencies' Actions In Response To The Trump And Hegseth Directives

22. Other agencies have also begun to take action in response to the Government's Actions. On Friday, February 27, 2026, following the Government's Actions, the GSA removed Anthropic from the agency's AI platform USAi.gov, the primary centralized means for federal agencies to access and adopt AI tools, as well as from the agency's Multiple Award Schedule contracts, through which

[15] Pete Hegseth (@SecWar), X (Feb. 27, 2026, 5:14 PM ET), https://x.com/SecWar/status/2027507717469049070.
[16] *Id.*
[17] *Id.*
[18] *Id.*

Anthropic provided $1 Claude subscriptions to the executive, legislative, and judicial branches of the Government as part of its "OneGov" agreement.[19] This decision cuts 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.

23. 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.[20]

24. 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.

25. 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.

[19] Press Release, Gen. Servs. Admin, G*SA Stands with President Trump on National Security AI Directive*, (Feb. 27, 2026), https://www.gsa.gov/about-us/newsroom/news-releases/gsa-stands-with-president-trump-on-national-security-ai-directive-02272026.
[20] Scott Bessent (@SecScottBessent), X (Mar. 2, 2026, 10:57 AM ET), https://x.com/secscottbessent/status/2028499953283117283?s=46; William Pulte (@pulte), X (Mar. 2, 2026, 11:12 AM ET), https://x.com/pulte/status/2028503809299779866.

26. 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.

27. 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.

28. 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.

Irreparable Harm To Anthropic

29. The Government's Actions immediately and irreparably harm Anthropic. The designation also impugns Anthropic's integrity and reputation as a trusted partner, having a real but incalculable effect on sales to non-governmental customers. All told, Anthropic will suffer considerable financial and reputational harm.

30. Being labeled as a "supply chain risk" affects our ability to sell our products and services to our U.S. government customers and others. Even before the Government's Actions, in response to the public threats from the DoW,

Anthropic had already started receiving requests from government customers to provide new contractual terms allowing those customers to terminate if Anthropic received a supply chain risk designation. The DoW has now accelerated this process by issuing the Government's Actions and directing firms to assess their reliance on Anthropic AI models.[21]

31. Anthropic maintains a sizable and accelerating public sector business, driven largely by fast adoption rates at the DoW and within the Intelligence Community. For example, from just December 2025 to the end of January 2026, we have experienced a fourfold increase in annual recurring revenue ("ARR")[22] run rate from public sector customers. Before the year began, we projected several hundred million dollars in Public Sector ARR in 2026 and have since revised these projections upward based on the pace of growth in the first two months alone. Based on current adoption rates, we project our Public Sector business in the next five years could increase to multiple billions in ARR.

32. The Government's Actions are an existential threat to all of this. By expressly excluding Anthropic from sales directly to the DoW or through its contractors, we estimate the immediate loss of more than $150 million ARR in

[21] Dave Lawler et al., *Scoop: Pentagon takes first step toward blacklisting Anthropic*, Axios (Feb. 25, 2026), https://www.axios.com/2026/02/25/anthropic-pentagon-blacklist-claude.
[22] ARR is the predictable yearly revenue associated with subscription-based software fees under active contracts.

existing and expected DoW contracts. That includes the $200 million CDAO agreement, which the Department has now cancelled, under which we anticipated over $50 million in ARR this year. It also includes Anthropic's substantial sales to the DoW through contractors, resellers, and systems integrators, which comprise a sizable portion of our Public Sector sales revenue.

33. Although I understand that there is no legal authority for the scope of the Government's Actions, we have seen, and expect to continue to see, the effects of these directives spread to the Intelligence Community and other U.S. federal agencies. Harm is already occurring. The dispute with the DoW has caused significant delays or pauses in six national security contracts or contract pipelines. In at least one instance, a customer at a strategic command center directed a partner to work with xAI or Google instead of Anthropic. These impacts are real and ongoing. In a sector where counterparties are risk-averse and dependent on government contracting, even the appearance of regulatory or political disfavor can be enough to cause disengagement. Defense contractors performing work under government contracts are assessing—and in many cases looking to terminate—their reliance on Anthropic, effectively eliminating an important market.[23] If that

[23] *See* Lora Kolodny, Ari Levy, Samantha Subin, *Defense tech companies are dropping Claude after Pentagon's Anthropic blacklist* (Mar. 4, 2026) (quoting the managing partner for a government and defense-focused venture capital firm as saying their portfolio companies involved in defense contracts "are very strict in their interpretation of the requirements" and

happens, we estimate Anthropic's more than half-a-billion dollar expected Public Sector ARR in 2026 to shrink substantially or disappear altogether. That loss of revenue includes the OneGov GSA agreement, which we anticipated generating close to $100 million in ARR this year. And it likely includes not only federal agencies, but state and local government customers as well, some of which also purchase through the GSA agreement, and others of which will reconsider purchasing Anthropic products after the company has been blacklisted as a purported "supply chain risk" by the federal government.

34. For similar reasons, the Government's Actions will also irreparably harm Anthropic's ability to forge new business partnerships and attract new customers in the future.

* * *

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

"have backed off their use of Claude for defense use cases and are in active processes to replace the service with another one"), *available at* https://www.cnbc.com/2026/03/04/pentagon-blacklist-anthropic-defense-tech-claude.html.

Executed on March 11, 2026.

Thiyagu Ramasamy

Thiyagu Ramasamy
Head of Public Sector, Anthropic

ADDENDUM F

Declaration of Paul Smith (Mar. 10, 2026)

No. 26-1049

UNITED STATES COURT OF APPEALS FOR THE DISTRICT OF COLUMBIA CIRCUIT

ANTHROPIC PBC,

Petitioner,

v.

U.S. DEPARTMENT OF WAR, PETER B. HEGSETH, in his official capacity as Secretary of War,

Respondents.

On Petition for Judicial Review of Department of War 41 U.S.C. § 4713 Notice

DECLARATION OF PAUL SMITH

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

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

KELLY P. DUNBAR
JOSHUA A. GELTZER
KEVIN M. LAMB
ANNEKE DUNBAR-GRONKE
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
(202) 663-6000
kelly.dunbar@wilmerhale.com
joshua.geltzer@wilmerhale.com
kevin.lamb@wilmerhale.com
anneke.dunbar-gronke@wilmerhale.com

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 March 10, 2026.

Paul Smith

Paul Smith
Chief Commercial Officer, Anthropic

ADDENDUM G

Declaration of Krishna Rao (Mar. 10, 2026)

No. 26-1049

UNITED STATES COURT OF APPEALS FOR THE DISTRICT OF COLUMBIA CIRCUIT

ANTHROPIC PBC,

Petitioner,

v.

U.S. DEPARTMENT OF WAR, PETER B. HEGSETH, in his official capacity as Secretary of War,

Respondents.

On Petition for Judicial Review of Department of War 41 U.S.C. § 4713 Notice

DECLARATION OF KRISHNA RAO

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

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

KELLY P. DUNBAR
JOSHUA A. GELTZER
KEVIN M. LAMB
ANNEKE DUNBAR-GRONKE
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
(202) 663-6000
kelly.dunbar@wilmerhale.com
joshua.geltzer@wilmerhale.com
kevin.lamb@wilmerhale.com
anneke.dunbar-gronke@wilmerhale.com

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." 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).

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.

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 March 10, 2026.

Krishna Rao

Krishna Rao
Chief Financial Officer, Anthropic

ADDENDUM H

Declaration of Kelly P. Dunbar (Mar. 10, 2026)

No. 26-1049

UNITED STATES COURT OF APPEALS FOR THE DISTRICT OF COLUMBIA CIRCUIT

ANTHROPIC PBC,

Petitioner,

v.

U.S. DEPARTMENT OF WAR, PETER B. HEGSETH, in his official capacity as Secretary of War,

Respondents.

On Petition for Judicial Review of Department of War 41 U.S.C. § 4713 Notice

DECLARATION OF KELLY P. DUNBAR IN SUPPORT OF PLAINTIFF ANTHROPIC PBC'S EMERGENCY MOTION FOR A STAY

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

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

KELLY P. DUNBAR
JOSHUA A. GELTZER
KEVIN M. LAMB
ANNEKE DUNBAR-GRONKE
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
(202) 663-6000
kelly.dunbar@wilmerhale.com
joshua.geltzer@wilmerhale.com
kevin.lamb@wilmerhale.com
anneke.dunbar-gronke@wilmerhale.com

March 11, 2026

I, Kelly P. Dunbar, 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. I represent Petitioner Anthropic PBC in this matter.

2. I am a member in good standing of the Bar of the United States Court of Appeals for the District of Columbia.

3. I have personal knowledge of and, if called as a witness, could and would competently testify to the factual matters asserted in this declaration, including based on my review of the documents, websites, and other items referenced in the declaration.

4. I submit this declaration in support of Petitioner's Emergency Motion for a Stay.

5. I certify that on March 9, 2026, Petitioner requested a stay of agency action from Respondent the Department of War pursuant to Rule 18(a) of the Federal Rules of Appellate Procedure by emailing Earl G. Matthews, General Counsel of the U.S. Department of War, at earl.g.matthews.civ@mail.mil; Colonel Anthony Fuscellaro at anthony.fuscellaro1.mil@war.mil; the Department's General Counsel's office at osd.pentagon.ogc.list.correspondence-staff@mail.mil; and the email address listed in the Department of War's March 3, 2026 § 4713 Notice—osd.mc-alex.ousd-a-s.mbx.10-usc-section3252-determinations@mail.mil.

6. I further certify that on March 9, 2026, in the same email, Petitioner notified Respondents of its intent to file this Motion if no response was received by Wednesday, March 11, at 12:00 p.m. ET.

7. Attached to this declaration as **Exhibit 1** is a true and correct copy of the Department of War's Notice to Anthropic Pursuant to Title 41 U.S.C. § 4713, dated Mar. 3, 2026.

8. Attached as **Exhibit 2** is a true and correct copy of a post by Secretary Hegseth (@SecWar), X (Feb. 27, 2026, 2:14 PM PT), https://tinyurl.com/yvtpje9b.

9. Attached as **Exhibit 3** is a true and correct copy of a Department of War memorandum attached to a CBS news story *Internal Pentagon Memo Orders Military Commanders to Remove Anthropic AI Technology from Key Systems*, CBS News (Mar. 10, 2026), https://tinyurl.com/mtyuanfd.

10. Attached as **Exhibit 4** is a true and correct copy of Dario Amodei, *The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI* (Jan. 2026), https://tinyurl.com/2cbcmfz7.

11. Attached as **Exhibit 5** is a true and correct copy of *A Statement from Dario Amodei on Anthropic's Commitment to American AI Leadership*, Anthropic (Oct. 21, 2025), https://tinyurl.com/4ncsm5w4.

12. Attached as **Exhibit 6** is a true and correct copy of *Claude Gov Models for U.S. National Security Customers*, Anthropic (Jun. 6, 2025), https://tinyurl.com/ynyc82bw.

13. Attached as **Exhibit 7** is a true and correct copy of *Statement from Dario Amodei on our Discussions with the Department of War*, Anthropic (Feb. 26, 2026), https://tinyurl.com/54pw9684.

14. Attached as **Exhibit 8** is a true and correct copy of a post by President Donald J. Trump (@realDonaldTrump), TruthSocial (Feb. 27, 2026, 12:47 PM PT), https://tinyurl.com/5n7ucwpw.

15. Attached as **Exhibit 9** is a true and correct copy of Joseph Menn, *Scrutiny Mounts over Tech Investments from Kremlin-connected Expatriates*, Wash. Post (Dec. 19, 2022), https://tinyurl.com/2jhz63vr.

16. Attached as **Exhibit 10** is a true and correct copy of Dave Lawler, Maria Curi & Mike Allen, *Exclusive: Pentagon Threatens Anthropic Punishment*, Axios (Feb. 16, 2026), https://tinyurl.com/2ayu3dkx.

17. Attached as **Exhibit 11** is a true and correct copy of Cade Metz, *A Guide to the Pentagon's Dance with Anthropic and OpenAI*, N.Y. Times (Mar. 7, 2026), https://tinyurl.com/4tv6zcah.

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

Executed on March 11, 2026.

/s/ *Kelly P. Dunbar*
Kelly P. Dunbar

EXHIBIT 1

SECRETARY OF WAR
1000 DEFENSE PENTAGON
WASHINGTON, DC 20301-1000

MAR - 3 2026

Mr. Dario Amodei
Chief Executive Officer
Anthropic, PBC
548 Market Street
San Francisco, CA 94104

Dear Anthropic, PBC Executive Leadership:

This letter provides notice to Anthropic, Public Benefit Corporation (PBC), and its subordinate, subsidiaries, or affiliated offices or entities, doing business under various names, and all subsidiaries, successors, or assigns thereof ("Covered Entity") that, pursuant to title 41, United States Code (U.S.C.), section 4713 ("Section 4713"), the Department of War (DoW) has determined that (i) the use of the Covered Entity's products or services in DoW covered procurements[1] presents a supply chain risk and that the use of the Section 4713 authority to carry out covered procurement actions[2] is necessary to protect national security by reducing supply chain risk, and (ii) less intrusive measures are not reasonably available to reduce such supply chain risk.

Scope of Authorized Covered Procurement Actions

This Determination is necessary to reduce supply chain risk and applies to the Covered Entity, Covered Products or Services, Covered Procurements, and Covered Procurement Actions as follows:

- **Covered Entity:** Anthropic, PBC, and its subordinate, subsidiaries, or affiliated offices or entities, doing business under various names, and all subsidiaries, successors, or assigns thereof.

- **Covered Products or Services:** All of the Covered Entity's products or services that meet the definition of a "covered article" or that are part of a "covered procurement," as those terms are defined at 41 U.S.C. § 4713(k), whether acquired as a product or service. This includes all of the Covered Entity's products or services offered by the Covered Entity that become available for procurement.

- **Covered Procurements:** All DoW procurements described in Section 4713(k)(3).

- **Covered Procurement Actions:** All actions described in Section 4713(k)(4).

[1] 41 U.S.C. § 4713(k)(3)
[2] 41 U.S.C. § 4713(k)(4)

Effective Date

This Determination is effective immediately and shall remain in effect until modified or terminated in writing by the Section 4713 Authorized Official.

Request for Reconsideration

If the Covered Entity wishes to request that the DoW reconsider this Determination, the Covered Entity must submit in writing to the undersigned within 30 days of receipt of this letter, notice of such request for reconsideration. For additional information, requirements, and procedures governing such request, see the enclosed Requirements and Procedures for Requesting Reconsideration of a Section 4713 Determination.

Sincerely,

Enclosure:
As stated

ATTACHMENT 1
Requirements and Procedures for Requesting Reconsideration of a Section 4713 Determination

The following requirements and procedures govern a Covered Entity's request for reconsideration of a Section 4713 determination:

1. **Notice of Request for Reconsideration:** Within thirty (30) days of receiving the Section 4713 Authorized Official's letter notifying the Covered Entity of the Section 4713 determination, the Covered Entity may submit in writing to the Section 4713 Authorized Official notice of the Covered Entity's request for reconsideration. Such notice should identify the specific relief or remedy being requested (e.g., specific modifications to, or termination in whole or in part, of any elements of the determination). All notices and written information must be submitted to osd.mc-alex.ousd-a-s.mbx.10-usc-section-3252-determinations@mail.mil.

2. **Opportunity to Submit and Present Information**: If the Covered Entity submits a timely notice of request for consideration, the Covered Entity will be afforded an additional thirty (30) calendar days (from the date the Section 4713 Authorized Official received such timely notice) to submit in writing, and to appear and present, additional information and arguments in support of such request for reconsideration. The Covered Entity must either send, or make arrangements to appear and present, the information to representatives of the Section 4713 Authorized Official within the thirty (30) day period. The Section 4713 Authorized Official may extend the time to appear and submit documentary evidence upon written request by the Covered Entity.

3. **Flexible Procedures**: The Section 4713 Authorized Official may use flexible procedures to allow the Covered Entity to submit and present information in support of the request for reconsideration. In so doing, the Section 4713 Authorized Official is not required to follow formal rules of evidence or procedures in creating an official record of the request for reconsideration and the Official's disposition of that and request.

4. **Content of Submissions/Presentations**: When submitting and presenting information in support of a request for reconsideration, the Covered Entity should, to the maximum extent practicable, identify specific facts that contradict statements contained in the Section 4713 Covered Entity notification, and provide detailed rationale for any arguments in support of the request and the remedy or relief being requested. A general denial is insufficient to support reconsideration of a Section 4713 determination.

5. **Appearing and Presenting Information:** An appearance and presentation is an informal meeting that is non-adversarial in nature. When electing to appear and present information, the representative(s) of the Covered Entity may choose to appear with counsel. Any information to be presented should be provided in written form at least 5 working days in advance of the presentation. Usually, all matters in opposition should be presented in a single proceeding. Any information not submitted in advance, but provided orally during an appearance, must also be submitted in writing after the appearance for the information to be

considered. The representative(s) of the Section 4713 Authorized Official, and/or other agency representatives, may ask questions of the Covered Entity's representative(s) making the presentation. Federal rules of evidence do not govern the appearance and presentation.

6. **Notice Regarding False Statements**: Any material information submitted in response to this action will be considered a statement or representation to a government official concerning a matter within the jurisdiction of the executive branch of the government. To that end, please note that an individual making any materially false, fictitious, or fraudulent statement or representation to a Government official may be subject to prosecution under 18 U.S.C. § 1001.

EXHIBIT 2

Post

Secretary of War Pete Hegseth
@SecWar

This week, Anthropic delivered a master class in arrogance and betrayal as well as a textbook case of how not to do business with the United States Government or the Pentagon.

Our position has never wavered and will never waver: the Department of War must have full, unrestricted access to Anthropic's models for every LAWFUL purpose in defense of the Republic.

Instead, @AnthropicAI and its CEO @DarioAmodei, have chosen duplicity. Cloaked in the sanctimonious rhetoric of "effective altruism," they have 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.

The Terms of Service of Anthropic's defective altruism will never outweigh the safety, the readiness, or the lives of American troops on the battlefield.

Their true objective is unmistakable: to seize veto power over the operational decisions of the United States military. That is unacceptable.

As President Trump stated on Truth Social, the Commander-in-Chief and the American people alone will determine the destiny of our armed forces, not unelected tech executives.

Anthropic's stance is fundamentally incompatible with American principles. Their relationship with the United States Armed Forces and the Federal Government has therefore been permanently altered.

In conjunction with the President's directive for the Federal Government to cease all use of Anthropic's technology, I am directing the Department of War to designate Anthropic a Supply-Chain Risk to National Security. Effective immediately, no contractor, supplier, or partner that does business with the United States military may conduct any commercial activity with Anthropic. Anthropic will continue to provide the Department of War its services for a period of no more than six months to allow for a seamless transition to a better and more patriotic service.

America's warfighters will never be held hostage by the ideological whims of Big Tech. This decision is final.

2:14 PM · Feb 27, 2026 · **12.8M** Views

EXHIBIT 3

DEPARTMENT OF WAR
6000 Defense Pentagon
Washington, D.C. 20301-6000

CHIEF INFORMATION OFFICER

MAR 0 6 2026

MEMORANDUM FOR SENIOR PENTAGON LEADERSHIP
COMMANDERS OF THE COMBATANT COMMANDS
DEFENSE AGENCY AND DOD FIELD ACTIVITY DIRECTORS

Subject: Removal of Anthropic, PBC Products in DoW Systems

References: (a) 10 United States Code (U.S.C.) § 3252
(b) Defense Federal Acquisition Regulations (DFAR) Subpart 239.73
(c) DoDI 3020.45, Mission Assurance Construct, August 14, 2018, as amended
(d) DoDI 3741.0, National Leadership Command Capabilities (NLCC) Configuration Management (CM), May 1, 2013, as amended
(e) DoDD 3020.26, DoD Continuity Programs, June 4, 2024
(f) National Defense Authorization Act for 2018, Section 1659, Evaluation and Enhanced Security of Supply Chain for Nuclear Command, Control, and Communications and Continuity of Government Programs
(g) Executive Order 13873, Securing the Information and Communications Technology and Services Supply Chain
(h) DoDI 8500.0, Cybersecurity, March 14, 2014, as amended
(i) DoDD 5144.02, DoD Chief Information Officer (DoD CIO), November 21, 2014
(j) DoDM 8530.0, Cybersecurity Activities Support Procedures, May 31, 2023
(k) Joint Capability Integration and Development System (JCIDS) Cyber Survivability Endorsement Implementation Guide, V3.0, July 2022
(l) 44 U.S.C. § 3554, Federal agency responsibilities
(m) 40 U.S.C. § 11331, Responsibility for acquisitions of information technology
(n) 41 U.S.C. § 4713, Enhancement of Contractor Protection from Reprisal for Disclosure of Certain Information

Information and communications technology (ICT) is essential to the daily operations and functionality of the Department of War (DoW) and facilitates DoW's ability to conduct business and execute its mission. Existing and emerging threats to weapon and information systems are often directly linked to the ICT supply chain. Adversaries can exploit vulnerabilities tied to hardware, software, and managed services from primary and third-party vendors, suppliers, and service providers. Successful exploitation of ICT supply chain vulnerabilities can lead to asymmetric adversary advantages, significant detriment to DoW critical systems (references a through d), and potentially catastrophic risks to the warfighter (references g and h). The DoW Chief Information Officer (CIO), in compliance with the authorities granted by references (g) through (n), has determined that the use of Anthropic, PBC, and its subordinate, subsidiaries, or affiliated offices or entities, doing business under various names, and all

subsidiaries, successors, or assigns thereof (Covered Company) products presents an unacceptable supply chain risk for use in all DoW systems and networks.

DoW Components will discontinue all use of the Covered Company's products across all DoW systems within 180 days. This action must be prioritized for systems supporting critical missions, including but not limited to:

- National security systems (NSS), strategic priorities, and nuclear weapons
- Nuclear command, control, and communications
- Continuity of government
- Ballistic missile defense (references a-f)
- Warfighting capabilities at high-risk cyber survivability levels (Categories 3-5), per reference (k)

DoW Components will remove the Covered Company's products from all DoW systems and networks, including government-furnished end-user devices such as desktops, laptops, and mobile devices, as soon as practical or through leveraging established technical refresh cycles. However, all products covered by this memorandum must be removed within 180 days from the date of this memorandum. To ensure continuity of operations during the migration, Components are authorized to procure temporary licenses and support tokens as required. All such procurements must be tied to an approved transition plan and are limited to the 180-day period defined in this memorandum.

Furthermore, Components must remove the Covered Company's products from all Component Approved Products Lists, Service Provider Equipment Lists, e-Commerce markets, and other similar functions that facilitate the procurement of IT hardware, software, and services, as these products are no longer authorized for installation in DoW systems and networks.

This memorandum directs the phased removal of all products and services from the Covered Company from the DoW enterprise. All DoW Components and Defense Industrial Base (DIB) partners must achieve full compliance within 180 days of this memorandum's date. As directed by the Under Secretary of War for Acquisition and Sustainment (USD(A&S)), this prohibition applies to all DIB contracts. Accordingly, DoW Components shall incorporate this restriction into all current and future contracts. Contracting officers shall notify contractors of this requirement within 30 days, and all DIB entities must represent their full compliance in writing to their contracting officer no later than the 180-day deadline.

.

While this memorandum prohibits waivers, the DoW CIO is the sole authority for granting temporary exemptions in rare and extraordinary circumstances. Exemptions will only be considered for mission-critical activities directly supporting national security operations where no viable alternative exists, and the requesting Component must submit a comprehensive risk mitigation plan for approval.

Kirsten A. Davies

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

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

Contents

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*

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."

Contents

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

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

Contents

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.

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

"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

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

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.

Contents

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

Contents

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

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

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

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.

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,

Contents

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

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

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

—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. ↩

[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

Contents

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

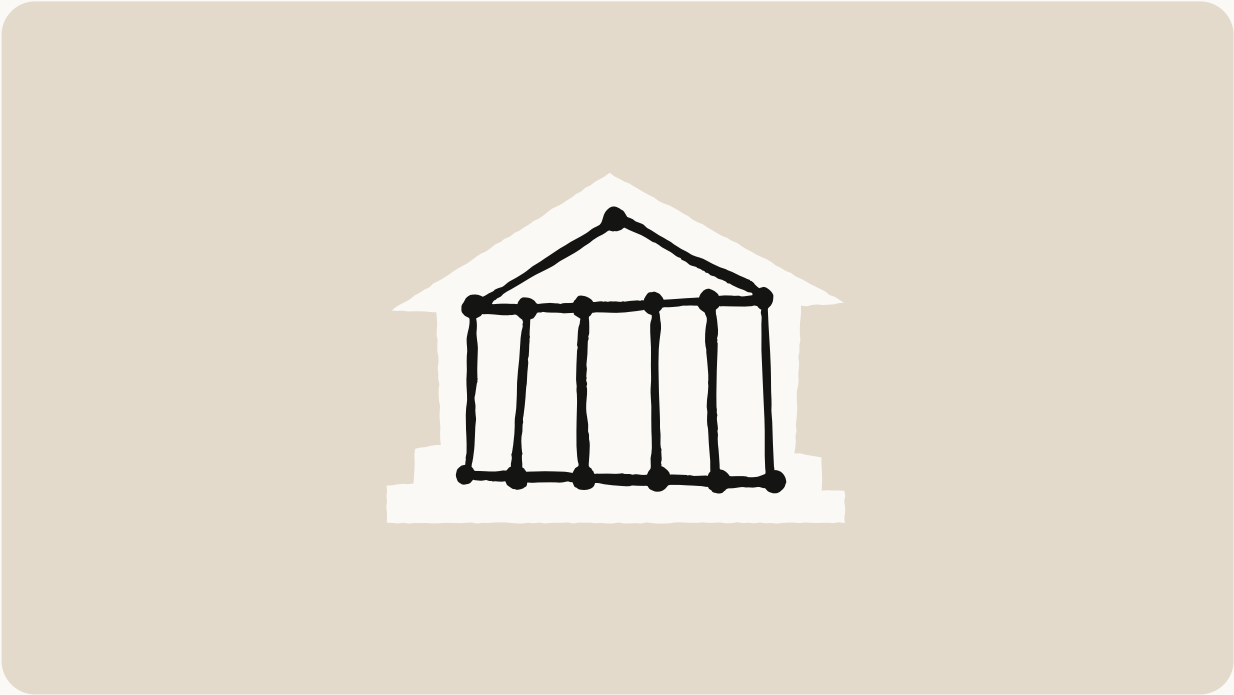
Contents

EXHIBIT 5

Announcements

A statement from Dario Amodei on Anthropic's commitment to American AI leadership

Oct 21, 2025



A statement from Anthropic CEO Dario Amodei on Anthropic's commitment to advancing America's leadership in building powerful and beneficial AI.

Anthropic is built on a simple principle: AI should be a force for human progress, not peril. That means making products that are genuinely useful, speaking honestly about risks and benefits, and working with anyone serious about getting this right. I strongly agree with Vice President JD Vance's recent comments on AI—particularly his point that we need to maximize applications that help people, like breakthroughs in medicine and disease prevention, while minimizing the harmful ones. This position is both wise and what the public overwhelmingly wants.

Anthropic is the fastest-growing software company in history, with revenue growing from a $1B to $7B run rate over the last nine months, and we've managed to do this while deploying AI thoughtfully and responsibly. There are products we will not build and risks we will not take, even if they would make money.

Our longstanding position is that managing the societal impacts of AI should be a matter of policy over politics. I fully believe that Anthropic, the administration, and leaders across the political spectrum want the same thing: to ensure that powerful AI technology benefits the American people and that America advances and secures its lead in AI development.

Despite our track record of communicating frequently and transparently about our positions, there has been a recent uptick in inaccurate claims about Anthropic's policy stances. Some are significant enough that they warrant setting the record straight.

Our alignment with the Trump administration on key areas of AI policy

- We work directly with the federal government in several ways. In July the Department of War awarded Anthropic a two-year, $200 million agreement to prototype frontier AI capabilities that advance national security. We have partnered with the General Services Administration to offer Claude for Enterprise and Claude for Government for $1 across the federal government. And Claude is deployed across classified networks through partners like Palantir and at Lawrence Livermore National Laboratory.
- Anthropic publicly praised President Trump's AI Action Plan. We have been supportive of the President's efforts to expand energy provision in the US in order to win the AI race, and I personally attended an AI and energy summit in Pennsylvania with President Trump, where he and I had a good conversation about US leadership in AI. Anthropic's Chief Product Officer attended a White House event where we joined a pledge to accelerate healthcare applications of AI, and our Head of External Affairs attended the White House's AI Education Taskforce event to support their efforts to advance AI fluency for teachers.
- Every major AI company has hired policy experts from both parties and recent administrations—Anthropic is no different. We've hired Republicans and Democrats alike, and built an advisory council that includes senior former Trump administration officials. Anthropic makes hiring decisions based on candidates' expertise, integrity, and competence, not their political affiliations.
- We (and many other organizations) respectfully disagreed with a single proposed amendment in the One Big Beautiful Bill: the 10-year moratorium on state-level AI laws, which would have blocked any action without offering a federal alternative. That specific provision was voted down by Republicans and Democrats in a 99-1 vote in the Senate. Our longstanding position has been that a uniform federal approach is preferable to a patchwork of state laws. I proposed such a standard months ago and we're ready to work with both parties to make it happen.

Our preference for a national AI standard

- While we continue to advocate for that federal standard, AI is moving so fast that we can't wait for Congress to act. We therefore supported a carefully designed bill in California where most of America's leading AI labs are headquartered, including Anthropic. This bill, SB 53, requires the largest AI developers to make their frontier model safety protocols public and is written to exempt any company with an annual gross revenue below $500M—therefore only applying to the very largest AI companies. Anthropic supported this exemption to protect startups and in fact proposed an early version of it.
- Some have suggested that we are somehow interested in harming the startup ecosystem. Startups are among our most important customers. We work with tens of thousands of startups and partner with hundreds of accelerators and

VCs. Claude is powering an entirely new generation of AI-native companies. Damaging that ecosystem makes no sense for us.

- I've heard arguments that state AI regulation could slow down the US AI industry and hand China the lead. But the real risk to American AI leadership isn't a single state law that only applies to the largest companies—it's filling the PRC's data centers with US chips they can't make themselves. We agree with leaders like Senators Tom Cotton and Josh Hawley that this would only help the Chinese Communist Party win the race to the AI frontier. We are the only frontier AI company to restrict the selling of AI services to PRC-controlled companies, forgoing significant short-term revenue to prevent fueling AI platforms and applications that would help the Chinese Communist Party's military and intelligence services.

Our progress on an AI industry-wide challenge: model bias

- Some have claimed that Anthropic's models are uniquely politically biased. This is not only unfounded but directly contradicted by the data. A January study from the Manhattan Institute, a conservative think tank, found Anthropic's main model (at the time, Claude Sonnet 3.5) to be less politically biased than models from most of the other major providers. Data from a Stanford study in May, on user perceptions of bias in AI models, shows no reason to single out Anthropic: many models from other providers were rated as more biased. The system cards for our latest models, Sonnet 4.5 and Haiku 4.5, show that we're making rapid progress towards our goal of political neutrality.

- As a broader point, no AI model, from any provider, is fully politically balanced in every reply. Models learn from their training data in ways that are not yet well-understood, and developers are never fully in control of their outputs. Anyone can cherry-pick outputs from any model to make it appear slanted in a particular direction.

Anthropic is committed to constructive engagement on matters of public policy. When we agree, we say so. When we don't, we propose an alternative for consideration. We do this because we are a public benefit corporation with a mission to ensure that AI benefits everyone, and because we want to maintain America's lead in AI. Again, we believe we share those goals with the Trump administration, both sides of Congress, and the public. We are going to keep being honest and straightforward, and will stand up for the policies we believe are right. The stakes of this technology are too great for us to do otherwise.

In his recent remarks, the Vice President also said of AI, "Is it good or is it bad, or is it going to help us or going to hurt us? The answer is probably both, and we should be trying to maximize as much of the good and minimize as much of the bad." That perfectly captures our view. We're ready to work in good faith with anyone of any political stripe to make that vision a reality.

Related content

Statement on the comments from Secretary of War Pete Hegseth

Anthropic's response to the Secretary of War and advice to customers.

Read more →

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

A statement from our CEO on national security uses of AI.

Read more →

Anthropic acquires Vercept to advance Claude's computer use capabilities

Read more →

Products

Claude
Claude Code
Claude Code Enterprise
Cowork
Claude in Chrome
Claude in Excel
Claude in PowerPoint
Claude in Slack
Skills
Max plan
Team plan
Enterprise plan
Download app
Pricing
Log in to Claude

Models

Opus
Sonnet
Haiku

Solutions

AI agents
Code modernization
Coding
Customer support
Education
Financial services
Government
Healthcare
Life sciences
Nonprofits

Claude Developer Platform

Overview
Developer docs
Pricing
Regional compliance
Amazon Bedrock
Google Cloud's Vertex AI
Console login

Learn

Blog
Claude partner network
Connectors
Courses
Customer stories
Engineering at Anthropic
Events
Plugins
Powered by Claude
Service partners
Startups program
Tutorials
Use cases

Company

Anthropic
Careers
Economic Futures
Research
News
Claude's Constitution
Responsible Scaling Policy
Security and compliance
Transparency

Help and security

Availability
Status
Support center

Terms and policies

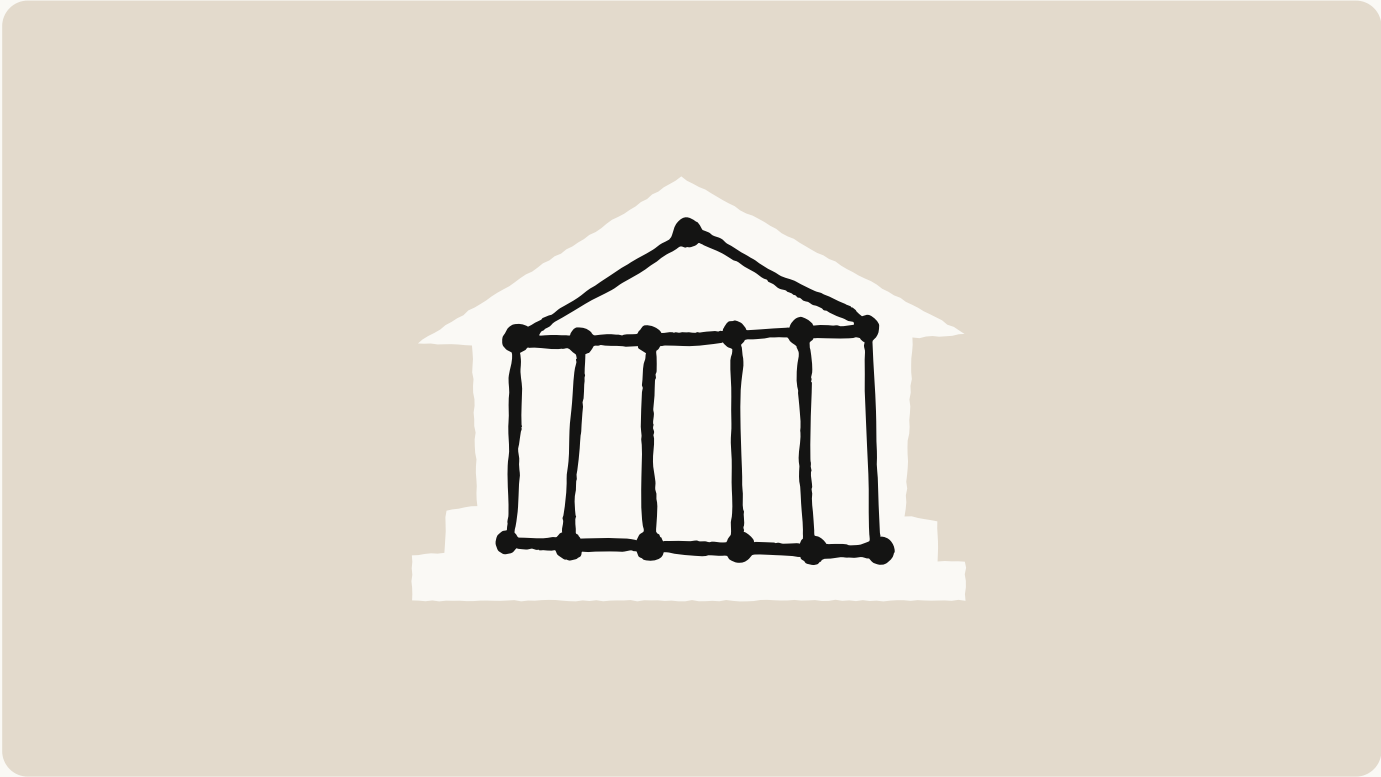
Privacy choices
Privacy policy
Consumer health data privacy policy
Responsible disclosure policy
Terms of service: Commercial
Terms of service: Consumer
Usage policy

EXHIBIT 6

Announcements

Claude Gov models for U.S. national security customers

Jun 6, 2025



We're introducing a custom set of Claude Gov models built exclusively for U.S. national security customers. The models are already deployed by agencies at the highest level of U.S. national security, and access to these models is limited to those who operate in such classified environments.

Claude Gov models were built based on direct feedback from our government customers to address real-world operational needs, and underwent the same rigorous safety testing as all of our Claude models. The result is a set of Claude models that understands our customers' unique national security requirements while maintaining Anthropic's unwavering commitment to safety and responsible AI development.

U.S. national security customers may choose to use our AI systems for a wide range of applications from strategic planning and operational support to intelligence analysis and threat assessment. Claude Gov models deliver

enhanced performance for critical government needs and specialized tasks. This includes:

- Improved handling of classified materials, as the models refuse less when engaging with classified information
- Greater understanding of documents and information within the intelligence and defense contexts
- Enhanced proficiency in languages and dialects critical to national security operations
- Improved understanding and interpretation of complex cybersecurity data for intelligence analysis

This builds on our commitment to bring responsible and safe AI solutions to our U.S. national security customers, with custom models that are built to address the unique needs of classified environments. To learn more about the Claude Gov models and see how they can support your agency's mission, you can reach our public sector team at pubsec@anthropic.com.

Related content

Statement on the comments from Secretary of War Pete Hegseth

Anthropic's response to the Secretary of War and advice to customers.

Read more →

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

A statement from our CEO on national security uses of AI.

Read more →

Anthropic acquires Vercept to advance Claude's computer use capabilities

Read more →

Products

Claude

Claude Code

Claude Code Enterprise

Cowork

Claude in Chrome

Claude in Excel

Claude in PowerPoint

Claude in Slack

Skills

Max plan

Team plan

Enterprise plan

Download app

Pricing

Log in to Claude

Models

Opus

Sonnet

Haiku

Solutions

AI agents

Code modernization

Coding

Customer support

Education

Financial services

Government

Healthcare

Life sciences

Nonprofits

Claude Developer Platform

Overview

Developer docs

Pricing

Regional compliance

Amazon Bedrock

Google Cloud's Vertex AI

Console login

Learn

Blog

Claude partner network

Connectors

Courses

Customer stories

Engineering at Anthropic

Events

Plugins

Powered by Claude

Service partners

Startups program

Tutorials

Use cases

Company

Anthropic

Careers

Economic Futures

Research

News

Claude's Constitution

Responsible Scaling Policy

Security and compliance

Transparency

Help and security

Availability

Status

Support center

Terms and policies

Privacy choices

Privacy policy

Consumer health data privacy policy

Responsible disclosure policy

Terms of service: Commercial

Terms of service: Consumer

Usage policy

EXHIBIT 7

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.

Related content

Statement on the comments from Secretary of War Pete Hegseth

Anthropic's response to the Secretary of War and advice to customers.

Read more →

Anthropic acquires Vercept to advance Claude's computer use capabilities

Read more →

Anthropic's Responsible Scaling Policy: Version 3.0

Read more →

Products

Claude

Claude Code

Claude Code Enterprise

Cowork

Claude in Chrome

Claude in Excel

Claude in PowerPoint

Claude in Slack

Skills

Max plan

Team plan

Enterprise plan

Download app

Pricing

Log in to Claude

Models

Opus

Sonnet

Haiku

Solutions

AI agents

Code modernization

Coding

Customer support

Education

Financial services

Government

Healthcare

Life sciences

Nonprofits

Claude Developer Platform

Overview

Developer docs

Pricing

Regional compliance

Amazon Bedrock

Google Cloud's Vertex AI

Console login

Learn

Blog

Claude partner network

Connectors

Courses

Customer stories

Engineering at Anthropic

Events

Plugins

Powered by Claude

Service partners

Startups program

Tutorials

Use cases

Company

Anthropic

Careers

Economic Futures

Research

News

Claude's Constitution

Responsible Scaling Policy

Security and compliance

Transparency

Help and security

Availability

Status

Support center

Terms and policies

Privacy choices

Privacy policy

Consumer health data privacy policy

Responsible disclosure policy

Terms of service: Commercial

Terms of service: Consumer

Usage policy

EXHIBIT 8

TRUTH.

← **Truth Details**
4809 replies

Donald J. Trump
@realDonaldTrump

THE UNITED STATES OF AMERICA WILL NEVER ALLOW A RADICAL LEFT, WOKE COMPANY TO DICTATE HOW OUR GREAT MILITARY FIGHTS AND WINS WARS! That decision belongs to YOUR COMMANDER-IN-CHIEF, and the tremendous leaders I appoint to run our Military.

The Leftwing nut jobs at Anthropic have made a DISASTROUS MISTAKE trying to STRONG-ARM the Department of War, and force them to obey their Terms of Service instead of our Constitution. Their selfishness is putting AMERICAN LIVES at risk, our Troops in danger, and our National Security in JEOPARDY.

Therefore, I am directing EVERY Federal Agency in the United States Government to IMMEDIATELY CEASE all use of Anthropic's technology. We don't need it, we don't want it, and will not do business with them again! There will be a Six Month phase out period for Agencies like the Department of War who are using Anthropic's products, at various levels. Anthropic better get their act together, and be helpful during this phase out period, or I will use the Full Power of the Presidency to make them comply, with major civil and criminal consequences to follow.

WE will decide the fate of our Country — NOT some out-of-control, Radical Left AI company run by people who have no idea what the real World is all about. Thank you for your attention to this matter. MAKE AMERICA GREAT AGAIN!

PRESIDENT DONALD J. TRUMP

13.7k ReTruths **56.7k** Likes Feb 27, 2026, 12:47 PM

EXHIBIT 9

U.S. intelligence probing Russian investors in U.S. tech

washingtonpost.com/technology/2022/12/19/russia-expatriates-links-probed

Joseph Menn December 19, 2022

Western intelligence officials are investigating whether a network of wealthy and well-connected expatriate Russian investors is part of a covert effort to aid their native country in developing cutting-edge technologies such as quantum computing and artificial intelligence through start-ups they funded in the United States, according to people familiar with the inquiries.

Before moving abroad and backing high-tech companies in the United States and allied nations, several of the expatriates were affiliated with one or more of three high-profile Russian tech initiatives: the government-subsidized Skolkovo technology area intended to rival Silicon Valley in suburban Moscow; the Russian Venture Co., a government investment vehicle to help Russian businesses develop innovative technology; and the nonprofit research administrator Russian Quantum Center, which operates 12 laboratories near Moscow.

Russian government money was included in venture funds managed by oneof the expats at least as recently as 2019, according to fund postings, interviews and Russian media, anissue that has drawn investigators' attention.

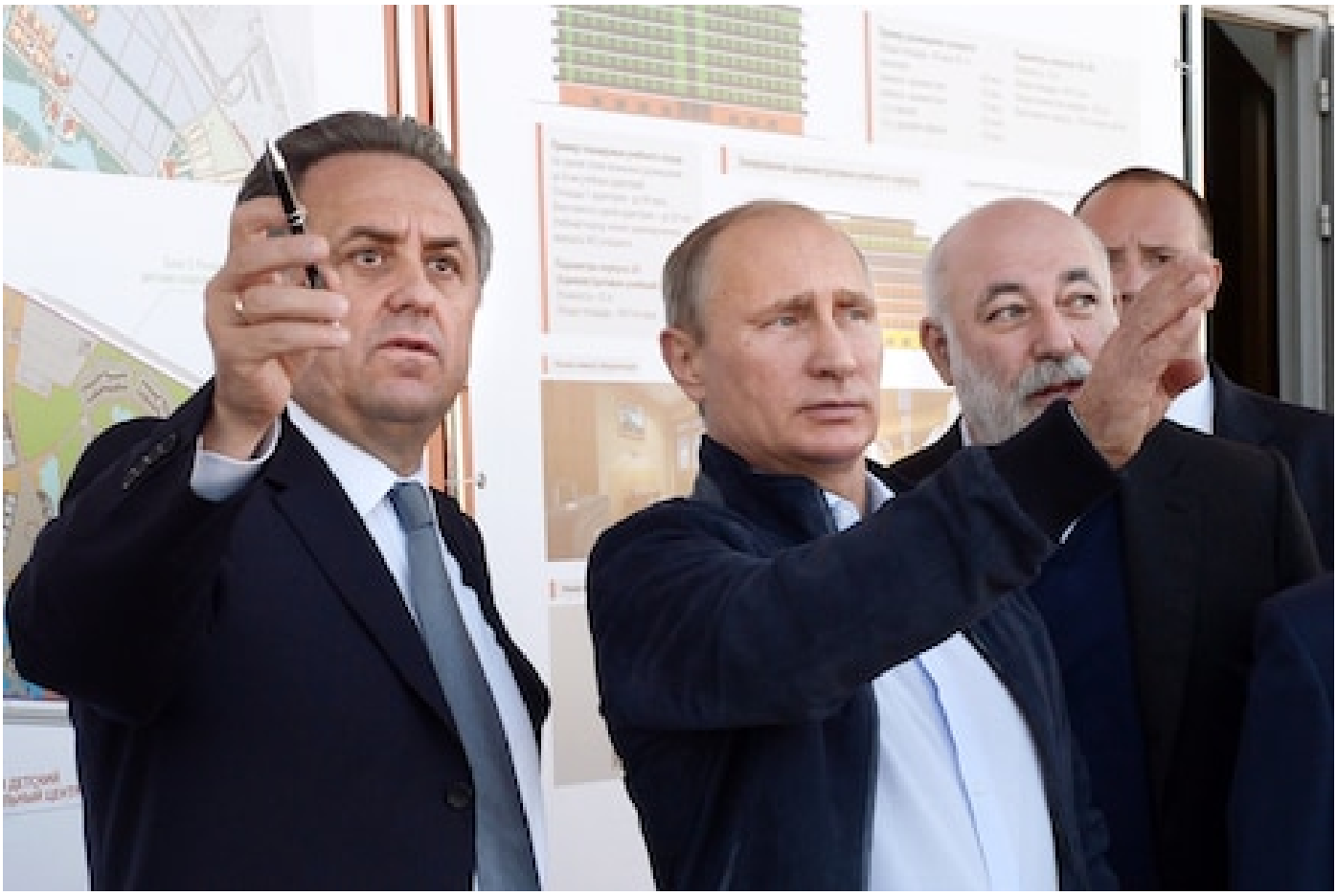
Western authorities expressed concerns over Russian technology funding as far back as 2014, when the Boston FBI publicly cautioned the Massachusetts Institute of Technology over an alliance with Skolkovo and its founding president, Viktor Vekselberg. Those concerns have intensified in the wake of Russia's war in Ukraine, as authorities expand their lists of Kremlin allies to sanction with restrictions on assets or business dealings.

The Russian Venture Co. was sanctioned by the U.S. government in February, and the Russian Quantum Center was added to the list in September. Also that month, the government said it might add anyone who had worked on quantum computing for Russia in the past.

Some of the expatriates have denounced the invasion of Ukraine, and many say they ended or reduced ties to Russia years ago. But those claims are not being accepted at face value, according to the people familiar with the inquiries. It is unclear what conclusions have been drawn by counterintelligence and other officials, since intelligence cases are rarely made public. The FBI declined to comment.

The investigations have proved challenging because of sparse or contradictory public disclosures, the role of shell companies, and the closed nature of venture capital and private equity firms, which have far fewer regulations requiring disclosures than publicly traded companies.

Beyond that, records show some of the people at issue have changed not only their professed political beliefs but also the records of their past positions, their company names and rosters, and their ownnames. But the network's high-level Russian government connections and its focus on strategic technologies have unnerved investigators, the people familiar with the probes said.



Russian President Vladimir Putin with Skolkovo Foundation President Viktor Vekselberg, right, and Russian Sport Minister Vitaly Mutko in 2014. (Alexey Nikolsky/AFP/Getty Images)

"If they were going to put Russian money into strategic technology in the U.S., this is exactly how you would do it," said one U.S. intelligence officer familiar with the inquiries. "Dark money going into [venture capital firms] in tech and politics we care about."

The reach of Russian technology has surprised officials in the past. In 2017, Russian security company Kaspersky Lab disclosed that its software had taken secret code for a U.S. hacking tool from an American customer. The discovery led to a directive that Kaspersky software be removed from U.S. government computer systems and promptedKasperskytoabandon plans for expansion in the United States.

An official at another intelligence agency confirmed that one of the people at the center of the new network, tech security company founder and investor Serguei Beloussov, was being tracked, but said the United States had not found proof of a security breach of the sort it found with Kaspersky. Beloussov said in an interview that he had not been contacted by U.S. authorities and that he is not close to the Kremlin.

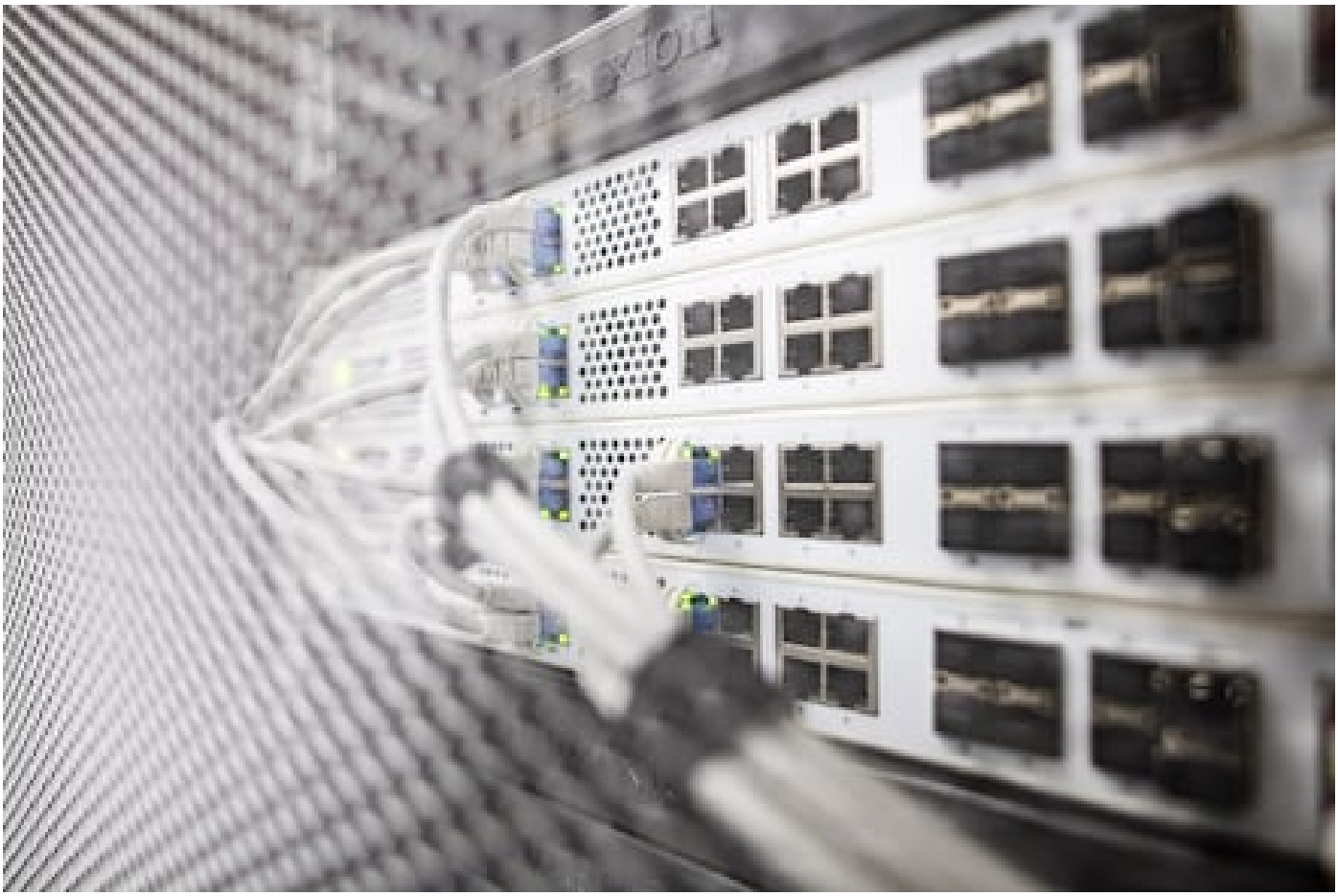
Of particular interest to investigators is Vekselberg, who has used a Northern California venture firm, Maxfield Capital, to invest in technology companies. He was named president of the Skolkovo Foundation on its launch in 2010.

Sanctioned in 2018 and again this year, Vekselberg had two U.S. properties raided in September and a yacht seized in April by authorities who said he had committed money laundering and bank fraud. Vekselberg has not been publicly charged with a crime. He could not be reached for comment.

Joining Vekselberg on the Skolkovo board was Alexander Galitsky, a talented engineer and inventor behind early virtual private networks and other gear. Long a partner with U.S. technologists, Galitsky served as coordinator for the government's Russian Venture Company and helped start the Russian Quantum Center, which took in millions from Skolkovo to put Russia ahead in the race for next-generation computing. He also created a venture firm in California, Almaz Capital, drawing such prominent co-investors as Cisco Systems and Andreessen Horowitz and pouring hundreds of millions of dollars into tech companies. A person close to Cisco said it last invested in an Almaz fund in 2013 and does not intend to do so again. Someone close to Andreessen said there had been one small investment.

The concerns about foreign pursuit of technology know-how go back decades. Attention focused on Russia not long after Skolkovo launched, when Galitsky helped set up the partnership between Skolkovo and MIT that drew the rare warning back in 2014.

In the Boston Business Journal, a deputy agent in charge of the city's FBI bureau wrote in a guest column: "The FBI recently released a notification to technology companies and research facilities, which include colleges and universities in the Boston area, warning them of the possible perils of entering into joint partnerships with foreign venture capital firms from Russia.



The Kaspersky Lab in Zurich. (Adrian Bretscher/Getty Images/Kaspersky Lab)

“The warning was based on the FBI’s growing concern that the purported reasons offered by the Russian partners mask their true intentions. The FBI believes the true motives of the Russian partners, who are often funded by their government, is to gain access to classified, sensitive and emerging technology from the companies.”

The article said Skolkovo “may be a means for the Russian government to access our nation’s sensitive or classified research, development facilities and dual-use technologies with military and commercial applications. This analysis is supported by reports coming out of Russia itself.”

Galitsky said by email that he had left Skolkovo and Quantum Center boards long ago and quit the board of the sanctioned Alfa Bank, the biggest private Russian bank, one day after Russia invaded Ukraine.

As the war in Ukraine escalated, Galitsky’s activities drew increased scrutiny in the United States. In April, Almaz said that despite a track record dating to 2008, it was getting revetted by Silicon Valley Bank. “We needed to prove that we are a good fund,” Galitsky told the venture capital publication PitchBook.

Galitsky told The Post that review had ended without a change in the fund's relationship to the bank and that Almaz was continuing to invest in U.S. companies. The bank declined to comment.

American and Swiss officials have also made inquiries about Soviet-era emigre Beloussov, according to two people informed of the matters. Beloussov has led a series of major companies, including the large security company Acronis, which won U.S. government contracts, including for backing up Pentagon computers, through at least 2017. Beloussov's first big company, a computing firm named Parallels, was backed by Vekselberg's Maxfield Capital and others. Acronis, officially Swiss, spun off from Parallels.

Though neither his corporate nor his personal webpage mentions it in recounting his career, Beloussov helped start the Russian Quantum Center, which worked with partners as sensitive as the national nuclear authorities. Beloussov took the role of chairman of the board of trustees.

In a 2019 interview, Russian President Vladimir Putin's special envoy for digital development, Dmitry Peskov, paid tribute to Beloussov's work for Russia and blessed his move abroad, saying that he could do more for the country from outside its borders.

"Acronis teams, even in private, do a lot for the country. The role, for example, of Serguei Beloussov, the leader of the Acronis team, in launching a large state-owned quantum computing system in Russia is not very public now, but it cannot be overestimated. He understood this before others, drove ahead of others and did a lot to make this story go. Therefore, they will find how to pay their debts to the Motherland," Peskov told the Russian business news site BFM.ru.

"If they are purely Russian, they will take on all the risks that Kaspersky took. Why would they fall into this trap? … A variety of income, profits, feedback for the country will be much greater than if they remained in Russia as a small 100 million [dollar] company."

Beloussov said he was not paid for his work at the Russian Quantum Center, formally known as the International Center for Quantum Optics and Quantum Technologies, and that he had helped because "at the time, it seemed to everyone that scientific collaboration was a good thing." He said his business ventures since then had no ulterior motive.

Beloussov, who has Singapore citizenship and changed his name to Serg Bell, also founded Runa Capital, a venture firm with offices in multiple cities, including, until this year, Moscow. After Russia invaded Ukraine, Beloussov criticized the attack on Twitter and told The Post he was moving Acronis employees out of Russia.

He said Acronis had already stopped selling in Russia. But a company called Akronis-InfoProtection modified and resold Acronis's security wares, which were government-approved for use in ministries. When Akronis changed its name last year to Cyber Protect, its chief executive said Acronis "remains the key technology partner." Beloussov said Acronis had ended the licensing deal.

Runa's portfolio includes many companies in cutting-edge technology, including quantum computer maker Pasqal of France, Swiss "quantum-safe" security and encryption company ID Quantique, and Enteria, a German company making an operating system software for industrial devices.

As previously reported, Beloussov also hired Masha Drokova, now married and known as Masha Bucher, who was once an ardently pro-Putin teenager who starred in a documentary that featured her kissing the autocrat. Drokova became a spokesperson for Nashi, a youth group that physically harassed Putin opponents. She has since publicly repudiated Putin, adding that her disavowal put her Russian family at risk.

Drokova worked for Beloussov at Acronis, at Runa, and at his 2012 fund for investing in quantum computing, called Quantum Wave Fund, which aimed at Silicon Valley. Then Drokova began investing through her own new fund, Day One Ventures, which also took money from Beloussov. This year, she denied acting for Russia or even taking money from Russians. Fundraising pitches to potential Day One investors, seen by The Post, touted connections to Russian billionaires who were later sanctioned. Drokova said the pitches were fakes.

Whileindependent, Quantum Wave Fund had connections to the government-funded Quantum Center besides Beloussov. The president and then chairman of the nonprofit, Sergey Viktorovich Kuzmin, became managing partner of the Quantum Wave Fund's initial effort, according to Russian regulatory filings. Kuzmin told The Post he had never worked with the Russian government and that his name is better translated as Kouzmine.

Beloussov served as an adviser to the Russian Venture Company through 2015. After the Quantum Wave Fund, Beloussov started a new investment fund, Phystech Ventures, with others including Galitsky and former Skolkovo investment manager Petr Lukyanov.

Lukyanov said Phystech took over management of the Quantum Wave Fund. Then it launched another fund, called TF II for Terra Fund, with money from the Russian Venture Company and others. The Russian fund put in about $15 million, according to its 2015 annual report. That was still in play at least in 2019, when it combined with another fund.

Beloussov said he did not participate in that fund, and Lukyanov said he resigned from it this year. Instead, the men are focused on a new firm, registered in December as Terra.VC.

Most of the money for it was to come from Russian investors, according to internal documents reviewed by The Post. But after the invasion of Ukraine, Lukyanov and the rest of the management knocked them out.

Beloussov said he has not been to Russia since 2017, and he has spoken out against the attack on Ukraine.

But he has stood by multiple previous high-level Putin supporters, including Robert Schlegel, who spoke for Nashi when it claimed credit for cyberattacks on Estonia, then served in the Russian parliament in Putin's United Russia Party. While in parliament, Schlegel traveled abroad to facilitate alliances with movements in other countries, including Germany's far-right Alliance for Democracy Party.

Schlegel disappeared from the Duma and the spotlight in 2016 until he was rediscovered in 2019 by a German newspaper, which found him working in Munich as a director of strategic projects for Acronis.

After the newspaper story, Acronis had one of its law firms interview Schlegel. A partner at that firm, former U.S. federal judge Eugene Sullivan, told The Post that Schlegel had committed no crime and that if he had been an intelligence risk, he would not have been granted German citizenship.

Schlegel resigned from the company but has continued to consult for it, Sullivan said. He did not respond to an interview request.

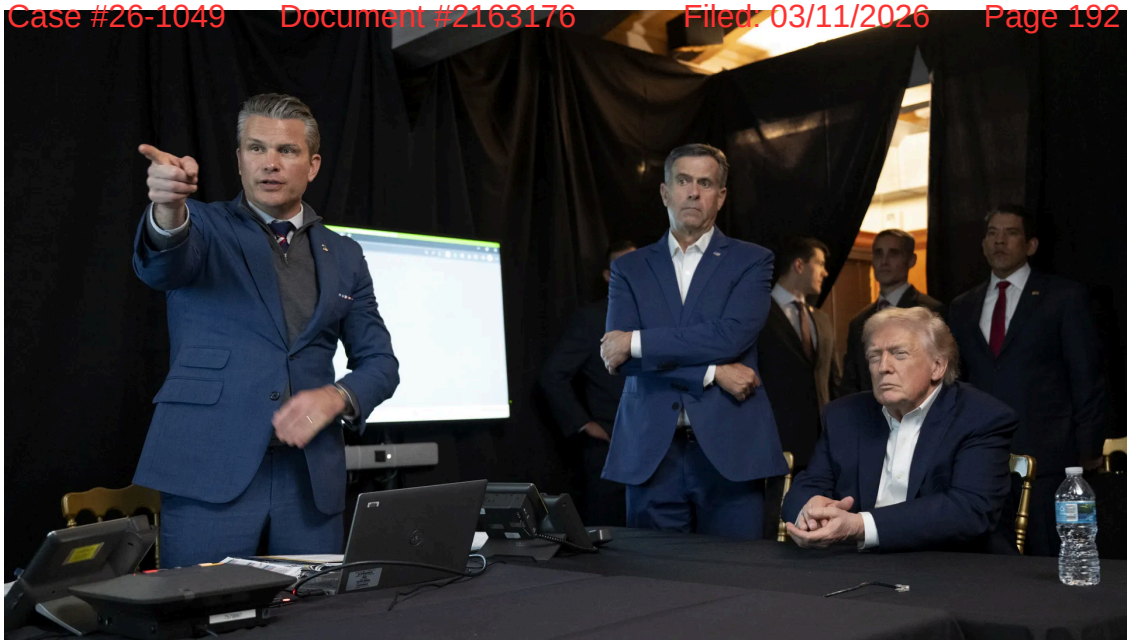
He joined Beloussov on a trip to Montenegro this year as Beloussov looked for possible business locations, according to the capital city's investment office.

EXHIBIT 10

Feb 16, 2026 - Technology

Exclusive: Pentagon threatens Anthropic punishment

Dave Lawler, Maria Curi, Mike Allen



Defense Secretary Hegseth *(left)*, CIA Director Ratcliffe and President Trump during the Maduro raid in January. Photo: Molly Riley/White House via Getty Images

Defense Secretary Pete Hegseth is "close" to cutting business ties with Anthropic and designating the AI company a "supply chain risk" — meaning anyone who wants to do business with the U.S. military has to cut ties with the company, a senior Pentagon official told Axios.

- The senior official said: "It will be an enormous pain in the ass to disentangle, and we are going to make sure they pay a price for forcing our hand like this."

Why it matters: That kind of penalty is usually reserved for foreign adversaries.

Chief Pentagon spokesman Sean Parnell told Axios: "The Department of War's relationship with Anthropic is being reviewed. Our nation requires that our partners be willing to help our warfighters win in any fight. Ultimately, this is about our troops and the safety of the American people."

The big picture: Anthropic's Claude is the only AI model currently available in the military's classified systems, and is the world leader

for many business applications. Pentagon officials heartily praise Claude's capabilities.

- As a sign of how embedded the software already is within the military, Claude was used during the Maduro raid in January, as Axios reported on Friday.

Breaking it down: Anthropic and the Pentagon have held months of contentious negotiations over the terms under which the military can use Claude.

- Anthropic CEO Dario Amodei takes these issues very seriously, but is a pragmatist.

- Anthropic is prepared to loosen its current terms of use, but wants to ensure its tools aren't used to spy on Americans en masse, or to develop weapons that fire with no human involvement.

The Pentagon claims that's unduly restrictive, and that there are all sorts of gray areas that would make it unworkable to operate on such terms. Pentagon officials are insisting in negotiations with Anthropic and three other big AI labs — OpenAI, Google and xAI — that the military be able to use their tools for "all lawful purposes."

- A source familiar with the dynamics said senior defense officials have been frustrated with Anthropic for some time, and embraced the opportunity to pick a public fight.

The other side: Existing mass surveillance law doesn't contemplate AI. The Pentagon can already collect troves of people's information, from social media posts to concealed carry permits, and there are privacy concerns AI can supercharge that authority to target civilians.

- An Anthropic spokesperson said: "We are having productive conversations, in good faith, with DoW on how to continue that

work and get these new and complex issues right."

- The spokesperson reiterated the company's commitment to using frontier AI for national security, noting Claude was the first to be used on classified networks.

The stakes: Designating Anthropic a supply chain risk would require the plethora of companies that do business with the Pentagon to certify that they don't use Claude in their own workflows.

- Some of them almost certainly do, given the wide reach of Anthropic, which recently said eight of the 10 biggest U.S. companies use Claude.

- The contract the Pentagon is threatening to cancel is valued at up to $200 million, a small fraction of Anthropic's $14 billion in annual revenue.

Friction point: A senior administration official said that competing models "are just behind" when it comes to specialized government applications, complicating an abrupt switch.

The intrigue: The Pentagon's hardball with Anthropic sets the tone for its negotiations with OpenAI, Google and xAI, all of which have agreed to remove their safeguards for use in the military's unclassified systems, but are not yet used for more sensitive classified work.

- A senior administration official said the Pentagon is confident the other three will agree to the "all lawful use" standard. But a source familiar with those discussions said much is still undecided.

Add Axios on Google

What to read next

Feb 14, 2026

Exclusive: Pentagon threatens to cut off Anthropic in AI safeguards dispute

Go deeper (2 min. read) →

Feb 25, 2026

Scoop: Pentagon takes first step toward blacklisting Anthropic

Go deeper (3 min. read) →

EXHIBIT 11

Anthropic's and OpenAI's Dance With the Pentagon: What to Know

nytimes.com/2026/03/07/technology/anthropic-openai-pentagon-dario-amodei-sam-altman.html

Cade Metz March 7, 2026

Sign up for the On Tech newsletter. Get our best tech reporting from the week.

Late last month, Defense Secretary Pete Hegseth delivered an ultimatum to Anthropic, the only company that had provided the Pentagon with artificial intelligence technologies for use on classified systems.

If Anthropic did not allow the Pentagon to deploy these technologies for "all lawful uses," Mr. Hegseth said, he would sever ties with the San Francisco start-up.

The threat set off a chain of events that resulted in the Defense Department's labeling Anthropic a "supply chain risk," which would prevent all military contractors from using the company's technologies, and signing an agreement with OpenAI, its biggest rival.

The negotiations were, to say the least, confusing.

Here is a guide to the unusual discussions involving the Defense Department, Anthropic and OpenAI.

How does the Pentagon use Anthropic's technology?

Anthropic's technologies are widely used inside the Defense Department because the start-up agreed last year to integrate its systems with technology from Palantir, a data analytics company that is approved for classified operations.

Separately from Anthropic's partnership with Palantir, the Pentagon also uses Anthropic's technology to analyze imagery and other intelligence data as part of a $200 million A.I. pilot program.

Anthropic's technology is being used as U.S. military forces engage in a widening war against Iran, two people familiar with the technology said on the condition of anonymity.

Google, OpenAI and Elon Musk's xAI are also part of the pilot program, but are not yet used on classified systems. Anthropic was a step ahead of its rivals thanks to its partnership with Palantir.

Why did the Pentagon get angry at Anthropic?

On Feb. 15, The Wall Street Journal reported that Anthropic had raised concerns with Palantir about the role its technologies played in the U.S. military operation to capture Venezuela's president, Nicolás Maduro. The story inflamed earlier tensions, as Mr. Hegseth and others at the Pentagon argued that Anthropic was resisting the military's use of these A.I. systems.

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

The Defense Department was already in talks with Anthropic to establish new contractual language that allowed the Pentagon to use the company's technologies for any lawful purpose. But Anthropic was reluctant to agree to those terms.

Why was Anthropic reluctant?

Anthropic wanted contractual language that prevented the Pentagon from using its technology with autonomous weapons or for mass surveillance of Americans. It argued that specific language was needed to ensure that the technologies were used only in ways that aligned with what they could “reliably and responsibly do.”

The Pentagon said private companies should not try to control how the military operated.

On Feb. 24, Mr. Hegseth met with Anthropic’s chief executive, Dario Amodei, and said that if Anthropic failed to agree to the Pentagon’s demands by 5:01 p.m. on the next Friday, he would designate the company a supply chain risk.

Image

Anthropic resisted the terms the Pentagon wanted in its contract.Credit...Marissa Leshnov for The New York Times

What does it mean to be a supply chain risk?

It means that a company’s technology cannot be used by the Pentagon or any of its contractors in their work with the government. The designation is typically applied only to firms with ties to the government of China.

Did cooler heads prevail?

No. The company published a blog post saying it could not “accede” to the Pentagon.

Minutes after the deadline passed, Mr. Hegseth deemed Anthropic a supply chain risk in a post to social media.

He added that “no contractor, supplier or partner that does business with the United States military may conduct any commercial activity” with the company. But the Pentagon planned to continue to use Anthropic’s technologies for up to six months as it arranged for alternatives.

The Pentagon later sent a letter to Anthropic saying it had officially designated the company as a supply chain risk.

Trump Administration: Live Updates

- Trump says he’s open to Iran playing at the World Cup this year.
- The Trump administration is restarting the Global Entry program.
- Bondi is said to move to military housing because of threats.

Does Hegseth have the power to do that?

A court will probably decide. Anthropic has said it intends to sue the government, and legal scholars say a suit would most likely be successful.

“Anthropic’s case is very strong,” said Alan Rozenshtein, a professor of law at the University of Minnesota.

Legal scholars also say the Pentagon does not have the power to bar its contractors from commercial activity with the start-up beyond just using its technology. For instance, it cannot prevent contractors from investing in Anthropic, they said.

“The commercial activity language is flatly illegal,” Mr. Rozenshtein said.

That is an important point because Amazon and Google — two of Anthropic’s biggest investors — are also Defense Department contractors.

In a statement on Anthropic’s website, Dr. Amodei said Anthropic was still in discussions with the Pentagon over their contract. But Emil Michael, chief of technology for the Defense Department, quickly responded on social media that there were “no active” negotiations between the two.

Why didn’t the Pentagon just stop using Anthropic?

That would have been an easier solution to the dispute. “The correct response is to just cancel the contract and walk away,” Mr. Rozenshtein said.

Instead, the Pentagon appeared to make a political statement by labeling Anthropic a supply chain risk.

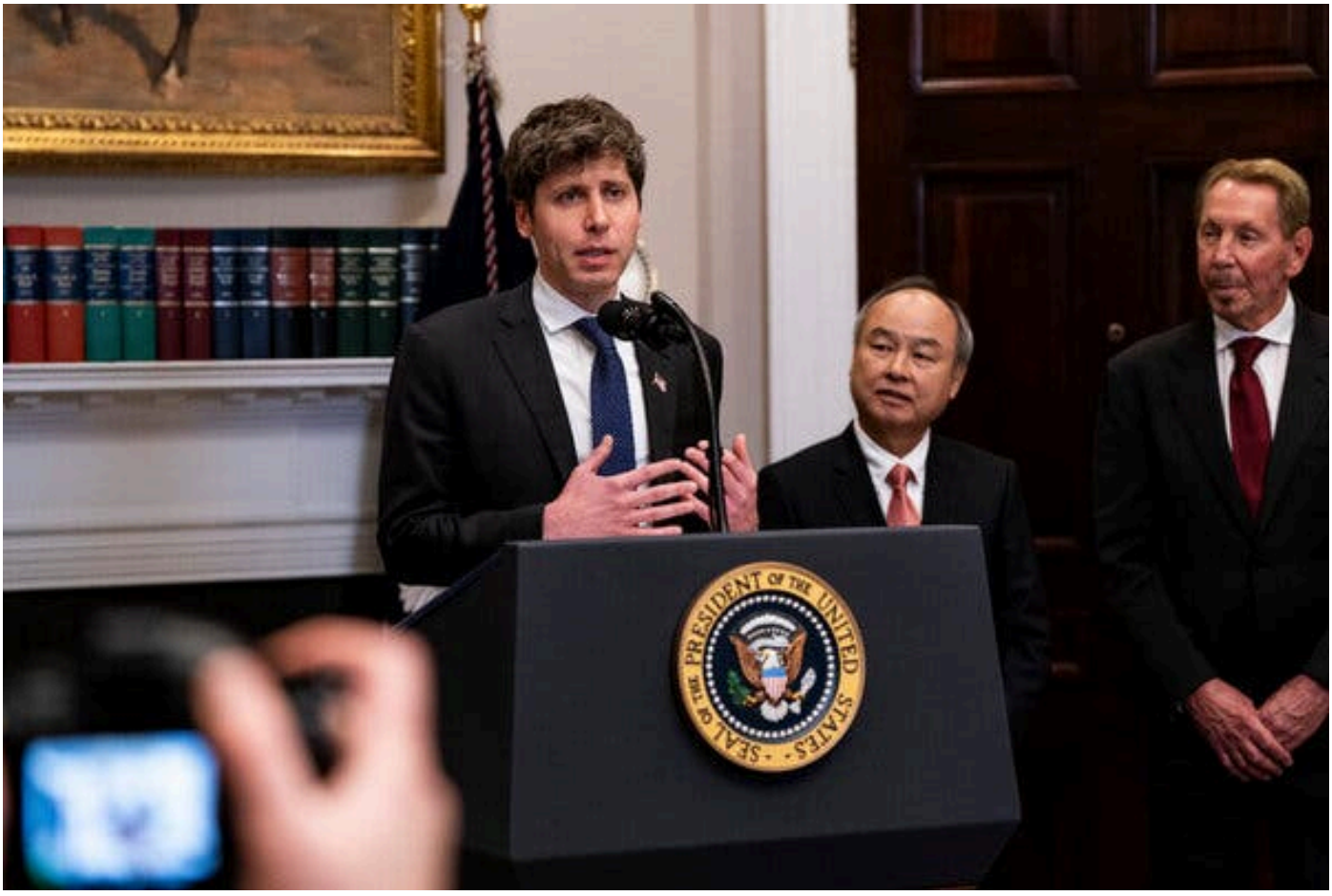
“It seems like the Pentagon just does not like Anthropic’s general political vibe and wants to destroy its entire business,” said Dean Ball, a senior fellow at the Foundation for American Innovation who was previously a policy adviser for A.I. under President Trump. “That is beyond the pale.”

How did OpenAI get involved?

A day after Mr. Hegseth met with Dr. Amodei, OpenAI’s chief executive, Sam Altman, started his own talks with the Defense Department.

Mr. Altman told the Pentagon that it should not give Anthropic the supply chain risk label because it would have a chilling effect on the department’s relationship with the tech industry. Like Anthropic, he said, OpenAI did not want its technologies used for mass surveillance of Americans or with autonomous weapons.

Image



OpenAI’s chief executive, Sam Altman, at the White House last year.Credit...Haiyun Jiang for The New York Times

But Mr. Altman and OpenAI also worked on their own contract with the Pentagon. Just hours after Anthropic missed its deadline, he announced that they had reached an agreement.

OpenAI agreed to let the Pentagon use its A.I. systems for any lawful purpose. But OpenAI also said it had negotiated terms that allowed the company to uphold its safety principles by installing specific technical guardrails on its systems.

Can technical guardrails prevent A.I. from being used for mass surveillance?

No. The guardrails built into today's A.I. do not always work as they are designed. And even when these guardrails hold firm, there are many ways A.I. systems could still be used to feed surveillance or the use of autonomous weapons.

Three days later, OpenAI announced that it had amended its agreement with the Pentagon. It added language saying its A.I. systems "shall not be intentionally used for domestic surveillance of U.S. persons and nationals."

People following this odd contract shuffle argued that the Pentagon had made an agreement with OpenAI that it refused to make with Anthropic. This was another sign, they said, that the Pentagon's response to Anthropic was politically motivated.

Does the amendment uphold OpenAI's safety principles?

Maybe not. Legal experts point out that the Pentagon could inadvertently collect data about Americans as it worked to monitor foreigners and that it would still be allowed to analyze this data under the terms of the contract.

A contract like this is also difficult for a private company to enforce, because a violation of the terms may not be obvious, Mr. Rozenshtein said. In other words, whether a technology has been used for mass surveillance is sometimes open to debate.

Even if the government breaches the contract, OpenAI can at most cancel service and sue for damages, but it cannot force the government to live up to its end of the bargain, Mr. Rozenshtein said.

Mr. Altman and OpenAI also said the Pentagon had assured the company that its technology would not be used by defense intelligence agencies, including the National Security Agency. But OpenAI could, of course, sign a separate agreement that allows the N.S.A. to use its technologies.

So, what does all this mean?

“This is not just some dispute over a contract. This is the first conversation we have had as a country about control over A.I. systems,” Mr. Ball said. “What should the limitations be? And who gets to decide?”

But he and other experts said this was not the best way to decide these questions. They say Congress should step in to set firmer laws.

“Congress should be asking hard questions about this,” said David Bader, a professor at the New Jersey Institute of Technology. “We need deliberate bipartisan framework for the governance of A.I.”