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

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

Attorneys for Plaintiff Anthropic PBC

KELLY P. DUNBAR (*pro hac* forthcoming)
kelly.dunbar@wilmerhale.com
JOSHUA A. GELTZER (*pro hac* forthcoming)
joshua.geltzer@wilmerhale.com
KEVIN M. LAMB (*pro hac* forthcoming)
kevin.lamb@wilmerhale.com
SUSAN HENNESSEY (*pro hac* forthcoming)
susan.hennessey@wilmerhale.com
LAUREN MOXLEY BEATTY (SBN 308333)
(application pending)
lauren.beatty@wilmerhale.com
LAURA E. POWELL (*pro hac* forthcoming)
laura.powell@wilmerhale.com
SONIKA R. DATA (*pro hac* forthcoming)
sonika.data@wilmerhale.com
WILMER CUTLER PICKERING
HALE AND DORR LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ANTHROPIC PBC,

Plaintiff,

v.

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

Defendants.

Case No. 3:26-cv-01996

DECLARATION OF JARED KAPLAN

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

Personal Background

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

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

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

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

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

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

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

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

Anthropic's Background As An AI Safety Company

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

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

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

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

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

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

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

Anthropic's LLM Claude And The Role Of Guardrails

14. Anthropic's signature model is a general-purpose LLM called Claude. We make Claude available to individual users, small businesses, and large organizations through a variety of offerings. We continually develop and release increasingly capable versions of Claude, 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 9, 2026.

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

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

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

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

Michael J. Mongan