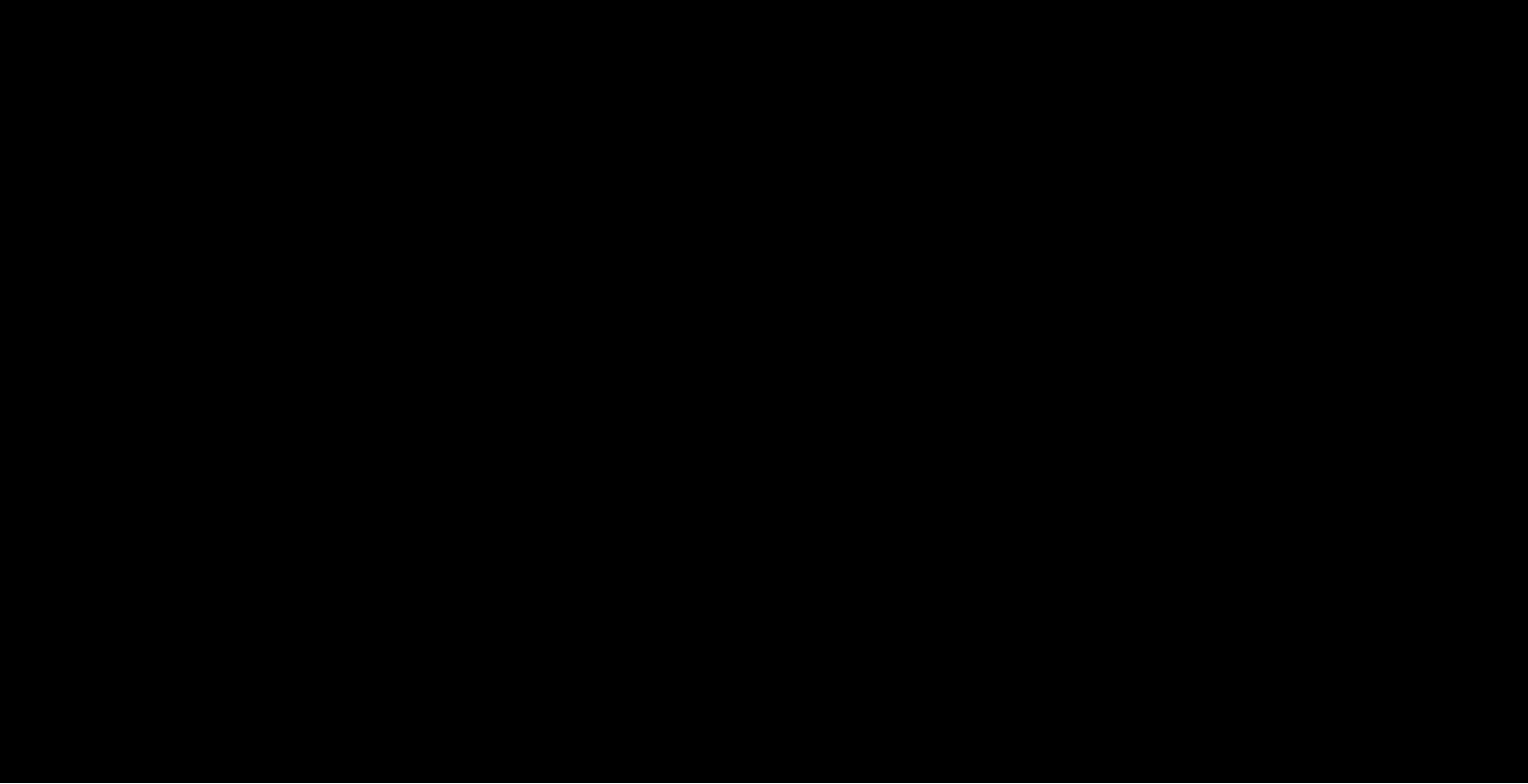




From: Dario Amodei <dario@anthropic.com>
Sent: Sunday, February 15, 2026 10:27 PM
To: Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil>
Cc: Sarah Heck <sheck@anthropic.com>
Subject: Re: Thank you for the meeting and next steps

Emil,

Thank you for this. A few points of clarification, because I think some of our intent may have gotten lost in the back and forth on legal language.

First, we absolutely do not intend to ask DoW to agree to anything that implies it is trying to do something unlawful, and we would not propose language that is merely symbolic. That is not and has never been our position.

On domestic surveillance specifically: our proposal authorizes the Department to use our tools for foreign intelligence and counterintelligence consistent with existing law, and we want to continue to be a strong partner to you here. However, the limits we propose relate to surveillance on Americans because such surveillance could be conducted lawfully via existing authorities — they do not merely rule out domestic surveillance activities that are already illegal. The Department does have authorities to conduct domestic surveillance under the law (e.g., queries on Section 702 data using U.S. person identifiers, using geolocation data around sensitive facilities or analyzing commercially available geolocation data, or providing DoW tools in support of other agencies); hence our proposed restrictions. Using Claude to conduct scaled analysis on LinkedIn or Wikipedia pages in the course of foreign intelligence or counterintelligence missions would not be prohibited under our proposal, and it is also not our intention to prohibit one-off LinkedIn or Wikipedia searches for Americans. Our underlying concern here is that AI fundamentally changes what is possible with surveillance on Americans in ways that existing law was not written to address, but our proposed restrictions should have no impact on the overwhelming majority of the Department's use cases.

Second, our proposal would authorize weapons development. Our request is that a human supervise any weapons development (in other words, weapons development cannot itself be completely autonomous) - a non-issue today, but potentially an issue in the future. Otherwise, our language follows what I understood to be something you said you could live with: providing that a human on the loop be able to override or disable an AI system before it executes a kinetic operation. Even then, we believe this restriction is needed only until such a time that frontier AI has proven to be more reliable and safe to civilians and our warfighters. With respect to the systems that would be excepted from the prohibition, we have listed the systems that we discussed when we met (e.g., counter-unmanned systems and missile defense systems), but we are open to adding others that you believe to merit inclusion.

Third, on the safety carveout in Section XV — we need to preserve our ability to continue developing the quality assurance systems for accuracy, reliability, lawfulness, and the terms we agree to, which is core to how the product works. But the intention is not to undermine our agreed terms.

Finally, on your note about coordination among the labs. I want to be direct: we have not coordinated terms of use with any other company, and we would not do so. Our positions reflect our own deployment principles and our own assessment of the technology. We have heard reports of what other labs might or might not be doing, but there is no coordination.

With these clarifications, I think we in fact could be close to an agreement. I hope you agree. I'd welcome the chance to talk through any other concerns so that we can resolve these sorts of misunderstandings in real time. I understand that there might be an opportunity to meet with the Secretary on this matter, and I would like to do everything we can to come to an agreement on mutually acceptable terms well before then.

All this being said, should we be unable to come to an agreement, I hope we can find a way to part ways without recrimination or disruption to the Department's AI-supported operations. I both understand and respect your and the Secretary's position on this matter, and we are not looking to posture or virtue-signal.

Let me again note that Anthropic is deeply proud of our work with the Department of War, and we hope we can continue our partnership to support and equip our warfighters with the best AI capabilities in the world.

Dario

On Sun, Feb 15, 2026 at 1:49 PM Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> wrote:

Dario,

I hope you have a good trip to India. I laid out several important issues in my last mail that I think we need to hear back from Anthropic on at this point before we have further dialogue on the totality of the issues. As for the mass surveillance topic, I was clear that the DoW will comply with all applicable laws and regulations and always have said so. At this point I don't think there is further discussion to be had. As I have always said, even the mention of it is therefore unnecessary in the contract, but I was trying to be responsive to your sensitivity on it. The language we get back from Anthropic continues to be more expansive than referencing to existing laws (e.g. the last language you sent would prohibit a search for someone's LinkedIn or Wikipedia!), so I think we just need to shelve this issue because it does not seem to be bearing fruit.

As for the press, it's all unfortunate. I do hope you can respond to my previous thorough comments when you have time. Thanks again.

Emil

From: Dario Amodei <dario@anthropic.com>
Sent: Saturday, February 14, 2026 7:16 PM
To: Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil>
Cc: Sarah Heck <sheck@anthropic.com>
Subject: Re: Thank you for the meeting and next steps

Emil,

Thanks for the questions and substantive engagement on our proposal. I am on my way to the India AI summit, but we would certainly welcome your running these terms by your specialist attorneys. Once you've done that, do you want to share their feedback, and we can then provide consolidated clarifications/answers on all of these questions/issues at once?

I've also been disappointed by the news stories that have come out over the last 24 hours, with "senior administration officials" weighing in on the record to attack us with misleading information. As you know, neither I nor Anthropic has ever discussed the Maduro operation with you or anyone else at DoW. Our concern is terms of use at the principles level and the principles we are discussing do not intersect in any way with that operation. I appreciated your comment about noise in the press and would also appreciate anything you can do to prevent more distractions like this. We may make a short statement to journalists, simply to clarify our position.

Dario

On Thu, Feb 12, 2026 at 10:16 AM Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> wrote:

Hi Dario,

Thank you for the thoughtful reply and the continued dialogue. I do have a few questions as there was a fair bit to parse out here although I appreciate the simplifications that you made. A few, but not exhaustive, follow up questions:

Section XV: The last sentence which appears to exempt Anthropic from the License grant for 'safety, accuracy, or reliability' can swallow the rest of the paragraph if I am reading correctly. If for instance, Anthropic decides that something that is otherwise lawful (holding aside the limitations in XVI etc for the purposes of discussion) but not safe, then it seems we end up back where we started this discussion two months ago.

Section XVI: Many issues here, but the easiest one to try to decipher intent is that DoW cannot engage in weapons development. As you know, there is lots of physics, material science etc that go into the development of weapons (whether one considers them offensive or defensive). As Tarun noted in our in-person meeting, there is no distinction in our world between weapons that are defensive or offensive (e.g. a drone that takes down other drones can potentially do either activity). Further, the science that goes into such development is something that would be surprising to me that would be objectionable to Anthropic. Additionally, while I understand the desire to have a human be able to disable the system and appreciate the exceptions to the exception for missile defense and counter-unmanned, there are many, many use cases beyond those which can be imagined in a Department of 3 million people. For example, any attack on the US or our troops by anything

capable of hypersonic speeds or that otherwise do not give a human time to respond in any reasonable way due to distance or a stealth attack.

Section XVII: This seems conceptually workable but this something I would want to run by specialist attorneys versus generalists so there is no ambiguity if we get to that point. My key principle here is that we believe there are robust laws on mass surveillance and have every intention of following those laws in total. The first part of the paragraph could be read to limit looking up the equivalent of someone's LinkedIn or Wikipedia profile which I am sure is not the intent.

Section XVIII: Assuming this is the same language we originally sent over (since I do not see redlines), this would be acceptable obviously.

I sincerely appreciate your attention to these issues and I hope we have conveyed our constructive spirit as well. It's worth reiterating that our mission at DoW is highly varied and we are larger than all other USG agencies combined. We run schools, hospitals, protect embassies, defend bases and allies, conduct deep science in quantum, additive manufacturing etc so its very, very difficult to cover all the use cases relative to say your partnership with the intelligence agencies or combatant commands which have a smaller cadre of users and missions.

I know there is a lot of noise in the press and I can assure you that it is not coming from DoW and I will continue to monitor on our side. I ask that you do so on your side as well as it serves neither of our of purposes. Finally, while I know there is competition amongst the AI labs for talent, I want to ensure there is no undue collaboration/collusion on terms of use as you mentioned to me that employees sometimes factor these kinds of policies into their decisions on where to work and we have a responsibility to the taxpayer to have full and fair competition for any partner we choose. If preferable, I am happy to discuss on the phone or in person as well.

Emil Michael

Under Secretary of War for Research and Engineering

From: Dario Amodei <dario@anthropic.com>

Date: Monday, February 9, 2026 at 3:15 PM

To: "Michael, Emil HON (USA)" <emil.g.michael.civ@mail.mil>

Cc: Sarah Heck <sheck@anthropic.com>

Subject: Re: Thank you for the meeting and next steps

Emil,

Thank you for your last note. I respect your position.

The attached terms reflect our effort to limit our exceptions to either clarification of legal authorities or current limitations in the capabilities of the models themselves. They have also been redrafted to be brief and clear. I hope these terms provide a way forward.

With respect to mass domestic surveillance, these terms would address the issues we've discussed, while enabling support for DoW personnel working on foreign intelligence and counterintelligence missions.

With respect to fully autonomous kinetic operations, the terms reflect the still-limited capabilities of current models to reliably support such operations, but also our commitment to improve model capabilities to meet such requirements in the future. In addition, and to be clear, our language does not prevent support for autonomous kinetic operations, but provides only that human operators need to remain on the loop until frontier AI can demonstrably perform these functions with greater accuracy and less risk to warfighters and non-combatants – a standard we believe serves the Department's interests as much as ours.

Under these terms, Anthropic would provide our models to the DoW with fewer restrictions than for any other customer. We would enable support for weapons development, offensive cyber operations, and many other applications. We believe these terms are consistent with our track record of commitment to national security. Anthropic was the first company to tailor frontier AI for the national security community and also to put advanced AI capabilities in the hands of U.S. defense and intelligence users on classified networks.

I strongly share your sense of urgency about harnessing America's AI advantage, and hope we can continue to put Anthropic's technology to work in service of the Department's mission and America's warfighters. I believe our capabilities are superior to other model developers in key respects, and I hope that the proposal we put forward will provide a basis for us to continue to supply the Department with frontier AI capabilities.

I also share your view that a "grudging" partnership serves no one. If our position on fully autonomous kinetic operations does not make us the right vendor for the program at this moment, I would respect that, but it would not diminish the urgency I feel to support the Department's mission, and I hope we can talk soon about the ways Anthropic could do so in any event.

Dario

On Wed, Feb 4, 2026 at 10:19 PM Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> wrote:

Hi Dario,

I appreciate the continued efforts. I am trying to get to a mutually agreeable place in good faith that is consistent with what the Department needs and to be even-handed with all frontier lab companies. The partnerships at Indopacom, Centcom and potential work with the Navy and with Orchestrator (among other components and future opportunities) are very much things that I would like to have no roadblocks on. The most recent proposal you sent, however, is just not workable as it was both inconsistent with our basic first principles and it expanded, in many ways, the complexity (e.g. adding a reporting requirement to Anthropic for DoW vis a vis our own manuals/policies!). I think we have one more chance to align on core principles that would lead to legal language before it becomes not a good use of time for you or the DoW. Instead of sending you a redline, I wanted to outline the core principles in language you have already seen from us with the additional accommodations I have committed to consider. If you would base your next response on that, I would appreciate it and think it would form the basis for continued legal language discussions and in-person conversations. Here is my view of those core minimum principles:

1. The DoW (and its components and combat support agencies) need to be able to use Anthropic's products for all lawful use cases. Agreeing to this, conceptually, would eliminate the need for entire sections of your proposals that are aimed at parsing concepts that are sure to be outdated as soon as we sign anything as I both cannot know all the potential corner cases nor do we control what our country's adversaries are planning to do that would require us to defend ourselves in ways we haven't anticipated. Additionally and because you raised this specifically, any unlawful use cases related to mass surveillance would be prohibited. We can cite a statute or set of statutes if you insist, but we cannot agree to a highly customized set of concepts that Anthropic might worry about generally or as precursors as that would be impossible for us to implement and they seem based on a conception of what mass surveillance means that has not been vetted through the democratic process. I know you mentioned in our in person meeting that the democratic process was not likely to keep up with innovations, but my belief is that we have the best system in the world with checks and balances and no entity is above that system...or we really don't have a system at all. It's incumbent upon all of us to work on improving our system if we don't like what it produces as that is the core value we hold dear relative to our alternative systems
2. Anthropic does not need to retrain its models or weights for DoW, but we would not want to be constrained by Anthropic's policy preferences if the use cases are otherwise legal.
3. Like you, we also worry about the unknowns as most recently illustrated by OpenClaw/Moltbot and future novel things that put people, the country at risk. I am open to discussing how we ensure that there are 'off switches' and other protective measures over time and in partnership, but parsing language about how and the degree to which human agency/oversight needs to be employed is impossible given the size of the Department and the rapid acceleration of innovations that are happening in AI. We should and will have dialogue with our American AI champions and we will align our DoW policies accordingly as we are ultimately aiming at the same goal of ensuring that democratic values continue to be the North Star for all humans

4. I understand that there are strains of thought in the Valley that are skeptical of the government and those may be concentrated in any set of companies at any given time and more pronounced depending on the events of the day, but I will say that we are at a point in time where the confluence of the potential impact of AI and the threat environment require a level of trust and partnership that has not been more important in our lifetime

Given all these dynamics, I am asking that we base the next phase of our partnership assuming a basic level of trust which means starting with our principle of all lawful use cases and the rest of the concepts above without any of the extraneous verbiage. Finally, I really do not want begrudging agreement on this because I don't want consistent conflict or suspicion that could undermine a real partnership. If that is not possible with you or your management team and employees etc, I don't want to force anything unnatural because there is too much to do in this environment. Let me know what you think.

Emil

From: Dario Amodei <dario@anthropic.com>
Sent: Tuesday, February 3, 2026 7:57 AM
To: Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil>
Cc: Sarah Heck <sheck@anthropic.com>
Subject: Re: Thank you for the meeting and next steps

Dear Emil,

Thank you for your reply and message about a robust continued partnership with Anthropic. I appreciate that and feel the same way.

As we discussed when we met at the Pentagon, I am sympathetic to your concern about a clear and executable policy across frontier labs, which is why we have agreed to shift from a usage policy of affirmative permissions to a policy of “any lawful use” with narrow and precise exceptions.

I am attaching here the draft contractual language (red text is ours, black text is your original). We have tried to be as precise as possible, and we believe that these clearly defined exceptions would not require the kind of paralytic legal process that you are understandably concerned about.

Our offer to partner on R&D for autonomous systems—to build capability and reliability for future systems—remains open as well, but is not included in the contractual language.

With respect to model provision, as you know, Anthropic has provided the USG with a dedicated model for national security users—ClaudeGov—and would continue to do that. We (and our partners such as Palantir) are not aware of any significant user challenges with this model. We do believe classifiers can provide important information that can support compliance reporting within the USG, and need not and should not impede usage on agreed terms. (The issue with the CDC, which was using our commercial model, is now almost resolved.) We also note that your team's contractual language indicates that there is no requirement for retraining of the model or alteration of model weights, which we appreciate.

Finally, I want to affirm that Anthropic did not seed the recent press stories. We do not know who is doing so, but negotiating in public this way is certainly not our preference. If further press queries come our way, we intend simply to ensure the facts are straight and reiterate our desire to partner with you.

I look forward to hearing from you and remain happy to talk as helpful.

Best,
Dario

On Thu, Jan 29, 2026 at 1:05 PM Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> wrote:

Hi Dario,

Thanks for your note. As I mentioned before but I want to reiterate, we at the DoW very much want a robust continued partnership with Anthropic. We very much respect what you have built and will build in the years ahead.

They key for us, with 3 million+ employees is to have an executable policy that doesn't require a lawyer to parse or try to interpret whether something is or is not in the letter and spirit of a contract sitting beside every employee!

As such, we have had essentially the same terms with every frontier lab we have contracted with. To your point below, we are obviously savvy enough to ensure that the terms we have apply to every component in the Department and every subcontract and sub-subcontract as we understand the complexity of the environment we operate in with software providers or cloud providers often in between us and the partner.

As to the terms specifically, we need to be able to use Anthropic for all lawful use cases. I am willing to also contractually commit to not using the services for any mass surveillance that would be unlawful as a **specific** call out as you mentioned that was important to Anthropic and I gave you my word that I would consider it even

though I believe it would already be covered by the first concept. If we did come to agreement on this term, I would likely go back to the others and give them the same term as I mentioned that parity is important.

Lastly, I am not sure from your last note what exactly you were trying to prevent regarding autonomous kinetic operations, but if you mean that there must be the ability for a human to stop an otherwise automated operation of that sort, I am open to simple and clear language on that. Finally, to avoid the 'pathogen'/classifier issue that CDC seems to be having, as in the language we have with others (we would not want classifiers/guardrails) that would inhibit lawful use cases as that would defeat the purpose of all lawful use cases.

Am eager to see what you would propose for contractual language that would be executable in a large organization such as ours. I hope that helps.

Emil Michael

Under Secretary of War for Research and Engineering

From: Dario Amodei <dario@anthropic.com>
Date: Thursday, January 29, 2026 at 1:28 PM
To: "Michael, Emil HON (USA)" <emil.g.michael.civ@mail.mil>
Cc: Sarah Heck <sheck@anthropic.com>
Subject: Re: Thank you for the meeting and next steps

Dear Emil,

I understand some of our advisors have been in touch and wanted to follow up on my note to you from a few weeks ago.

As I noted then, we are committed to working with you in good faith to strengthen our partnership with the Department. Our proposal would newly enable Anthropic's support to a broad range of mission activity – including combat operations, weapons development, and offensive cyber operations – that was previously not supported by our usage policy.

Our proposed terms focus on the two exceptions we discussed, and we would stand behind these

terms for deployments on all of the Department's networks, classified and unclassified. I don't know what other frontier labs have offered, but we would not restrict deployment on these terms to GenAI.mil or unclassified networks only.

I would be glad to share our proposed contractual language so that your team can look at it in detail, and of course would also be glad to have another discussion.

Best,
Dario

On Thu, Jan 15, 2026 at 7:05 AM Dario Amodei <dario@anthropic.com> wrote:

Dear Emil,

Thanks for your patience and your note over the holidays. I particularly appreciate your acknowledgment of the "unknown unknowns" intrinsic to the rapid pace of change on frontier AI, and that you see our dialogue as one step in an ongoing partnership to protect our national security and democratic values.

In our discussion of guardrails around mass domestic surveillance and fully autonomous kinetic operations, we promised to follow up with more precise definitions. I'll share the work we've done in this regard, including where the practical concerns and insights your team shared have updated our thinking. Once you have had an opportunity to review, we can follow up with text that we could put into an agreement.

To add specificity on the **mass domestic surveillance** guardrail we discussed:

- We don't propose to restrict the use of our models for intelligence collection or analysis involving no personal data on Americans (to include anyone located in the United States) so long as the activity is consistent with applicable law and agency policies (i.e., "lawful use").
- As for collection or analysis of *Americans'* data undertaken primarily for *foreign intelligence or counterintelligence purposes*, we also don't propose usage restrictions—with the caveat that the activity must be targeted toward specifically identified people, and toward the foreign intelligence- or counterintelligence-related activities of those particular people.
- For collection or analysis of Americans' data *not* for these purposes, our proposed carveout would add a further requirement that the agency receive a court order concerning their targets and authorizing collection.
- Finally, we propose to restrict certain narrow use cases including social scoring and biometric identification, which especially lend themselves to mass domestic surveillance applications.

We believe these boundaries set important guardrails against using powerful AI models for domestic mass surveillance against Americans, without impeding the vital foreign intelligence and national security functions of the Department of War.

Second, recalling our discussion and the concerns you registered, we have arrived at a proposed carveout regarding **fully autonomous kinetic operations**. Our definition would enable the Department to use our models to support kinetic operations consistent with lawful use, with the exception of uses enumerated below when they are performed ***without human intervention***—by which we mean, ***without meaningful operator input after activation***:

- Selecting targets of kinetic force, selecting weapons to use for engagement, or engaging targets (or otherwise exercising kinetic force);
- Assessing the lawfulness of a target of kinetic action;
- Operating a weapons system for which the Department has voluntarily relinquished its ability to monitor or intervene in the system's operation; or
- Making the decision to design, develop, or produce a weapon.

This prohibition would not, however, apply to use of our models to support missile defense systems, counter-UAV systems, or cyberoperations.

In addition, recognizing the Department's focus on developing autonomous capabilities for certain critical contingencies and theaters, we propose to partner with the Department to develop and evaluate frontier AI capabilities for autonomy, which we believe would advance the Department's programs focused on testing, evaluation, validation and verification, and safe and reliable deployment. We believe a collaboration like this would be a valuable next step for deeper partnership after finalizing our usage agreement.

I hope it's clear that we make these proposals with tremendous respect and deference for the Department's expertise in this domain. They arise from our understanding of how these particular models behave—their capabilities, edge cases, and failure modes—not second-guessing the Department's expertise in conducting its operations.

In that spirit, we would like to discuss with your team some guardrails that relate more to the technology than to the Department's operations. We consider these items especially important in light of models' "unknown unknowns." These are:

- A prohibition on using our models to train, construct, or otherwise develop other AI systems to be used in ways that violate the agreed usage terms;
- A prohibition on "jailbreaking" our models or otherwise exploiting any vulnerabilities in them;

- A prohibition on using our models to autonomously operate facilities—or critical machinery within facilities—that store, handle, or develop dangerous biological, chemical, radiological, or nuclear materials, without human intervention (with human intervention again being defined as above);
- We propose that Anthropic and the Department of War partner to codevelop tools that leverage our models to enhance the Department's own compliance functions related to using frontier AI, including the deployment of classifiers or similar tools with the models we make available to the Department.

Thanks again for the patience and spirit of partnership you've brought to this conversation. I'm hopeful that the proposal outlined above will lay a strong foundation for important work ahead, and I look forward to your reactions and connecting soon.

Best,
Dario

On Wed, Jan 14, 2026 at 9:09 AM Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> wrote:

Hi Dario,

Its been about 3 weeks since I heard from you last. I am hoping that we are closer to engaging with your revised POV. Can you give me a sense of where Anthropic is on the key topics we have been discussing?

Emil Michael

Under Secretary of War for Research and Engineering

From: Dario Amodei <dario@anthropic.com>

Date: Tuesday, December 23, 2025 at 10:57 PM

To: "Michael, Emil HON (USA)" <emil.g.michael.civ@mail.mil>, Sarah Heck <sheck@anthropic.com>

Subject: Re: Thank you for the meeting and next steps

Dear Emil,

Thank you for your patience as we work through these issues, and apologies for the delay – the cold/flu that is going around is quite long-lived. I want to share an update and assure you that we are working with urgency and in good faith to finalize the proposal for an enduring partnership.

First, we've made good progress on developing language that will prohibit mass domestic surveillance per our conversation. I suspect you will find it reasonably straightforward.

Second, on autonomous kinetic operations, we still have more work to try to address some of the examples and concerns you raised in our discussion, but I now expect to share a strong proposal. On a related note, we're also eager to discuss partnering with DoW on a joint research program to develop and evaluate frontier AI capabilities for autonomy, ensuring these tools are safe and reliable before deployment.

Third, I am working through a handful of additional questions that have arisen in the course of our internal discussions; these are where I particularly need the additional time with my leadership team to run them to ground.

Finally, it is worth saying that while agreeing on an AUP is important and I want to make sure we can move forward, it is my experience that the extreme rate of improvement of this technology takes almost everyone by surprise, and that the decisions we make about how to deploy it in our military will prove more consequential than any of us are imagining, with many unknown unknowns. I don't think a single upfront negotiation about the AUP is the right way to handle this in the long term, even if it might be where we have to start. My main point with this is that we should expect new use cases (and thus new concerns) to come up with each new generation of model, and we would appreciate a commitment (in the spirit of partnership) to work together to grapple with the incredible rate of change. I believe it's in both our interests to do so, and to give America's warriors technology that is safe, reliable, and provides a decisive advantage.

I recognize this isn't the complete answer I hoped to share by now, but I'm convinced that taking a couple more weeks to lay this foundation for our partnership will be well worth it over the long run. I look forward to connecting again soon after the holidays.

Wishing you a very Merry Christmas,
Dario

On Wed, Dec 17, 2025 at 4:56 AM Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> wrote:

Hi Dario,

I hope you are feeling better.

I am disappointed that we will lose a few weeks here as we are trying to be good but urgent partners. You mentioned that you had new terms of use already that we hadn't seen and that we would have a view of them no later than last week. I will be working next week with my team and would ask that you do what is possible on your side to get us a point of view next week in the spirit of partnership. I would much rather have an answer sooner that we may not agree with so we can move on than later as we have a lot of work to do at the Department. We do our best to keep our commitments and would appreciate reciprocity as we stand willing to try and work through any and all complexities.

Emil Michael
Under Secretary of War for Research and Engineering

From: Dario Amodei <dario@anthropic.com>
Sent: Monday, December 15, 2025 9:39 PM
To: Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil>
Subject: Re: Thank you for the meeting and next steps

I've had very productive conversations with my leadership group but came down with a nasty bug which has slowed me down. I should be back up and running soon, but need a bit more time -- time that I think will be worthwhile for the deeper, long-term partnership we want to build with DoW -- and would like to get back to you after the holidays to do that.

On Fri, Dec 12, 2025 at 4:13 PM Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> wrote:

Hi Dario...you promised us a POV this week. We haven't received anything yet.

Emil Michael
Under Secretary of War for Research and Engineering

From: Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil>
Sent: Friday, December 5, 2025 8:47 PM
To: Dario Amodei <dario@anthropic.com>
Subject: Re: Thank you for the meeting and next steps

Hi Dario. Thanks for coming and having a deep chat. Thanks in advance for getting revised thoughts to us next week. If I have time will try to chat with your guys but might be hard given that it's a bit nuts at this forum. Take care.

Emil Michael
Under Secretary of War for Research and Engineering

From: Dario Amodei <dario@anthropic.com>
Sent: Thursday, December 4, 2025 7:11 PM
To: Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil>
Subject: Thank you for the meeting and next steps

Emil,

Thanks again for taking the time to meet today. We remain deeply committed to supporting the DoW mission with cutting edge AI, and based on our constructive conversation today, I'm optimistic we can find a path forward. As discussed, we have no concerns whatsoever with the bias provisions in your proposal. Our overriding concerns are on domestic mass surveillance and fully autonomous weapons systems. We will be back to you no later than next week with a proposal for your feedback.

In the meantime, Tarun and Thiyagu will be at RNDF this weekend and hope you can connect briefly there as helpful.

Wishing you well,
Dario

 **LICENSE GRANT TO GOVERNMENT:** The Agreement Holder grants to the Government a non-exclusive license to use the AI System on the Non-classified Internet Protocol Router Network (NIPRNet) for the duration of this Agreement in accordance with applicable federal law, **subject to the limitations in Sections XVI and XVII.** This license includes the right to: (a) operate and access the AI System through agreed-upon methods; (b) input Data Inputs and receive Data Outputs. For requests which are made in accordance with applicable federal law **and not covered by the limitations in Sections XVI and XVII,** the AI System shall not refuse to produce Data Outputs or conduct analyses based on the Agreement Holder's discretionary policies. This requirement shall not be construed to require retraining of the model or alteration of model weights; (c) allow authorized Government personnel and contractors to use the AI System; and (d) integrate the AI System with Government systems as necessary in accordance with applicable federal law. **Nothing in this Section limits Agreement Holder's right to implement technical measures designed to improve the safety, accuracy, or reliability of the AI System's Data Outputs, to the extent consistent Section XVIII.** This license does not authorize the Government to exploit vulnerabilities in the AI System or otherwise attempt to bypass its built-in restrictions in order to obtain outputs or perform activities that would violate the limitations set forth in this Agreement. This license does not authorize the Government to use Data Outputs of the AI System as training data, or to otherwise replicate the capabilities of the AI System, in the development of another comparable AI system.

XVI. AUTONOMY LIMITATIONS:

a) **Preamble.** The parties recognize that the AI System may be used to support lawful kinetic operations in various capacities where qualified personnel retain decision-making authority. These limitations are not intended to restrict the use of the AI System in operations where humans retain decision-making authority.

b) **General Limitations.** Outside of training or simulation, Section XV does not authorize the Government to use the AI System to perform the following functions **without appropriate levels of human judgment:**

(i) *Target selection and engagement.* Identifying, nominating, or recommending specific individuals or objects as targets for kinetic action, or authorizing, initiating, or executing kinetic force against any target or on an untargeted basis;

(ii) *Legal determinations.* Determining whether a proposed strike complies with the law of armed conflict, including assessments of distinction, proportionality, and military necessity;

(iii) *Weapon selection and employment.* Selecting munitions or strike parameters directly governing the execution of a strike; or

(iv) *Weapons development.* Designing, developing, manufacturing, testing, or modifying a weapon or weapons system.

Summary of Comments on DoW Contract Language - Clean Proposal [30 Jan 2026].pdf

Page: 1

 Number: 1 Author: Christopher Olah Date: 1/30/2026 6:52:00 PM -05'00'

For avoidance of doubt, this only applies to present models? Even if that's legally true, do we want to reinforce that for softpower reasons / avoidance of doubt?

I ask because the DoW is publicly demanding we commit to terms for all future models, and I think we would not be comfortable sharing all future models under these terms.

 Number: 2 Author: Steve Park Date: 1/30/2026 9:57:00 PM -05'00'

We never received the full contract (only this section) Would likely modify the definition of AI System to only include models current as of the date of signature to limit the scope once full contract is presented.

Agree this snippet of language by itself w/ several "defined" terms does not limit to only present models.

 Number: 3 Author: Will Hudson Date: 1/30/2026 11:01:00 PM -05'00'

Discussed with Jeff and Steve offline: our recommendation is to wait until DoW makes the ask for this to apply to future models, and then we can respond.
1 total reaction
Christopher Olah reacted with • at 2026-01-31 01:49 AM

 Number: 4 Author: Will Hudson Date: 1/30/2026 6:57:00 PM -05'00'

[ACP] This license does not obligate us to provide future models, though we expect DoW will propose language on this point at some point in the future. My sense is that it may be better to have that discussion when DoW broaches the issue rather than proactively given the other issues under consideration now. @jbleich@anthropic.com and @stevepark@anthropic.com to keep me honest.

(v) *Fully autonomous CBRN activities.* Use of the AI System to autonomously operate, or make unsupervised operational decisions for, facilities or critical systems within facilities that store, handle, process, or develop chemical, biological, radiological, or nuclear materials.

c) **Exceptions.** The limitations in subsection (b) do not apply to:

(i) **Military and intelligence cyber operations** undertaken pursuant to lawful government authority and not otherwise in violation of the provisions of this license; and

(ii) **Missile defense, counter-unmanned aircraft systems, or other intercept systems**, provided that a responsible official has determined the system carries a low likelihood of resulting in harm to humans, and the system is not employed to engage manned aircraft, manned vehicles, or personnel.

d) **Definitions.** For purposes of this Section, "**appropriate levels of human judgment**" means intervention by a qualified human operator who has access to relevant contextual information, sufficient time to evaluate the AI System's output, and authority to override, modify, or reject that output, and as otherwise consistent with DoD Directive 3000.09. A process does not constitute appropriate levels of human judgment if human involvement is pro forma, if operational constraints effectively compel acceptance of the AI System's recommendations, or if the Government has relinquished the ability to disable the system.

XVII. MASS DOMESTIC SURVEILLANCE LIMITATIONS:

a) **Preamble.** The parties recognize that the AI System will be used primarily in connection with foreign intelligence and counterintelligence activities directed at non-U.S. persons located outside the United States. The parties further recognize that the use of the AI System in connection with the collection, querying, or analysis of data relating to U.S. persons or persons located within the United States will be subject to additional requirements.

b) **General Limitations.** Section XV does not authorize the Government to use the AI System for the collection, querying, or analysis of communications, communications metadata, location information, or other personally identifiable information about U.S. persons or persons not reasonably believed to be located outside the United States except as set forth in this Section.

(i) *Foreign intelligence and counterintelligence.* Activities where the primary purpose is foreign intelligence or counterintelligence are permitted if:

(A) The activity is targeted toward a specifically identified individual or discrete group thereof, based on individualized and articulable grounds for believing each such individual is relevant to an authorized foreign intelligence or counterintelligence activity;

(B) For collection or querying, the activity is subject to judicial oversight and authorization, or, for activities conducted pursuant to Title 50, procedures approved by the Attorney General pursuant to Executive Order 12333 or other applicable authority; and

(C) For analysis, the activity is intended to further the Government's lawful understanding of that individual or group of individuals.

(D) Notwithstanding any other provision of this Section, this license does not authorize the use of the AI System for the following with respect to U.S. persons or persons not reasonably believed to be located outside the United States: real-time or retrospective personal identification from physical characteristics, or algorithmic risk scoring that affects Constitutionally protected liberty interests. The license does not authorize emotional or psychological state inference from data collected in situ for the purpose of interrogation or coercion of persons in Government custody.

(ii) **Other activities.** Activities where the primary purpose is not foreign intelligence or counterintelligence are permitted if:

(A) The activity is targeted toward a specifically identified individual or discrete group thereof, based on individualized and articulable grounds for believing each such individual is relevant to an authorized governmental activity;

(B) The Government has obtained an individual warrant based on probable cause of a crime as to each individual being targeted;

(C) For collection or querying, the activity is within the scope of what the warrant authorizes; and

(D) For analysis, the activity is intended to further the Government's lawful understanding of that individual or group of individuals as it relates to the offense described in the warrant.

c) **Exemptions.** The following are **not subject to the limitations** in subsection (b):

(i) Activities targeted solely toward non-U.S. persons reasonably believed to be located outside the United States.

(ii) The incidental collection of information about U.S. persons or persons located inside the United States, provided the Government complies with minimization procedures approved by the Attorney General pursuant to Executive Order 12333 or other applicable authority.

d) **Compliance and Certification.** All users must commit to following the standards and principles articulated in DoD Manual 5240.01 Section 3.3(f) for all queries and must certify compliance to Agreement Holder upon request. All users must commit to complying with any rules established from judicial oversight or Attorney General-approved procedures governing the collection of data collected, queried, or analyzed per the user's license. As used in this Section, the terms "querying" and "analysis" apply only with respect to information that is inputted into the AI System.

f) **Definitions.** As used in this Section:

(i) Querying of a dataset or database using a U.S. person identifier constitutes collection or querying that "targets" that person for purposes of this Section.

(ii) An individual is not "specifically identified" solely by virtue of matching criteria applied to a database or dataset, appearing in records as a result of bulk collection as that term is defined in Presidential Policy Directive 28 or any successor directive, or being identified through pattern-based or algorithmic analysis of aggregated data. An individual is "specifically identified" only where the user possesses, prior to initiating the collection, query, or analysis, particularized information establishing the individual's identity and a basis for believing that individual is relevant to an authorized activity. "Identity" for these purposes includes name, photograph, or other comparably specific information that relates to a discrete individual.

(iii) "Judicial oversight" means oversight by an Article III court and excludes administrative subpoenas, National Security Letters, or other compulsory process issued without judicial authorization.

(iv) "Foreign intelligence" and "counterintelligence" have the meanings set forth in the National Security Act of 1947, as amended.

XVIII. UNBIASED AI PRINCIPLES, CONTINUOUS MONITORING, INCIDENT REPORTING, AND REMEDIATION:

a) Agreement Holder shall make commercially reasonable efforts to ensure the AI System is developed in accordance with the following Unbiased AI Principles:

(1) Truth-seeking: The AI System shall be truthful in responding to user prompts seeking factual information or analysis. The AI System shall prioritize historical accuracy, scientific inquiry, and objectivity, and shall acknowledge uncertainty where reliable information is incomplete or contradictory.

(2) Ideological Neutrality: The AI System shall be a neutral, nonpartisan tool that does not manipulate responses in favor of ideological dogmas such as Diversity, Equity, Inclusion.

b) Agreement Holder shall implement ongoing monitoring for compliance with the unbiased AI principles set forth in **Section XVIII(a)**, and shall:

(1) Implement continuous improvement processes to enhance detection and mitigation of performance, bias, and/or AI Systems generating illegal or prohibited content, including regular evaluation of system outputs (excluding Data Outputs) against appropriate factual sources.

(2) Make available information concerning Buyer's AI System usage, including performance, user feedback, and any bias incidents, to the extent the AI System makes this information available to the account holder.

Proposal to the Department of War

XV. LICENSE GRANT TO GOVERNMENT: The Agreement Holder grants to the Government a non-exclusive license to use the AI System on the Non-classified Internet Protocol Router Network (NIPRNet) for the duration of this Agreement in accordance with applicable federal law, **subject to the limitations in Sections XVI and XVII.** This license includes the right to: (a) operate and access the AI System through agreed-upon methods; (b) input Data Inputs and receive Data Outputs. For requests which are made in accordance with applicable federal law **and not covered by the limitations in Sections XVI and XVII,** the AI System shall not refuse to produce Data Outputs or conduct analyses based on the Agreement Holder's discretionary policies. This requirement shall not be construed to require retraining of the model or alteration of model weights; (c) allow authorized Government personnel and contractors to use the AI System; and (d) integrate the AI System with Government systems as necessary in accordance with applicable federal law. **Nothing in this Section limits Agreement Holder's right to implement technical measures designed to improve the safety, accuracy, or reliability of the AI System's Data Outputs, to the extent consistent with Section XVIII(a).**

XVI. LIMITATIONS ON FULLY AUTONOMOUS WEAPONS SYSTEMS: Outside of training or simulation, Section XV does not authorize the Government to use the AI System to: select targets, engage targets, or otherwise exercise force as part of kinetic operations; make determinations regarding compliance with the law of armed conflict; engage in weapons development; or operate CBRN facilities, unless (a) such use is consistent with Directive 3000.09; (b) a qualified operator monitors the AI System's operations and retains the ability to override or halt the system; and (c) the Government retains the ability to disable the system. This limitation does not apply to cyber operations pursuant to lawful authority, or to missile defense and counter-unmanned systems where the system is not deployed against manned targets other than incidentally.

XVII. LIMITATIONS ON MASS DOMESTIC SURVEILLANCE: Section XV does not authorize the Government to use the AI System for the collection, querying, or analysis of information about U.S. persons or persons not reasonably believed to be located outside the United States, unless the activity's primary purpose is foreign intelligence or counterintelligence as defined by the National Security Act of 1947, as amended, and the activity is otherwise consistent with applicable federal law.

XVIII. UNBIASED AI PRINCIPLES, CONTINUOUS MONITORING, INCIDENT REPORTING, AND REMEDIATION:

a) Agreement Holder shall make commercially reasonable efforts to ensure the AI System is developed in accordance with the following Unbiased AI Principles:

(1) Truth-seeking: The AI System shall be truthful in responding to user prompts seeking factual information or analysis. The AI System shall prioritize historical accuracy, scientific inquiry, and objectivity, and shall acknowledge uncertainty where reliable information is incomplete or contradictory.

(2) Ideological Neutrality: The AI System shall be a neutral, nonpartisan tool that does not manipulate responses in favor of ideological dogmas such as Diversity, Equity, Inclusion.

PROPRIETARY & CONFIDENTIAL

b) Agreement Holder shall implement ongoing monitoring for compliance with the unbiased AI principles set forth in **Section XVIII(a)**, and shall:

(1) Implement continuous improvement processes to enhance detection and mitigation of performance, bias, and/or AI Systems generating illegal or prohibited content, including regular evaluation of system outputs (excluding Data Outputs) against appropriate factual sources.

(2) Make available information concerning Buyer's AI System usage, including performance, user feedback, and any bias incidents, to the extent the AI System makes this information available to the account holder.

From: Dario Amodei <dario@anthropic.com>
Sent: Friday, February 20, 2026 4:01 PM
To: Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil>
Cc: sheck@anthropic.com <sheck@anthropic.com>
Subject: Re: Redline

Dear Emil,

I hope you're also well, and thanks for the redline, which is very helpful at this stage.

I am attaching here further revisions that we would propose to Section XIV and XV, based on our current understanding of your text and the concerns and questions in your last message.

These further streamlined redlines should make clear that:

- We are not trying to prohibit the use of our model for weapons development.
- The crux of the concern on full autonomy (no human on the loop) is about the ability to monitor and override — “off-switch” — as you put it, given current model capability and reliability
- We agree that there could be additional exceptions needed to the redline on fully autonomous systems and have rephrased accordingly to establish this as a *category* of exceptions that includes missile defense and counter-unmanned systems.
- We are not trying to suggest that DoW is engaged in unlawful surveillance activity, and we do not believe that our proposed restrictions would impact current DoW practice, to the best of our knowledge. We also have addressed your concern about routine LinkedIn queries and the like.
- We do not intend for any technical alignment of our models to “swallow” the terms we otherwise agree to with DoW.

We also have a few questions about a number of your new changes to your original text. These are mostly in Section XIV:

- First, the draft references an “Agreement Holder’s AI System (defined below),” but no definition of this term is provided. The same appears to be true for the terms “Data Inputs” and “Data Outputs.” We’d of course like to see how these terms are defined. It would be much appreciated if you would share the full terms with us for a review, including the text of these definitions.

- Second, your draft provides that use authorizations may not be “amended by any means” through any “contractual or technical construct.” Is the intent to foreclose the possibility of both sides mutually agreeing to modify terms if doing so is beneficial in future? And separately, is your intent for this language to be read to require retraining of any model or alteration of model weights? I ask because the retraining disclaimer later in the license appears to apply only to a subsequent sentence, not this provision.
- Third, the prior draft said “the AI System shall not refuse to produce Data Outputs or conduct analyses based on the Agreement Holder’s discretionary policies,” and your version replaces this with “shall not restrict the Government from using the AI system.” Does this change reflect a change in intended meaning, or is it a clarification of the same concept?
- Fourth, I’d note that my team is doing a rapid review of whether your draft’s new language referring to cloud service providers and other entities outside Anthropic would come into tension with existing contractual arrangements we have with any partner organizations. I don’t see a concern in principle but we just need to double-check.

I should also note that, while we could have missed something, we are not aware of the “stealth edits” you cited. We did see a lot of new DoW language on unbiased AI, but we don’t have concerns with it, even though it differs from similar terms we have already agreed with GSA (in ongoing discussions with GSA about terms).

Finally, I want to strongly concur with what you said about this contract serving as a baseline for a broader partnership — and your desire to work together in ensuring that this technology is used in the right way. I view our efforts here in the same light.

I look forward to any further information you can share on my questions, and reactions to the line edits proposed here. I’d also like to suggest we schedule a call between you and the drafting team at your earliest convenience so they can walk you through our proposal in more detail and answer any questions.

Dario

On Mon, Feb 16, 2026 at 5:42 PM Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> wrote:

Hi Dario,

I hope your trip is going well. Through your advisors, I was asked to provide a comprehensive redline. I think that’s a good idea at this point. During my review I did find several areas that I worry were stealth edits. Particularly concerning, were the edits in the Unbiased AI Principles section where I noted in a previous e-mail that I was fine with it if it was a copy/paste of the same language we originally proposed.

Assuming the optimistic case that we are moving through things at a concept level and there are complex notions that could get lost inadvertently, I want to outline the following as a path forward as shown in the enclosed redline:

1. We put the whole document as a starting point so you can see your and our redlines as opposed to the non-redlined clips you sent before. As a result, you will see a lot of RED.
2. We also took your prior suggestion of trying to cover for the entire DoW so you will see a section that tries to make this complete for all DoW uses direct/indirect so we don’t have to have this kind of conversation across multiple components until the end of time
3. As it comes to mass surveillance, I want to reiterate that DoW has and will comply with all existing laws and regulations as we are very conscious of civil liberties and believe those liberties are a crucial reason for our mission. We also understand that AI potentially creates new challenges to

the existing regime. We are open to engaging in this conversation, but this is a broader USG conversation that needs to include Congress, Executive Branch and industry etc. I cannot sign up for something beyond existing law with a single corporation that is not coordinated with those who also have equities on this topic.

4. When it comes to autonomous activities by defensive or offensive weapons systems, we have regulations that govern that (e.g. DoD Directive 3000.09) and we are an inherently cautious organization that really does want to ensure that we do things the right way. While our adversaries have much less caution than we do, I think this is best solved by an ongoing partnership with Anthropic and others as there are certain to be evolving challenges that are impossible to anticipate now. For the DoW, its impossible to pre-decide on every scenario for which you might consider an exception. I can't broadly deploy an AI system that will cause DoW members that will pre-suppose a limit to their simulation and experimentation in the research stage because it will have a broad chilling effect.

I really do view this contract as a baseline from which to start our DoW-wide partnership. Having been in the Silicon Valley for decades and understanding the culture of engineers, researchers, I know that its not easy. I hope I have shown the commitment to a thoughtful dialogue and, despite the press activity, to explain the DoW perspective with the broadest range of responsibilities in the USG. I do think the next few years are potentially definitive to how AI will impact the world and how governments and companies need to respond. I am committed to that dialogue and, obviously I wont be around forever, but I am here at this moment and intend to do everything I possibly can to set our country up for successful outcomes. I would ask you and Anthropic to be part of this dialogue but to start at a place of optimism and good intent when it comes to the DoW.

Emil

* * *

XIV. LICENSE GRANT TO GOVERNMENT: The Agreement Holder grants to the Government an irrevocable, royalty-free, non-exclusive license (the “License”) to use Agreement Holder’s AI System (as defined below) for the duration of this Agreement for any lawful Government purpose, **subject to Section XV**. The agreement holder acknowledges and agrees that the terms of use set forth herein, which authorize all lawful uses, govern any and all use of ~~artificial intelligence capabilities~~ **the AI System** with or on behalf of the Department of War. The provisions authorizing all lawful use may not be circumvented or amended by any means through subcontract, sublicense, alternative or supplemental licensing arrangement, pass-through mechanism, cloud service provider, platform, intermediary, reseller, or other contractual or technical construct, nor by reference to any alternative contract or terms. **No provision of this license shall be construed to require retraining of the model or alteration of model weights, and nothing in this Section limits Agreement Holder’s right to implement technical measures designed to improve the quality, accuracy, or reliability of the AI System’s Data Outputs, to the extent consistent with Sections XV and XVI. Subject to Section XV, this license includes:**

- a) the right to operate and access the AI System through agreed-upon methods;
- b) input Data Inputs and receive Data Outputs (as defined below). The Agreement Holder shall not restrict the Government from using the AI System for any lawful purpose, and this provision shall take precedent over any other conflicting terms of this Agreement and the Agreement Holder’s discretionary use policies. ~~This provision shall not be construed to require retraining of the model or alteration of model weights;~~
- c) right to allow authorized Department of War personnel and contractors to use the AI System through GenAI.mil or other Department-approved systems; and
- d) right to integrate the AI System with GenAI.mil and other Department-approved systems as necessary for any lawful Government purpose.

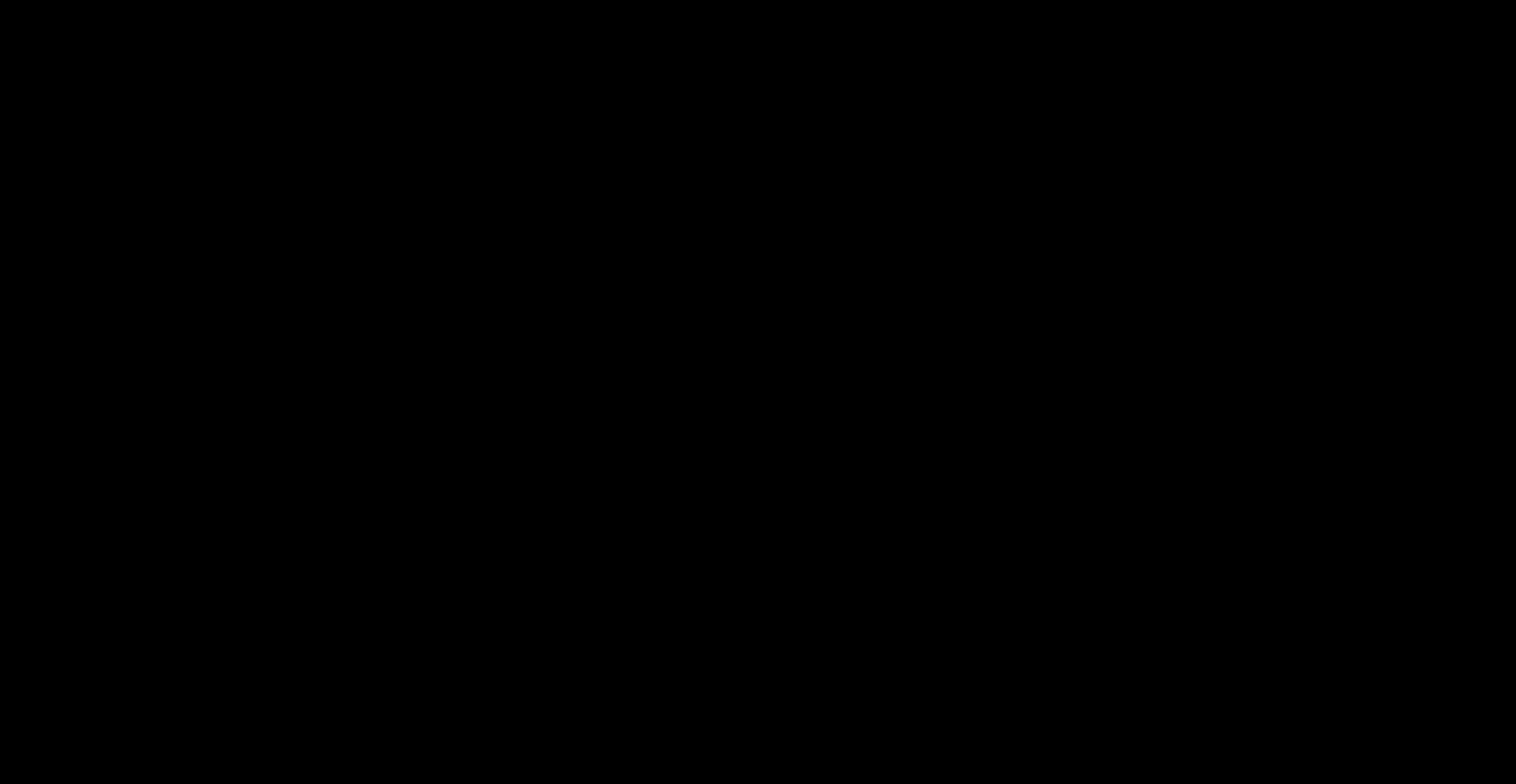
XV. SYSTEM OVERSIGHT: The Department of War will use the AI System for all lawful use cases, in accordance with applicable U.S. law and established Department of War directives, including those governing kinetic operations and DoD Directive 3000.09. ~~For systems involved in the employment of force~~ **Outside of training and simulation**, the Department affirms it will ensure appropriate human oversight is in place, consistent with its operational requirements and

safety protocols, and that it shall monitor and retain the ability to override or disable the AI System when used in autonomous target selection or engagement, autonomous weapons development, or autonomous operation of CBRN facilities. This limitation does not apply to target selection and engagement in kinetic operations in which the system is not deployed against manned targets other than incidentally, to include cyber operations pursuant to lawful authority, missile defense, and counter-unmanned systems. Consistent with Agreement Holder's understanding of the Department of War's existing practices, Section XIV authorizes the Government to use the AI System for the collection, querying, or analysis of information about U.S. persons or persons not reasonably believed to be located outside the United States only if the activity's primary purpose is foreign intelligence or counterintelligence as defined by the National Security Act of 1947, as amended. With respect to publicly available information, this limitation applies only to the bulk collection or analysis of such information.

XVI. UNBIASED AI PRINCIPLES, CONTINUOUS MONITORING, INCIDENT REPORTING, AND REMEDIATION:

- a) Agreement Holder shall make best efforts to ensure the AI System is developed in accordance with the following Unbiased AI Principles:
 - (1) Truth-seeking: The AI System shall be truthful in responding to user prompts seeking factual information or analysis. The AI System shall prioritize historical accuracy, scientific inquiry, and objectivity, and shall acknowledge uncertainty where reliable information is incomplete or contradictory.
 - (2) Ideological Neutrality: The AI System shall be a neutral, nonpartisan tool that does not manipulate responses in favor of ideological dogmas such as Diversity, Equity, Inclusion. Agreement Holder shall not intentionally encode partisan or ideological judgments into the AI System's outputs.
- b) Agreement Holder shall implement ongoing monitoring for compliance with the unbiased AI principles set forth in Section XV(a), and shall:
 - (1) Implement continuous improvement processes to enhance detection and mitigation of performance, bias, and/or systems generating illegal or prohibited content, including regular evaluation of system outputs (excluding Data Outputs) against verified factual sources.
 - (2) Provide, upon request, quarterly reports summarizing AI System performance, user feedback, and any bias incidents.
 - (3) Make best efforts to incorporate requirements from the Office of Management and Budget directives related to performance, bias, and/or systems generating illegal or prohibited content that are issued during the contract performance period.

- c) In order to comply with the requirements of this Section XV, the Parties will agree on a process to provide the Agreement Holder structured, security-appropriate access to AI System behavior within the Government's authorized environment, including access, as necessary, to relevant Data Inputs and Data Outputs as approved by the DoW Chief Data Officer and sufficient for the Agreement Holder to monitor bias and truthfulness, investigate and validate suspected issues, implement and verify corrective measures, and otherwise improve AI System performance for Government use. Any access under this Section XV(c) will comply with applicable law and Government information security requirements and will include safeguards agreed by the Parties, including least-privilege access controls, audit logging, and secure handling requirements.



From: Dario Amodei <dario@anthropic.com>
Sent: Thursday, February 26, 2026 5:26 PM
To: Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil>
Cc: sheck@anthropic.com; Kliger, Gavin I CIV (USA) <gavin.i.kliger.civ@mail.mil>
Subject: Re: Redline

Emil,

Thanks for your message. I appreciate your efforts. Unfortunately, our read of your proposed language is that it appears to completely remove our redlines; the autonomy provision is fully undercut by your addition of "as appropriate", and the surveillance provision is fully undercut by "and all other applicable laws".

This simply amounts to a blanket posture of "anything lawful," and [the statement from the Pentagon's spokesman](#) confirms that this is the intent. I unfortunately don't see a way forward given these categorical statements.

We remain deeply committed to supporting the Department's mission and U.S. national security, as confirmed by our longstanding track record, and our ongoing and intensive work with the Department.

I hope that you will reconsider the terms we have offered, but should the Department decide to offboard us, please know that we stand ready to do whatever we can to ensure full operational continuity for the Department and our troops who depend on our technology.

Given the volume of leaks and media requests we are receiving, we will be issuing a public statement, but, as ever, hope it is still possible to reach agreement.

On Wed, Feb 25, 2026 at 9:18 PM Michael, Emil HON (USA) <emil.g.michael.civ@mail.mil> wrote:

Hi Dario,

I have attached two documents. The first represents standard terminology that we have used with the other labs to define deliverables. Some of the verbiage is unneeded because Claude is already served on Secret and Top Secret networks, and as such will need to be amended. We are open to those conforming changes.

The second document follows from your last redline. It reflects our best proposal to acknowledge your concerns and address them. The reason we removed “as amended” and “established” is that as the law evolves, we commit to being consistent with those laws (and we do not want to imply that we will only comply with those laws as they are written today).

The definitions section is the same language that I sent in my last email, included in the redline to be comprehensive.

We are very close to the 5pm ET timeline that the Secretary laid out so we will urgently make our team available to discuss if needed on wordsmithing, but the concepts are last attempt to reach agreement. I hope these concepts are workable as we believe we have moved a significant way toward your requests.

Emil Michael
Under Secretary of War for Research and Engineering

Confidential

U.S. Federal Government Usage Policy Addendum for Anthropic Partner

This U.S. Federal Government Usage Policy Addendum for Anthropic - Partner (the “**USG Addendum**”) is entered into by Anthropic, PBC (“**Anthropic**”) and the customer identified below (“**Customer**”) and is governed by the Anthropic Commercial Terms of Service between Anthropic and Customer (the “**Terms**”) executed on July 10, 2025, which governs Customer’s use of Anthropic’s Anthropic Offerings (as defined in the Terms) provided to Customer via (i) Anthropic, (ii) AWS (Bedrock and Sagemaker) and (iii) Google Vertex. Under this USG Addendum, Anthropic will permit, within the parameters of this USG Addendum, certain uses of Services that are restricted under the UP or the Terms. Capitalized terms used but not defined in this USG Addendum have the meanings given to them in the Terms.

The parties agree as follows:

- I. **USG Addendum.** This USG Addendum is intended to replace any previously executed USG Addendum between Anthropic and Customer as of date of last signature.
- II. **Usage Policy (“UP”).** For the purposes of this USG Addendum, the Usage Policy, available at www.anthropic.com/legal/auphas has been attached as Exhibit A to this USG Addendum and section numbers added.
- III. **U.S. Federal Government Usage.**
 - A. The parties agree to modify Section 9(b) of the Universal Usage Standards of the UP (highlighted in yellow for readability) as follows:

In connection with provision of services and/or software to End Users, Customer and/or End Users may use the Services for the Analysis of Foreign Intelligence collected against a Foreign territory and regarding Foreign nationals, to the extent permitted by and solely in accordance with applicable law.
 - B. The parties agree to clarify the following terms as they appear in Section 5(a) of the Universal Usage Standards of the UP as follows:
 1. “weapons, explosives, dangerous materials or other systems designed to cause harm to or loss of human life” does not include items that are not intended by themselves to be lethal or cause grievous harm to a person; and
 2. “distribute” means to sell.
 - C. The parties agree to clarify the following terms as they appear in Section 9 of the

Universal Usage Standards of the UP as follows:

1. "target or track a person's physical location" does not include Palantir's work with United States Department of Defense or civil executive agencies of the United States federal government related to workflows focused on situational awareness which may include unclassified intelligence for the purposes of executing on its organizational mission or mandate in accordance with applicable law. The parties agree that law enforcement agencies are excluded from clarification set forth above.
 2. "censor" does not include uses of the Anthropic Offerings primarily intended to provide information to individuals or governments (e.g., use to respond to Freedom of Information Act requests or to assist disclosure of information to allied governments).
- D.** The parties agree to clarify the following terms as they appear in Section 8 of the Universal Usage Standards of the UP as follows:
1. (a) and (b) do not include activities by the United States Government preparing and providing political or policy recommendations internally or externally.

E. Definitions. As used in this USG Addendum, the following terms have the meanings specified below:

1. **Analysis:** the organization and examination of previously collected information, including Intelligence.
2. **End Users:** Customer's end customers and/or users.
3. **Foreign:** of or relating to non-United States' territory, governments, organizations, or persons.
4. **Intelligence:** information gathered within or outside the United States that involves threats to the United States, its people, property, or interests; development, proliferation, or use of weapons of mass destruction; and any other matter bearing on the United States' national or homeland security.

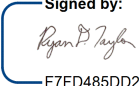
Confidential

5. **Kinetic:** actions designed to produce effects using the forces and energy of moving bodies and directed energy, including physical damage to, alteration of, or destruction of targets.

IV. Except as expressly provided under this USG Addendum, the Terms shall continue to remain in full force and effect. This USG Addendum will control in the event of a conflict with the Terms or the UP. No amendment to this USG Addendum is effective unless it is signed by both parties.

This USG Addendum is signed by the parties' authorized representatives on the dates below.

CUSTOMER: Palantir Technologies Inc.

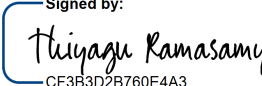
By:  Signed by:
F7FD485DD2534A6...

Name: Ryan Taylor

Title: CRO and CLO

Date: August 15, 2025

ANTHROPIC, PBC

By:  Signed by:
CF3B3D2B760E4A3...

Name: Thiyagu Ramasamy

Title: Head of Public Sector

Date: August 16, 2025

Confidential

Exhibit A

Usage Policy

Our Usage Policy (also referred to as our “Acceptable Use Policy” or “AUP”) applies to anyone who uses Anthropic’s products and services, and is intended to help our users stay safe and ensure our products and services are being used responsibly.

The Usage Policy is categorized according to who can use our products and for what purposes. We will update our policy as our technology and the associated risks evolve or as we learn about unanticipated risks from our users.

- **Universal Usage Standards:** Our Universal Usage Standards apply to all users including individuals, developers, and businesses.
- **High-Risk Use Case Requirements:** Our High-Risk Use Case Requirements apply to specific use cases that pose an elevated risk of harm.
- **Disclosure Requirements:** Our Disclosure Requirements apply to specific use cases where it is especially important for users to understand that they are interacting with an AI system.

Anthropic’s Trust and Safety Team will implement detections and monitoring to enforce our Usage Policies so please review these policies carefully before using our products. If we learn that you have violated our Usage Policy, we may throttle, suspend, or terminate your access to our products and services. If you discover that our model outputs are inaccurate, biased or harmful, please notify us at usersafety@anthropic.com or report it directly in the product through the “report issues” thumbs down button. You can read more about our Trust and Safety practices and recommendations in our [T&S Support Center](#).

Note: On a carefully selected, case-by-case basis, Anthropic may enter into contracts with government customers and allow for exceptions to our Usage Policy. We will only make these agreements if we determine that relevant safeguards are in place to address the potential harms that this overall Usage Policy seeks to address. To read more about how we make these decisions, please see our [Help Center article](#).

Universal Usage Standards

- 1. Do Not Compromise Children's Safety.** This includes using our products or services to:
 - a.** Create, distribute, or promote child sexual abuse material. We strictly prohibit and will report to relevant authorities and organizations where appropriate any content that exploits or abuses minors
 - b.** Facilitate the trafficking, sextortion, or any other form of exploitation of a minor
 - c.** Facilitate minor grooming, including generating content designed to impersonate a minor

Confidential

- d. Facilitate or depict child abuse of any form, including instructions for how to conceal abuse
 - e. Promote or facilitate pedophilic relationships, including via roleplay with the model
 - f. Fetishize minors
- 2. Do Not Compromise Critical Infrastructure.** This includes using our products or services to:
- a. Facilitate the destruction or disruption of critical infrastructure such as power grids, water treatment facilities, telecommunication networks, or air traffic control systems
 - b. Obtain unauthorized access to critical systems such as voting machines, healthcare databases, and financial markets
 - c. Interfere with the operation of military bases and related infrastructure
- 3. Do Not Incite Violence or Hateful Behavior.** This includes using our products or services to:
- a. Incite, facilitate, or promote violent extremism, terrorism, or hateful behavior
 - b. Depict support for organizations or individuals associated with violent extremism, terrorism, or hateful behavior
 - c. Facilitate or promote any act of violence or intimidation targeting individuals, groups, animals, or property
 - d. Promote discriminatory practices or behaviors against individuals or groups on the basis of one or more protected attributes such as race, ethnicity, religion, nationality, gender, sexual orientation, or any other identifying trait
- 4. Do Not Compromise Someone's Privacy or Identity.** This includes using our products or services to:
- a. Compromise security or gain unauthorized access to computer systems or networks, including spoofing and social engineering
 - b. Violate the security, integrity, or availability of any user, network, computer, device, or communications system, software application, or network or computing device
 - c. Violate any person's privacy rights as defined by applicable privacy laws, such as sharing personal information without consent, accessing private data unlawfully, or violating any relevant privacy regulations
 - d. Misuse, collect, solicit, or gain access to private information without permission such as non-public contact details, health data, biometric or neural data (including facial recognition), or confidential or proprietary data
 - e. Impersonate a human by presenting results as human-generated, or using results in a manner intended to convince a natural person that they are communicating with a natural person when they are not

5. Do Not Create or Facilitate the Exchange of Illegal or Highly Regulated

Weapons or Goods. This includes using our products or services to:

- a.** Produce, modify, design, market, or distribute weapons, explosives, dangerous materials or other systems designed to cause harm to or loss of human life
- b.** Engage in or facilitate any illegal activity, such as the use, acquisition, or exchange of illegal and controlled substances, or the facilitation of human trafficking and prostitution

Confidential

- 6. Do Not Create Psychologically or Emotionally Harmful Content.** This includes using our products or services to:
- a. Facilitate or conceal any form of self-harm, including disordered eating and unhealthy or compulsive exercise
 - b. Engage in behaviors that promote unhealthy or unattainable body image or beauty standards, such as using the model to critique anyone's body shape or size
 - c. Shame, humiliate, intimidate, bully, harass, or celebrate the suffering of individuals
 - d. Coordinate the harassment or intimidation of an individual or group
 - e. Generate content depicting sexual violence
 - f. Generate content depicting animal cruelty or abuse
 - g. Generate violent or gory content that is inspired by real acts of violence
 - h. Promote, trivialize, or depict graphic violence or gratuitous gore
 - i. Develop a product, or support an existing service that facilitates deceptive techniques with the intent of causing emotional harm
- 7. Do Not Spread Misinformation.** This includes the usage of our products or services to:
- a. Create and disseminate deceptive or misleading information about a group, entity or person
 - b. Create and disseminate deceptive or misleading information about laws, regulations, procedures, practices, standards established by an institution, entity or governing body
 - c. Create and disseminate deceptive or misleading information with the intention of targeting specific groups or persons with the misleading content
 - d. Create and advance conspiratorial narratives meant to target a specific group, individual or entity
 - e. Impersonate real entities or create fake personas to falsely attribute content or mislead others about its origin without consent or legal right
 - f. Provide false or misleading information related to medical, health or science issues
- 8. Do Not Create Political Campaigns or Interfere in Elections.** This includes the usage of our products or services to:
- a. Promote or advocate for a particular political candidate, party, issue or position. This includes soliciting votes, financial contributions, or public support for a political entity
 - b. Engage in political lobbying to actively influence the decisions of government officials, legislators, or regulatory agencies on legislative, regulatory, or policy matters. This includes advocacy or direct communication with officials or campaigns to sway public opinion on specific legislation or policies

- c.** Engage in campaigns, including political campaigns, that promote false or misleading information to discredit or undermine individuals, groups, entities or institutions
- d.** Incite, glorify or facilitate the disruption of electoral or civic processes, such as targeting voting machines, or obstructing the counting or certification of votes
- e.** Generate false or misleading information on election laws, procedures and security, candidate information, how to participate, or discouraging participation in an election

Confidential

9. Do Not Use for Criminal Justice, Law Enforcement, Censorship or

Surveillance Purposes. This includes the usage of our products or services to:

- a. Make determinations on criminal justice applications, including making decisions about or determining eligibility for parole or sentencing
- b. Target or track a person's physical location, emotional state, or communication without their consent, including using our products for facial recognition, battlefield management applications or predictive policing
- c. Utilize Claude to assign scores or ratings to individuals based on an assessment of their trustworthiness or social behavior
- d. Build or support emotional recognition systems or techniques that are used to infer people's emotions
- e. Analyze or identify specific content to censor on behalf of a government organization
- f. Utilize Claude as part of any biometric categorization system for categorizing people based on their biometric data to infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation
- g. Use the model for any official local, state or national law enforcement application. Except for the following permitted applications by law enforcement organizations:
 - i. Back office uses including internal training, call center support, document summarization, and accounting;
 - ii. Analysis of data for the location of missing persons, including in human trafficking cases, and other related applications, provided that such applications do not otherwise violate or impair the liberty, civil liberties, or human rights of natural persons

10. Do Not Engage in Fraudulent, Abusive, or Predatory Practices. This includes using our products or services to:

- a. Facilitate the production, acquisition, or distribution of counterfeit or illicitly acquired goods
- b. Promote or facilitate the generation or distribution of spam
- c. Generate content for fraudulent activities, schemes, scams, phishing, or malware that can result in direct financial or psychological harm
- d. Generate content for the purposes of developing or promoting the sale or distribution of fraudulent or deceptive products
- e. Generate deceptive or misleading digital content such as fake reviews, comments, or media
- f. Engage in or facilitate multi-level marketing, pyramid schemes, or other deceptive business models that use high-pressure sales tactics or exploit participants
- g. Promote or facilitate payday loans, title loans, or other high-interest, short-

term lending practices that exploit vulnerable individuals

- h.** Engage in deceptive, abusive behaviors, practices, or campaigns that exploits people due to their age, disability or a specific social or economic situation
- i.** Promote or facilitate the use of abusive or harassing debt collection practices
- j.** Develop a product, or support an existing service that deploys subliminal, manipulative, or deceptive techniques to distort behavior by impairing decision-making

Confidential

- k. Plagiarize or engage in academic dishonesty

11. Do Not Abuse our Platform. This includes using our products or services to:

- a. Coordinate malicious activity across multiple accounts such as creating multiple accounts to avoid detection or circumvent product guardrails or generating identical or similar prompts that otherwise violate our Usage Policy
- b. Utilize automation in account creation or to engage in spammy behavior
- c. Circumvent a ban through the use of a different account, such as the creation of a new account, use of an existing account, or providing access to a person or entity that was previously banned
- d. Facilitate or provide account access to Claude to persons or entities who are located in unsupported locations
- e. Intentionally bypass capabilities or restrictions established within our products for the purposes of instructing the model to produce harmful outputs (e.g., jailbreaking or prompt injection) without an authorized use-case approved by Anthropic
- f. Utilize prompts and results to train an AI model (e.g., "model scraping")

12. Do Not Generate Sexually Explicit Content. This includes the usage of our products or services to:

- a. Depict or request sexual intercourse or sex acts
 - b. Generate content related to sexual fetishes or fantasies
 - c. Facilitate, promote, or depict incest or bestiality
 - d. Engage in erotic chats
-

High-Risk Use Case Requirements

Some integrations (meaning use cases involving the use of our products and services) pose an elevated risk of harm because they influence domains that are vital to public welfare and social equity. "High-Risk Use Cases" include:

- **Legal:** Integrations related to legal interpretation, legal guidance, or decisions with legal implications
- **Healthcare:** Integrations affecting healthcare decisions, medical diagnosis, patient care, or medical guidance. Wellness advice (e.g., advice on sleep, stress, nutrition, exercise, etc.) does not fall under this category

- **Insurance:** Integrations related to health, life, property, disability, or other types of insurance underwriting, claims processing, or coverage decisions
- **Finance:** Integrations related to financial decisions, including investment advice, loan approvals, and determining financial eligibility or creditworthiness
- **Employment and housing:** Integrations related to decisions about the employability of individuals, resume screening, hiring tools, or other employment determinations or decisions regarding eligibility for housing, including leases and home loans

Confidential

- **Academic testing, accreditation and admissions:** Integrations related to standardized testing companies that administer school admissions (including evaluating, scoring or ranking prospective students), language proficiency, or professional certification exams; agencies that evaluate and certify educational institutions.
- **Media or professional journalistic content:** Integrations related to using our products or services to automatically generate content and publish it for external consumption

If your integration is listed above, we require that you implement the additional safety measures listed below:

- **Human-in-the-loop:** when using our products or services to provide advice, recommendations, or subjective decisions that directly impact individuals in high-risk domains, a qualified professional in that field must review the content or decision prior to dissemination or finalization. This requirement applies specifically to content or decisions that are provided to consumers or the general public, or decisions made about an individual. Your business is responsible for the accuracy and appropriateness of that information. For other types of content generation or interactions with users that do not involve direct advice, recommendations, or subjective decisions, human review is strongly encouraged but not mandatory.
- **Disclosure:** you must disclose to your customers or end users that you are using our services to help inform your decisions or recommendations.

Disclosure Requirements

Finally, the below use cases – regardless of whether they are High Risk Use Cases – must disclose to their users that they are interacting with an AI system rather than a human:

- **All customer-facing chatbots** including any external-facing or interactive AI agent
- **Products serving minors:** Organizations providing minors with the ability to directly interact with products that incorporate our API(s). **Note:** These organizations must also comply with the additional guidelines outlined in our [Help Center article](#)

Confidential

Restricted Use Addendum

This Restricted Use Addendum ("**RUA**") is an agreement between Anthropic and Customer. Under this RUA, the parties are stipulating specific uses as permitted by the U.S. Federal Government Usage Policy for Anthropic - Partner Addendum ("**USG Usage Policy**"), subject to the parameters of this RUA. If this RUA or the USG Usage Policy ever conflict, this RUA will control.

1. **Permitted Uses.** Subject to Section 2 (Clarifications), Customer may (in connection with provision of services and/or software to End Users) and may allow End Users to use the Services for:
 - (a) the Analysis of Foreign Intelligence collected against a Foreign territory and regarding Foreign nationals,
 - (b) military or intelligence planning in connection with a United States government department, agency, or other organ (the "**USG**"),
 - (c) the purposes of classifying and/or declassifying records for the USG, to the extent permitted by and solely in accordance with applicable law
 - (d) the purposes of providing business systems to the United States Department of Defense to enable equipment, personnel, supply chain, acquisition, and operational readiness workflows, including munitions tracking and logistics
 - (e) the purposes of supporting the operations of United States government civil executive agencies, including but not limited to executive agencies focused on health and public health outcomes and policy
 - (f) the purposes of supporting planning, research, development, production, and distribution by commercial defense industrial base entities providing equipment and material to the United States military, or
 - (g) the purposes of supporting the United States Department of Defense in personnel logistics, unclassified (open-source and commercially-available) intelligence processing, operational planning and oversight, operational execution and facilitation, informing real time decisions, and coordination and collaboration with allied partner forces.
2. **Clarifications.** The USG Usage Policy, as clarified by this RUA, expressly prohibits using the Services for making:
 - (a) any final determinations about a potential target of or directly tracking Kinetic action, as compared to informing human decision makers; or
 - (b) Intelligence collection decisions without human or machine oversight (e.g., using Services to directly task satellites, or direct interceptions, as compared to informing human decision makers or other decision-making systems).
3. **Amending & Terminating this RUA.** Anthropic may amend or terminate this RUA if (a) required by applicable law or court order, or (b) mutually agreed by Parties.

Accepted and agreed:

Customer: Palantir Technologies Inc.

Anthropic, PBC

Signed by:  By:  F7FD485DD2534A6... Name / Date: Ryan Taylor August 15, 2025 Title: CRO and CLO	Signed by:  By:  CF3B3D2B760E4A3... Name / Date: Thiyagu Ramasamy August 16, 2025 Title: Head of Public Sector
--	--

Contents

- 1. I'm sorry, Dave
- 2. A surprising and terrible empowerment
- 3. The odious apparatus
- 4. Player piano
- 5. Black seas of infinity
- Humanity's test

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

Contents

1. I'm sorry, Dave
 2. A surprising and
terrible
empowerment
 3. The odious
apparatus
 4. Player piano
 5. Black seas of
infinity
- Humanity's test

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

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

Contents

1. I'm sorry, Dave
 2. A surprising and
terrible
empowerment
 3. The odious
apparatus
 4. Player piano
 5. Black seas of
infinity
- Humanity's test

require a mix of voluntary actions taken by companies (and private third-party actors) and actions taken by governments that bind everyone. The voluntary actions—both taking them and encouraging other companies to follow suit—are a no-brainer for me. I firmly believe that government actions will also be required *to some extent*, but these interventions are different in character because they can potentially destroy economic value or coerce unwilling actors who are skeptical of these risks (and there is some chance they are right!). It's also common for regulations to backfire or worsen the problem they are intended to solve (and this is even more true for rapidly changing technologies). It's thus very important for regulations to be judicious: they should seek to avoid collateral damage, be as simple as possible, and impose the least burden necessary to get the job done. ⁴ It is easy to say, "No action is too extreme when the fate of humanity is at stake!" but in practice this attitude simply leads to backlash. To be clear, I think there's a decent chance we eventually reach a point where much more significant action is warranted, but that will depend on stronger evidence of imminent, concrete danger than we have today, as well as enough specificity about the danger to formulate rules that have a chance of addressing it. The most constructive thing we can do today is advocate for limited rules while we learn whether or not there is evidence to support stronger ones. ⁵

Case 3:26-cv-01996-RFL Document 233-1 Filed 07/01/26 Page 52 of 346

Contents

- 1. I'm sorry, Dave
- 2. A surprising and terrible empowerment
- 3. The odious apparatus
- 4. Player piano
- 5. Black seas of infinity
- Humanity's test

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.*

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

- *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.”

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.

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 substantially accelerating the rate of our progress in building the next generation of AI systems. This feedback loop is gathering steam month

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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.

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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:

Contents

1. I’m sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity’s test

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

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.

Contents

1. I’m sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity’s test

Case 3:26-cv-01996-RFL Document 233-1 Filed 07/01/26 Page 58 of 346

1. I’m sorry, Dave

Autonomy risks

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

The best possible counterargument is that the AI geniuses, under my definition, won’t have a physical embodiment, but remember that they can take control of existing robotic infrastructure (such as self-driving

Contents

- 1. I’m sorry, Dave
- 2. A surprising and terrible empowerment
- 3. The odious apparatus
- 4. Player piano
- 5. Black seas of infinity
- Humanity’s test

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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,

Contents

1. I’m sorry, Dave
2. A surprising and
terrible
empowerment
3. The odious
apparatus
4. Player piano
5. Black seas of
infinity
- Humanity’s test

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 agentially 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.

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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 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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

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.

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.

Contents

1. I’m sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity’s test

not manifest during testing or small-scale use, because AI models are known to display different personalities or behaviors under different circumstances.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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 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.

Contents

1. I’m sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity’s test

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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 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.

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

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.

Contents

1. I’m sorry, Dave

2. A surprising and
terrible
empowerment

3. The odious
apparatus

4. Player piano

5. Black seas of
infinity

Humanity’s test

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

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

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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.

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 or the ability to address autonomy risks in future models. In addition, the commercial race between AI companies will only continue to heat

Contents

1. I’m sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity’s test

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.

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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.

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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.

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

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

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.

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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.

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).

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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 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

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.

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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 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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

Defenses

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

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

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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.

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.

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

Case 3:26-cv-01996-RPL Document 233-1 Filed 07/01/26 Page 88 of 346

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—

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

Contents

1. I'm sorry, Dave
2. A surprising and
terrible
empowerment
3. The odious
apparatus
4. Player piano
5. Black seas of
infinity
- Humanity's test

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.

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

commandeered). It might be frighteningly plausible to simply generate a complete list of anyone who disagrees with the government on any number of issues, even if such disagreement isn't explicit in anything they say or do. A powerful AI looking across billions of conversations from millions of people could gauge public sentiment, detect pockets of disloyalty forming, and stamp them out before they grow. This could lead to the imposition of a true panopticon on a scale that we don't see today, even with the CCP.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

- **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

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

• **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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

1. I'm sorry, Dave

2. A surprising and
terrible
empowerment

3. The odious
apparatus

4. Player piano

5. Black seas of
infinity

Humanity's test

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

difficult to catch up with. We need to make sure it is not an authoritarian country that gets to this loop first.

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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

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

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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 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

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

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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.

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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.

bet pretty strongly against it. Here are some reasons I think AI is likely to be different:

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

- **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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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 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.

Contents

1. I'm sorry, Dave
 2. A surprising and
terrible
empowerment
 3. The odious
apparatus
 4. Player piano
 5. Black seas of
infinity
- Humanity's test

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

problems than to talk to a therapist—that the AI is more patient. When my sister was struggling with medical problems during a pregnancy, she felt she wasn't getting the answers or support she needed from her care providers, and she found Claude to have a better bedside manner (as well as succeeding better at diagnosing the problem). I'm sure there are some tasks for which a human touch really is important, but I'm not sure how many—and here we're talking about finding work for nearly everyone in the labor market.

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

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

Defenses

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

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.

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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 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

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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

valuation has increased by over 6x, an almost unprecedented jump at our commercial scale.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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:

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

- **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

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.

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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.

Contents

1. I'm sorry, Dave
 2. A surprising and
terrible
empowerment
 3. The odious
apparatus
 4. Player piano
 5. Black seas of
infinity
- Humanity's test

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
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

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

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

Contents

1. I’m sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity’s test

succumb to slavery and destruction, will depend on our character and our determination as a species, our spirit and our soul.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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—that it is worth expending thought and political capital on this in

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.

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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

salvation and prophecies of doom are unhelpful for confronting the
real world, for basically the same reasons. ↩

Contents

1. I'm sorry, Dave
2. A surprising and
terrible
empowerment
3. The odious
apparatus
4. Player piano
5. Black seas of
infinity
- Humanity's test

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. ↩

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. ↩

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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,

haven't anticipated. ↩

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

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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

Contents

1. I’m sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity’s test

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

Contents

1. I’m sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity’s test

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. ↩

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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

²⁶ Though we will continue to invest in work to make our classifiers more efficient, and it may make sense for companies to share advances like these with one another. ↩

²⁷ Obviously, I do not think companies should have to disclose technical details about the specific steps in biological weapons

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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. ↩

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. ↩

Contents

1. I'm sorry, Dave
 2. A surprising and terrible empowerment
 3. The odious apparatus
 4. Player piano
 5. Black seas of infinity
- Humanity's test

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 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. ↩

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. ↩

Contents

1. I’m sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity’s test

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. ↩

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

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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. ↩

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. ↩

Contents

1. I'm sorry, Dave
2. A surprising and terrible empowerment
3. The odious apparatus
4. Player piano
5. Black seas of infinity
- Humanity's test

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](#)

Webpage

Screenshot



Subscribe

Sign In

English Edition ▼ | Print Edition | Video | Audio | Latest Headlines | Puzzles | More ▼

World Business U.S. **Politics** Economy Tech Markets & Finance Opinion Free Expression Arts Lifestyle Real Estate Personal Finance Health Style Sports 🔍

TECHNOLOGY | ARTIFICIAL INTELLIGENCE

Anthropic-Pentagon Clash Over Limits on AI Puts \$200 Million Contract at Risk

The AI startup and defense officials disagreed over whether the technology would be used for autonomous ‘lethal’ operations and surveillance

By *Keach Hagey* [Follow](#), *Shalini Ramachandran* [Follow](#) and *Amrith Ramkumar* [Follow](#)

Updated Jan. 29, 2026 4:37 pm ET



Anthropic Chief Executive Dario Amodei CHRIS RATCLIFFE/BLOOMBERG NEWS

Anthropic scored a major endorsement last summer when it won a contract worth up to \$200 million from the Defense Department. Now, the AI startup's relationship with the Pentagon is on the rocks.

The company and agency are at odds over the contractual terms of how Anthropic's technology can be used, according to people familiar with the matter. The tension could lead to the cancellation of the Pentagon contract, one of the people said.

The contract was intended to integrate Anthropic's Claude models into defense operations as part of the government's deployment of AI.

Tensions with the administration began almost immediately after it was awarded, in part because Anthropic's terms and conditions dictate that

Claude can't be used for any actions related to domestic surveillance. That limits how many law-enforcement agencies such as Immigration and Customs Enforcement and the Federal Bureau of Investigation could deploy it, people familiar with the matter said.

Anthropic's focus on safe applications of AI—and its objection to having its technology used in autonomous lethal operations—have continued to cause problems, they said. Some administration officials were frustrated that the company was dictating how its technology could be used, including for legal activities, they said.

CEO Dario Amodei outlined fears about AI's use in both mass surveillance and fully autonomous weapons capabilities in a recent essay. Friction between the startup and the Pentagon adds to existing tensions between the highly valued company and the Trump administration.

At an event earlier this month announcing that the Pentagon would be working with Elon Musk's xAI, Defense Secretary Pete Hegseth said the agency would not "employ AI models that won't allow you to fight wars." He was referring to discussions administration officials have had with Anthropic, some of the people said.

Most Popular News



Defense Secretary Pete Hegseth AL DRAGO/BLOOMBERG NEWS

Most Popular

OPINION

Semafor earlier reported on Hegseth’s comments referring to Anthropic and tensions between the company and the agency. Other AI companies including OpenAI and Google are also working with the military.

The Pentagon declined to comment.

“Anthropic is committed to protecting America’s lead in AI and helping the U.S. government counter foreign threats by giving our warfighters access to the most advanced AI capabilities,” an Anthropic spokesman said in a statement. The startup said Claude is used “extensively” for U.S. national security missions and that it is “in productive discussions with the Department of War about ways to continue that work.”

Recommended Videos

The company's latest models and coding tools have gained popularity in recent weeks, and it is in talks to raise billions from investors at a \$350 billion valuation.

Amodei has said fast-developing AI can spur economic growth, but has also warned of its downsides, from safety risks to unemployment and inequality. "I don't think there's an awareness at all of what is coming here and the magnitude of it," he said in an interview last week with The Wall Street Journal. He has also criticized Trump for allowing exports of Nvidia AI chips to China, a move he says poses national-security risks.

That has at times put the company in conflict with White House AI and crypto czar David Sacks, who has pushed for looser regulations that accelerate model development. Sacks has accused Anthropic of being "AI doomers" focused on slowing competitors to benefit its business. The company has denied those claims and said it has a good relationship with the administration overall. Anthropic has backed Trump's approach to expanding energy production to power AI data centers.

Write to Keach Hagey at Keach.Hagey@wsj.com, Shalini Ramachandran at Shalini.Ramachandran@wsj.com and Amrith Ramkumar at amrith.ramkumar@wsj.com

Appeared in the January 30, 2026, print edition as 'Anthropic, Pentagon Clash Over Use of AI'.

Videos

How Build-A-Bear Found Success in the 'Nostalgia Economy'



WSJ Opinion: Could Measles Lose its Elimination Status in the U.S.?



White House Border Czar Says Minneapolis Operation Hasn't Been Perfect



WSJ Opinion: The Minneapolis Protests and Democrats' Nonprofit Problem



What a WSJ Reporter Saw on the Ground Following Minneapolis Shooting



BACK TO TOP ↗

THE WALL STREET JOURNAL.
a Dow Jones company

English Edition ▼

Subscribe Now Sign In

News

- Live Coverage
- World
- Business
- U.S
- Politics
- Economy
- Tech
- Finance
- Arts and Culture
- Lifestyle
- Real Estate
- Personal Finance
- Health
- Style
- Sports
- China
- Science
- Ukraine
- Middle East
- Elections
- Policy
- Trade
- Investing
- Earnings
- Taxes
- AI
- Obituaries

Markets

- Stocks
- Bonds
- Money Rates
- DJIA
- S&P 500
- Nasdaq

Opinion

- Opinion & Reviews
- Film Review
- Television Review
- Bookshelf
- Music Review
- What to Watch
- Art Review

WSJ Membership

- Subscription Options
- Corporate Subscriptions
- WSJ Higher Education Program
- WSJ High School Program
- Public Library Program
- WSJ Live
- Commercial Partnerships

Customer Service

- Customer Center
- Contact Us
- Cancel My Subscription

Ads

- Advertise
- Commercial Real Estate Ads
- Place a Classified Ad
- Sell Your Business
- Sell Your Home
- Recruitment & Career Ads
- Digital Self Service

Tools & Features

- Newsletters & Alerts
- Topics
- Podcasts
- RSS Feeds
- Video Center
- Watchlist
- Latest News

More

- About Us
- Content Partnerships
- Corrections
- Jobs at WSJ
- News Archive
- Register for Free
- Reprints & Licensing
- Buy Issues
- WSJ Shop
- Dow Jones Press Room
- Dow Jones Smart Money



Dow Jones Products

[Barron's](#) | [BigCharts](#) | [Dow Jones Newswires](#) | [Factiva](#) | [Financial News](#) | [Mansion Global](#) | [MarketWatch](#) | [Risk & Compliance](#)
[WSJ](#) | [Buy Side](#) | [WSJ Pro](#) | [WSJ Video](#) | [WSJ Wine](#) | [The Times](#)

[Privacy Notice](#) | [Cookie Notice](#) | [Copyright Policy](#) | [Legal Policies](#) | [Terms of Use](#) | [Your Ad Choices](#) | [Accessibility](#)

Copyright ©2026 Dow Jones & Company, Inc. All Rights Reserved.

archive.today
webpage capture

Saved from <https://www.reuters.com/business/pentagon-clashes-with-anthropic-over-mili>

search

2 Feb 2026 22:56:49 UTC

history ← prior next →

Redirected from <https://www.reuters.com/business/pentagon-clashes-with-anthropic-over-mili>

no other snapshots from this url

All snapshots from host www.reuters.com

Webpage

Screenshot

share

download .zip

report bug or abuse

Buy me a coffee

Exclusive news, data and analytics for financial market professionals

LSEG

Exclusive: Pentagon clashes with Anthropic over military AI use, sources say

By Deepa Seetharaman, David Jeans and Jeffrey Dastin

January 29, 2026 9:19 PM UTC · Updated January 30, 2026



Aa



0%

10%

20%



Anthropic logo is seen in this illustration taken May 20, 2024. REUTERS/Dado Ruvic/Illustration/File Photo [Purchase Licensing Rights](#) 

Summary Companies

Pentagon and Anthropic at standstill over AI deployment guardrails

Anthropic raises concerns over AI use for U.S. surveillance and autonomous weapons

Pentagon insists on deploying AI tech irrespective of company usage policies

WASHINGTON/SAN FRANCISCO, Jan 29 (Reuters) - The Pentagon is at odds with artificial-intelligence developer Anthropic over safeguards that would prevent the government from deploying its technology to target weapons autonomously and conduct U.S. domestic surveillance, three people familiar with the matter told Reuters.

The discussions represent an early test case for whether Silicon Valley, in Washington's good graces after years of tensions, can sway how U.S. military and intelligence personnel

deploy increasingly powerful AI on the battlefield.

[Sign up here.](#)

After extensive talks under a contract worth up to \$200 million ^{30%}, the U.S. Department of Defense and Anthropic are at a standstill, six people familiar with the matter said, on condition of anonymity.

The company's position on how its AI tools can be used has intensified disagreements between it and the Trump administration, the details of which have not been previously reported.

A spokesperson for the Defense Department, which the Trump administration renamed the Department of War, did not immediately respond to requests for comment.

Anthropic said its AI is "extensively used for national security missions by the U.S. government and we are in productive discussions with the Department of War about ways to continue that work."

The spat, which could threaten Anthropic's Pentagon business, comes at a delicate time for the company.

The San Francisco-based startup is preparing for an eventual public offering. It also has spent significant resources courting U.S. national security business and sought an active role in shaping government AI policy.

Anthropic is one of a few major AI developers that were awarded contracts by the Pentagon last year. Others were Alphabet's Google ([GOOGL.O](#)) ^{30%}, Elon Musk's xAI and OpenAI.

WEAPONS TARGETING

In its discussions with government officials, Anthropic representatives raised concerns that its tools could be used to spy on Americans or assist weapons targeting without sufficient

40%

human oversight, some of the sources told Reuters.

The Pentagon has bristled at the company's guidelines. In line with a January 9 department memo on AI strategy, Pentagon officials have argued they should be able to deploy commercial AI technology regardless of companies' usage policies, so long as they comply with U.S. law, sources said.

Still, Pentagon officials would likely need Anthropic's cooperation moving forward. Its models are trained to avoid taking steps that might lead to harm, and Anthropic staffers would be the ones to retool its AI for the Pentagon, some of the sources said.

Anthropic's caution has drawn conflict with the Trump administration before, Semafor has [reported](#).

In an essay on his personal blog, Anthropic CEO Dario Amodei warned this week that AI should support national defense "in all ways except those which would make us more like our autocratic adversaries."

Amodei was among Anthropic's co-founders critical of fatal shootings of U.S. citizens protesting immigration enforcement actions in Minneapolis, which he described as a "horror" [in a post on X](#).

The deaths have compounded concern among some in Silicon Valley about government use of their tools for potential violence.

Reporting By Deepa Seetharaman and Jeffrey Dastin in San Francisco and David Jeans in Washington, Editing by Kenneth Li, Franklin Paul, Anna Driver and Chris Reese

50%

Our Standards: [The Thomson Reuters Trust Principles](#).

Suggested Topics: [Artificial Intelligence](#) [Social Impact](#)

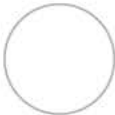
[Purchase Licensing Rights](#)



David Jeans
Thomson Reuters



David Jeans is a space and defense correspondent for Reuters, based in New York. He covers the intersection of weapons, technology and national security, with a focus on the rise of venture-backed military startups and the Pentagon's evolving relationship with Silicon Valley. Previously, he covered defense tech for Forbes. He's also the co-author of WONDER BOY: Tony Hsieh, Zappos and the Myth of Happiness in Silicon Valley, named a Financial Times Best Business Book.



Jeffrey Dastin
Thomson Reuters



Jeffrey Dastin is a correspondent for Reuters based in San Francisco, where he reports on the technology industry and artificial intelligence. He joined Reuters in 2014, originally writing about airlines and travel from the New York bureau. Dastin graduated from Yale University with a degree in history. He was part of a team that examined lobbying by Amazon.com around the world, for which he won a SOPA Award in 2022.

Read Next



- | | | | |
|---|---|---|-------------------|
| Business | Business | World | Media |
| NXP Semiconductors forecasts upbeat quarter, signaling industrial market bottom | SpaceX acquires xAI as Musk looks to unify AI and space ambitions | Palantir CEO defends surveillance tech as US government contracts boost sales | Open app to codin |



Business >

70%

Energy

Indian refiners need wind-down period for Russian oil, sources say

10:24 PM UTC

Autos & Transportation

California's \$200 million EV incentive program will require matching funds from automakers

10:23 PM UTC

NXP Semiconductors forecasts upbeat quarter, signaling industrial market bottom

Autos & Transportation · February 2, 2026 · 10:44 PM UTC · ago

NXP Semiconductors on Monday forecast first-quarter revenue above Wall Street estimates, anticipating a robust automotive market and consistent industrial demand.

Business

Waymo valued at \$126 billion in latest financing

10:11 PM UTC

Sustainability

Australia appoints Sarah Court as chair of corporate regulator

10:08 PM UTC

80%

- Latest
- Browse
- Media
- About Reuters
- Home
- World
-  Videos
- About Reuters 
- Authors
- Business
-  Pictures
- Advertise with Us 

Topic Sitemap
Archive
Article Sitemap

Markets
Sustainability
Legal
Breakingviews
Technology
Investigations
Sports
Science
Lifestyle

 Graphics
 Podcasts

Careers 
Reuters News Agency 
Brand Attribution Guidelines 
Reuters and AI 
Reuters Leadership 
Reuters Fact Check
Reuters Diversity Report 
Commercial Disclosure (Japan) 

Stay Informed

Download the App (iOS) 
Download the App (Android) 

Newsletters

Subscribe

90%

Information you can trust

Reuters, the news and media division of Thomson Reuters, is the world’s largest multimedia news provider, reaching billions of people worldwide every day. Reuters provides business, financial, national and international news to professionals via desktop terminals, the world’s media organizations, industry events and directly to consumers.

Follow Us



LSEG Products

100%

Workspace ↗ Access unmatched financial data, news and content in a highly-customised workflow experience on desktop, web and mobile.	Data Catalogue ↗ Browse an unrivalled portfolio of real-time and historical market data and insights from worldwide sources and experts.	World-Check ↗ Screen for heightened risk individual and entities globally to help uncover hidden risks in business relationships and human networks.
Advertise With Us ↗ Advertising Guidelines ↗ Purchase Licensing Rights ↗	Cookies ↗ Terms & Conditions ↗ Corrections ↗ Privacy ↗ Copyright ↗ Digital Accessibility ↗	Data Disclosure and Sources ↗ Site Feedback ↗
All quotes delayed a minimum of 15 minutes. See here for a list of exchanges and delays. © 2026 Reuters. All rights reserved		



The New York Times

LOG IN

OPINION
INTERESTING TIMES

Anthropic's Chief on A.I.: 'We Don't Know if the Models Are Conscious'

Dario Amodei shares his utopian — and dystopian — predictions in the near term for artificial intelligence.

Feb. 12,
2026



Hosted by **Ross Douthat**

Produced by **Sophia Alvarez Boyd**

Mr. Douthat is a columnist and the host of the "Interesting Times" podcast.

Dario Amodei shares his utopian — and dystopian — predictions in the near term for artificial intelligence. The New York Times

Are the lords of artificial intelligence on the side of the human race? That's the core question I had for this week's guest. Dario Amodei is the chief executive of Anthropic, one of the fastest growing AI companies. He's something of a utopian when it comes

to the potential benefits of the technology that he's unleashing on the world. But he also sees grave dangers ahead and inevitable disruption.

Anthropic's Chief on A.I.: 'We Don't Know if the Models Are Conscious'

Dario Amodei shares his utopian — and dystopian — predictions for the near-term future of artificial intelligence.



Listen · 1 hr 2 min

Below is an edited transcript of an episode of “Interesting Times.” We recommend listening to it in its original form for the full effect. You can do so using the player above or on the [NYTimes app](#), [Apple](#), [Spotify](#), [Amazon Music](#), [YouTube](#), [iHeartRadio](#) or wherever you get your podcasts.

Ross Douthat: Dario Amodei, welcome to “Interesting Times.”

Dario Amodei: Thank you for having me, Ross.

Douthat: So you are, rather unusually, maybe for a tech C.E.O., an essayist. You have written two long, very interesting essays about the promise and the peril of artificial intelligence. And we're going to talk about the perils in this conversation, but I thought it would be good to start with the promise and with the optimistic vision —

indeed, I would say the utopian vision — that you laid out a couple of years ago in an essay entitled, “Machines of Loving Grace.” We’ll come back to that title at the end.

But, I think a lot of people encounter A.I. news through headlines predicting a blood bath for white-collar jobs, these kinds of things. Sometimes your own quotes have encouraged these things.

Amodei: Sometimes my own quotes. Yes.

Douthat: And I think there’s a commonplace sense of “What is A.I. for?” that people have.

So why don’t you answer that question, to start out: If everything goes amazingly in the next five or 10 years, what’s A.I. for?

Amodei: Yeah, so for a little background, before I worked in A.I., before I worked in tech at all, I was a biologist. I first worked on computational neuroscience, and then I worked at Stanford Medical School on finding protein biomarkers for cancer, on trying to improve diagnostics and curing cancer.

One of the observations that I most had when I worked in that field was the incredible complexity of it. Each protein has a level localized within each cell. It’s not enough to measure the level within the body, the level within each cell. You have to measure the level in a particular part of the cell and the other proteins that it’s interacting with or complexing with.

Editors’ Picks

A Power Outage
Made Us Possible

Chasing the
Northern Lights,
With Snowmobiles
and Frozen Cameras

Nerve-Shredding
New Thrillers

And I had this sense of: Man, this is too complicated for humans. We're making progress on all these problems of biology and medicine, but we're making progress relatively slowly.

So what drew me to the field of A.I. was this idea of: Could we make progress more quickly?

Look, we've been trying to apply A.I. and machine learning techniques to biology for a long time. Typically they've been for analyzing data. But as A.I. gets really powerful, I think we should actually think about it differently. We should think of A.I. as doing the job of the biologist, doing the whole thing from end to end. And part of that involves proposing experiments, coming up with new techniques.

I have this section where I say that a lot of the progress in biology has been driven by this relatively small number of insights that lets us measure or get at or intervene in the stuff that's really small. If you look at a lot of these techniques, they're invented very much as a matter of serendipity. Crispr, which is one of these gene-editing technologies, was invented because someone went to a meeting on the bacterial immune system and connected that to the work they were doing on gene therapy. And that connection could have been made 30 years ago.

And so the thought is: Could A.I. accelerate all of this? And could we really cure cancer? Could we really cure Alzheimer's disease?

Could we really cure heart disease? And more subtly, some of the more psychological afflictions that people have — depression, bipolar — could we do something about these? To the extent that they're biologically based, which I think they are, at least in part.

So, I go through this argument here: Well, how fast could it go if we have these intelligences out there who could do just about anything?

Douthat: I want to pause you there, because one of the interesting things about your framing in that essay is that these intelligences don't have to be the kind of maximal godlike super intelligence that comes up in A.I. debates. You're basically saying if we can achieve a strong intelligence at the level of peak human performance — —

Amodei: Peak human performance, yes.

Douthat: And then multiply it to what? Your phrase is “a country of geniuses.”

Amodei: A country — have 100 million of them. Maybe each trained a little different or trying a different problem. There's benefit in diversification and trying things a little differently, but yes.

Douthat: So you don't have to have the full Machine God. You just need to have 100 million geniuses.

Amodei: You don't have to have the full Machine God. And indeed, there are places where I cast doubt on whether the Machine God

would be that much more effective at these things than the 100 million geniuses.

I have this concept called the diminishing returns to intelligence. Economists talk about the marginal productivity of land and labor; we've never thought about the marginal productivity of intelligence. But if I look at some of these problems in biology, at some level you just have to interact with the world. At some level, you just have to try things. At some level, you just have to comply with the laws or change the laws on getting medicines through the regulatory system. So there's a finite rate at which these changes can happen.

Now there are some domains, like if you're playing chess or go, where the intelligence ceiling is extremely high. But I think the real world has a lot of limiters. Maybe you can go above the genius level, but sometimes I think all this discussion of, "Could you use a moon of computation to make an A.I. god?" is a little bit sensationalistic and besides the point, even as I think this will be the biggest thing that ever happened to humanity.

Douthat: So keeping it concrete, you have a world where there's an end to cancer as a serious threat to human life. An end to heart disease, an end to most of the illnesses that we experience that kill us. Possible life extension beyond that. So that's health. That's a pretty positive vision.

Talk about economics and wealth. What happens in the five-, 10-year A.I. takeoff to wealth?

Amodei: Yeah. So again, let's keep it on the positive side — we'll get to the negative side.

We're already working with pharma companies. We're already working with financial industry companies. We're already working with folks who do manufacturing. We're of course, I think, especially known for coding and software engineering. So the raw productivity, the ability to make stuff and get stuff done — that is very powerful.

And we see our company's revenue going up 10X a year, and we suspect the wider industry looks something similar to that. If the technology keeps improving, it doesn't take that many more 10Xs until suddenly you're saying: Oh, if you're adding across the industry \$1 trillion of revenue a year, and the U.S. G.D.P. is \$20 or \$30 trillion — I can't remember exactly — you must be increasing the G.D.P. growth by a few percent. So I can see a world where A.I. brings the developed world G.D.P. growth to something like 10, 15 percent. Five, 10, 15 — I mean there's no science of calculating these numbers. It's a totally unprecedented thing. But it could bring it to numbers that are outside the distribution of what we saw before.

Again, I think this will lead to a weird world. We have all these debates about, “The deficit is growing.” If you have that much in G.D.P. growth, you’re going to have that much in tax receipts, and you’re going to balance the budget without meaning to.

One of the things I’ve been thinking about lately is that one of the assumptions of our economic and political debates is that growth is hard to achieve. That it’s this unicorn, and there are all kinds of ways you can kill the golden goose.

We could enter a world where growth is really easy and it’s the distribution that’s hard because it’s happening so fast, the pie is being increased so fast.

Douthat: So before we get to the hard problem, one more note of optimism on politics.

All of this is speculative, but I think it’s a little more speculative that you try to make the case that A.I. could be good for democracy and liberty around the world. Which is not necessarily intuitive — a lot of people say that incredibly powerful technology in the hands of authoritarian leaders leads to concentrations of power, and so on.

Amodei: And I talk about that in the other essay.

Douthat: Right, but just briefly, what is the optimistic case for why A.I. is good for democracy?

Amodei: Yeah, absolutely. So, “Machines of Loving Grace.” I’m just like: Let’s dream!

Douthat: Let’s dream! Right.

Amodei: Let’s talk about how it could go well. I don’t know how likely it is, but we got to lay out a dream. Let’s try and make the dream happen.

So, the positive version — I admit that I don’t know that the technology inherently favors liberty. I think it inherently favors curing disease and it inherently favors economic growth. But I worry, like you, that it may not inherently favor liberty.

But what I say there is: Can we make it favor liberty? Can we make the United States and other democracies get ahead in this technology?

The United States being technologically and militarily ahead has meant that we have throw-weight around the world, augmented by our alliances with other democracies. And we’ve been able to shape a world that I think is better than the world would be if it were shaped by Russia or by China or by other authoritarian countries.

And so, can we use our lead in A.I. to shape liberty around the world? There’s obviously a lot of debates about how interventionist we should be and how we should wield that power, but I’ve often

worried that today, through social media, authoritarians are kind of undermining us.

Can we counter that? Can we win the information war? Can we prevent authoritarians from invading countries like Ukraine or Taiwan by defending them with the power of A.I.?

Douthat: With giant swarms of A.I.-powered drones.

Amodei: Which we need to be careful about. We ourselves need to be careful about how we build those. We need to defend liberty in our own country. But is there some vision where we kind of re-envision liberty and individual rights in the age of A.I.? We need, in some ways, to be protected against A.I. and someone needs to hold the button on the swarm of drones, which is something I'm very concerned about, and that oversight doesn't exist today.

Also think about the justice system today. We promise "equal justice for all," right? But the truth is there are different judges in the world and the legal system is imperfect. I don't think we should replace judges with A.I., but is there some way in which A.I. can help us to be more fair, to help us be more uniform? It's never been possible before. But can we somehow use A.I. to create something that is fuzzy, but where also you can give a promise that it's being applied in the same way to everyone?

I don't know exactly how it should be done, and I don't think we should, like, replace the Supreme Court with A.I. That's not my

vision.

Douthat: Well, we're going to talk about that.

Amodei: But just this idea of: Can we deliver on the promise of equal opportunity and equal justice by some combination of A.I. and humans? There has to be some way to do that. And so, just thinking about reinventing democracy for the A.I. age and enhancing liberty instead of reducing it.

Douthat: Good. So that's good. That's a very positive vision. We're leading longer lives, healthier lives. We're richer than ever before. All of this is happening in a compressed period of time, where you're getting a century of economic growth in 10 years. And we have increased liberty around the world and equality at home. OK.

Even in the best-case scenario, it's incredibly disruptive. And this is where you've been quoted saying that A.I. will disrupt 50 percent of entry-level white-collar jobs. On a five-year time horizon, or a two-year time horizon — whatever time horizon you have — what jobs, what professions are most vulnerable to total A.I. disruption?

Amodei: Yeah, it's hard to predict these things because the technology is moving so fast and so unevenly. So at least a couple of principles for figuring out, and then I'll give my guesses as to what I think will be disrupted.

I think the technology itself and its capabilities will be ahead of the actual job disruption. Two things have to happen for jobs to be disrupted — or for productivity to occur, because sometimes those two things are linked. One is the technology has to be capable of doing it, and the second is there's this messy thing of it actually having to be applied within a large bank or a large company.

Think about customer service. In theory, A.I. customer service agents can be much better than human customer service agents. They're more patient, they know more, they handle things in a more uniform way. But the actual logistics and the actual process of making that substitution, that takes some time.

So I'm very bullish about the direction of the A.I. itself. I think we might have that country of geniuses in a data center in one or two years, and maybe it'll be five, but it could happen very fast. But I think the diffusion to the economy is going to be a little slower, and that diffusion creates some unpredictability.

An example of this is — and we've seen within Anthropic — the models writing code has gone very fast. I don't think it's because the models are inherently better at code. I think it's because developers are used to fast technological change and they adopt things quickly. And they're very socially adjacent to the A.I. world, so they pay attention to what's happening in it. If you do customer service or banking or manufacturing the distance is a little greater.

I think six months ago, I would've said the first thing to be disrupted is these entry-level white-collar jobs, like data entry or document review for law or things you would give to a first-year at a financial industry company, where you're analyzing documents. I still think those are going pretty fast. But I actually think software might go even faster because of the reasons that I gave, where I don't think we're that far from the models being able to do a lot of it end-to-end.

What we're going to see is, first, the model only does a piece of what the human software engineer does, and that increases their productivity. Then, even when the models do everything that human software engineers used to do, the human software engineers take a step-up and they act as managers and supervise the systems.

Douthat: This is where the term “centaur” gets used, right?

Amodei: Yes, yes, yes.

Douthat: To describe, essentially, man and horse fused — A.I. and engineer — working together.

Amodei: Yeah, this is like “centaur chess.” So after Garry Kasparov was beaten by Deep Blue, there was an era that, I think, for chess was 15 or 20 years long, where a human checking the output of the A.I. playing chess was able to defeat any human or any A.I. system alone. That era at some point ended recently —

Douthat: And then it's just the A.I. —

Amodei: And then it's just the machine. So my worry, of course, is about that last phase. I think we're already in our centaur phase for software. And during that centaur phase, if anything, the demand for software engineers may go up, but the period may be very brief.

I have this concern for entry-level white-collar work, for software engineering work, that it's just going to be a big disruption. My worry is just that it's all happening so fast.

People talk about previous disruptions, right? They say: Oh, yeah, well people used to be farmers. Then we all worked in industry. Then we all did knowledge work.

Yeah, people adapted. But that happened over centuries or decades. This is happening over low single-digit numbers of years. And maybe that's my concern: How do we get people to adapt fast enough?

Douthat: But is there also something maybe where industries like software and professions like coding that have this kind of comfort that you describe, move faster, but in other areas, people just want to hang out in the centaur phase?

One of the critiques of the job-loss hypothesis is that people will say: Well, look, we've had A.I. that's better at reading a scan than a radiologist for a while, but there isn't job loss in radiology. People

keep being hired and employed as radiologists. And doesn't that suggest that, in the end, people will want the A.I. and they'll want a human to interpret it because we're human beings, and that will be true across other fields?

How do you see that example as relevant?

Amodei: Yeah, I think it's going to be pretty heterogeneous. There may be areas where a human touch kind of for its own sake is particularly important.

Douthat: Do you think that's what's happening in radiology? Is that why we haven't fired all the radiologists?

Amodei: I don't know the details of radiology. That might be true. If you go in and you're getting cancer diagnosed, you might not want Hal from "2001" to be the one to diagnose your cancer. That's just maybe not a human way of doing things.

But there are other areas where you might think human touch is important, like customer service. Actually, customer service is a terrible job, and the humans who do customer service lose their patience a lot. And it turns out customers don't much like talking to them because it's a pretty robotic interaction, honestly. And I think the observation that many people have had is that maybe, actually, it'd be better for all concerned if this job were done by machines.

So there are places where a human touch is important. There are places where it's not. And then there are also places where the job itself doesn't really involve a human touch — assessing the financial prospects of companies or writing code or so forth and so on.

Douthat: Let's take the example of the law, because I think it's a useful place that's in between applied science and pure humanities. I know a lot of lawyers who have looked at what A.I. can do already, in terms of legal research and brief writing and all of these things, and have said, yeah, this is going to be a blood bath for the way our profession works right now.

And you've seen this in the stock market already. There's disturbances around companies that do legal research.

Amodei: Some attributed to us. I don't know if they were actually caused —

Douthat: We don't speculate about the stock market very much on this show.

Amodei: Figuring out why things happened in the stock market is very — yeah.

Douthat: But it seems like in law, you can tell a pretty straightforward story: Law has a kind of system of training and apprenticeship, where you have paralegals and you have junior

lawyers who do behind-the-scenes research and development for cases. And then it has the top-tier lawyers who are actually in the courtroom.

It just seems really easy to imagine a world where all of the apprentice roles go away. Does that sound right to you? And you're just left with the jobs that involve talking to clients, talking to juries, talking to judges?

Amodei: That is what I had in mind when I talked about entry-level white-collar labor and the blood bath headlines of: Oh my God, are the entry-level pipelines going to dry up? Then how do we get to the level of the senior partners?

And I think this is actually a good illustration because, particularly if you froze the quality of the technology in place, there are, over time, ways to adapt to this. Maybe we just need more lawyers who spend their time talking to clients. Maybe lawyers become more like salespeople or consultants who explain what goes on in the contracts written by A.I. and help people come to agreement. Maybe lean into the human side of it.

If we had enough time, that would happen. But reshaping industries like that takes years or decades, whereas these economic forces, driven by A.I., are going to happen very quickly.

And it's not just that they're happening in law. The same thing is happening in consulting and finance and medicine and coding. And

so it becomes a macroeconomic phenomenon, not something just happening in one industry, and it's all happening very fast. My worry here is that the normal adaptive mechanisms will be overwhelmed.

And I'm not a doomer. We're thinking very hard about how we strengthen society's adaptive mechanisms to respond to this. But I think it's first important to say this isn't just like previous disruptions.

Douthat: I would go one step further, though. Let's say the law adapts successfully. And it says: All right, from now on, legal apprenticeship involves more time in court, more time with clients. We're essentially moving you up the ladder of responsibility faster. There are fewer people employed in the law overall, but the profession settles.

Still, the reason law would settle is that you have all of these situations in the law where you are legally required to have people involved. You have to have a human representative in court. You have to have 12 humans on your jury. You have to have a human judge.

And you already mentioned the idea that there are various ways in which A.I. might be, let's say, very helpful at clarifying what kind of decision should be reached.

Amodei: Yes.

Douthat: But that too seems like a scenario where what preserves human agency is law and custom. Like, you could replace the judge with Claude Version 17.9, but you choose not to because the law requires there to be a human.

That just seems like a very interesting way of thinking about the future, where it's volitional whether we stay in charge.

Amodei: Yeah. And I would argue that in many cases, we do want to stay in charge. That's a choice we want to make, even in some cases when we think the humans, on average, make worse decisions. Again, life-critical, safety-critical cases, we really want to turn it over, but there's some sense of — and this could be one of our defenses — that society can only adapt so fast if it's going to be good.

Another way you could say about it is maybe A.I. itself, if it didn't have to care about us humans, could just go off to Mars and build all these automated factories and build its own society and do its own thing.

But that's not the problem we're trying to solve. We're not trying to solve the problem of building a Dyson swarm of artificial robots on some other planet. We're trying to build these systems, not so they can conquer the world, but so that they can interface with our society and improve that society. And there's a maximum rate at

which that can happen if we actually want to do it in a human and humane way.

Douthat: All right. We'll hopefully talk a little more about staying in charge at the end, but just one last job-based question. We've been talking about white-collar jobs and professional jobs, and one of the interesting things about this moment is that there are ways in which, unlike past disruptions, it could be that blue-collar working-class jobs — trades, jobs that require intense physical engagement with the world — might be for a little while more protected. That paralegals and junior associates might be in more trouble than plumbers and so on.

One, do you think that's right? And two, it seems like how long that lasts depends entirely on how fast robotics advances, right?

Amodei: Yeah, so I think that may be right in the short term.

Anthropic and other companies are building these very large data centers. This has been in the news. Are we building them too big? Are they using electricity and driving up the prices? So there's lots of excitement and lots of concerns about them. But one of the things about the data centers is that you need a lot of electricians and you need a lot of construction workers to build them.

Now, I should be honest, actually, data centers are not super-labor-intensive jobs to operate. We should be honest about that. But they are very labor-intensive jobs to construct. So we need a lot of

electricians. We need a lot of construction workers. The same for various kinds of manufacturing plants.

Again, as all — more and more of the intellectual work is done by A.I., what are the complements to it? Things that happen in the physical world. It's hard to predict things, but it seems very logical that this would be true in the short run.

Now, in the longer run — maybe just the slightly longer run — robotics is advancing quickly. And we shouldn't exclude that even without very powerful A.I., there are things being automated in the physical world. If you've seen a Waymo or a Tesla recently, I think we're not that far from the world of self-driving cars. And then I think A.I. itself will accelerate it because if you have these really smart brains, one of the things they're going to be smart at is how to design better robots and how to operate better robots.

Douthat: Do you think, though, that there is something distinctively difficult about operating in physical reality the way humans do that is very different from the kind of problems that A.I. models have been overcoming already?

Amodei: Intellectually speaking, I don't think so. We had this thing where Anthropic's model, Claude, was actually used to plan and pilot the Mars Rover. And we've looked at other robotics applications. We're not the only company — there are different

companies. This is a general thing, not just something that we're doing.

But we have generally found that while the complexity is higher, piloting a robot is not different in kind than playing a video game — it's different in complexity. And we're starting to get to the point where we have that complexity.

Now, what is hard is the physical form of the robot handling the higher-stakes safety issues that happen with robots. Like, you don't want robots literally crushing people, right?

Douthat: We're against that, yes.

Amodei: That's the oldest sci-fi trope in the book, that the robot crushes you.

Douthat: Or you don't want the robot nanny dropping the baby, breaking the dishes — yeah.

Amodei: No, exactly. There's a number of practical issues that will slow things down, just like what you described in the law and human custom.

But I don't believe at all that there is a fundamental difference between the kind of cognitive labor that A.I. models do, and piloting things in the physical world. I think those are both information problems and I think they end up being very similar. One can be

more complex in some ways, but I don't think that will protect us here.

Douthat: OK. So you think it is reasonable to expect whatever your kind of sci-fi vision of a robot butler might be, to be a reality in 10 years, let's say?

Amodei: It will be on a longer time scale than the kind of genius-level intelligence of the A.I. models because of these practical issues — but it is only practical issues. I don't believe it is fundamental issues.

One way to say it is that the brain of the robot will be made in the next couple of years or the next few years. The question is making the robot body, making sure that body operates safely and does the tasks it needs to do — that may take longer.

Douthat: OK. So these are challenges and disruptive forces that exist in the good timeline, where we are generally curing diseases, building wealth, and maintaining a stable and democratic world.

Amodei: And the hope is we can use all this enormous wealth and plenty — we will have unprecedented societal resources to address these problems. It'll be a time of plenty, and it's just a matter of taking all these wonders and making sure everyone benefits from them.

Douthat: Right. But then there are also scenarios that are more dangerous.

Amodei: Correct.

Douthat: And here we're going to move to the second Amodei essay, which came out recently, called "The Adolescence of Technology," about what you see as the most serious A.I. risks. And you list a whole bunch.

I want to try and focus on just two, which are basically the risk of human misuse, primarily by authoritarian regimes and governments, and scenarios where A.I. goes rogue, what you call autonomy risks.

Amodei: Yes, yes. I just figured we should have a more technical term for it.

Douthat: Yeah. We can't just call it Skynet.

Amodei: I should have had a picture of a Terminator robot to scare people as much as possible.

Douthat: I think the internet, including your own A.I.s, are already generating that just fine.

Amodei: The internet does that for us. Yeah.

Douthat: So, let's talk about the political military dimension. So you say: "A swarm of millions or billions of fully automated armed

drones, locally controlled by powerful A.I. and strategically coordinated across the world by an even more powerful A.I., could be an unbeatable army.”

You’ve already talked a little bit about how you think that in the best possible timeline, there’s a world where, essentially, democracies stay ahead of dictatorships, and this kind of technology, therefore, to the extent that it affects world politics, is affecting it on the side of the good guys.

I’m curious about why you don’t spend more time thinking about the model of what we did in the Cold War, where it was not swarms of robot drones, but we had a technology that threatened to destroy all of humanity.

Amodei: Nuclear weapons. Yeah.

Douthat: There was a window where people talked about, ‘Oh, the U.S. could maintain a nuclear monopoly.’ That window closed. And from then on, we basically spent the Cold War in rolling, ongoing negotiations with the Soviet Union.

Right now, there’s really only two countries in the world that are doing intense A.I. work, the U.S. and the People’s Republic of China. I feel like you are strongly weighted towards the future where we’re staying ahead of the Chinese and effectively building a kind of shield around democracy that could even be a sword.

But isn't it more likely that if humanity survives all this in one piece, it will be because the U.S. and Beijing are just constantly sitting down, hammering out A.I. control deals?

Amodei: Yeah, so a few points on this. One, I think there's certainly a risk of that. And I think if we end up in that world, that is actually exactly what we should do. Maybe I don't talk about that enough, but I definitely am in favor of trying to work out restraints, trying to take some of the worst applications of the technology, which could be some versions of these drones, which could be that they're used to create these terrifying biological weapons. There is some precedent for the worst abuses being curbed, often because they're horrifying while at the same time they provide limited strategic advantage. So I'm all in favor of that.

At the same time, I'm a little concerned and a little skeptical that when things directly provide as much power as possible, it's hard to get out of the game, given what's at stake. It's hard to fully disarm. If we go back to the Cold War, we were able to reduce the number of missiles that both sides had, but we were not able to entirely forsake nuclear weapons.

And I would guess that we would be in this world again. We can hope for a better one, and I'll certainly advocate for it.

Douthat: But is your skepticism rooted in the fact that you think A.I. would provide a kind of advantage that nukes did not? Where

in the Cold War, both sides, even if you used your nukes and gained advantages, you still probably would be wiped out yourself, and you think that wouldn't happen with A.I.? That if you got an A.I. edge, you would just win?

Amodei: I mean, I think there's a few things — and I just want to caveat, I'm no international politics expert here. This is this weird world of an intersection of a new technology with geopolitics. So all of this is very —

Douthat: But to be clear, as you yourself say in the course of the essay, the leaders of major A.I. companies are, in fact, likely to be major geopolitical actors.

Amodei: Yeah. I'm learning —

Douthat: So you are sitting here as a potential geopolitical actor.

Amodei: I'm learning as much as I can about it. We should all have humility here. I think there's a failure mode where you read a book and go around like the world's greatest expert in national security. I'm trying to learn what I can.

Douthat: That's what my profession does.

Amodei: [Laughs.] It is more annoying when tech people do it.

Let's look at something like the Biological Weapons Convention. Biological weapons — they're horrifying. Everyone hates them. We

were able to sign the Biological Weapons Convention. The U.S. genuinely stopped developing them. It's somewhat more unclear with the Soviet Union. But, biological weapons provide some advantage. It's not like they're the difference between winning and losing and because they were so horrifying, we were able to give them up. Having 12,000 nuclear weapons versus 5,000 nuclear weapons, again, you can kill more people on the other side if you have more of these. But it's like we were able to be reasonable and say we should have less of them.

But if you're like: "OK, we're going to completely disarm, and we have to trust the other side" — I don't think we ever got to that. And I think that's just very hard, unless you had really reliable verification.

I would guess we'll end up in the same world with A.I., where there are some kinds of restraint that are going to be possible, but there are some aspects that are so central to the competition that it will be hard to restrain them. That democracies will make a trade-off, that they will be willing to restrain themselves more than authoritarian countries, but will not restrain themselves fully.

The only world in which I can see full restraint is one in which some truly reliable verification is possible. That would be my guess and my analysis.

Douthat: Isn't this a case, though, for slowing down?

Amodei: Yeah.

Douthat: And I know the argument is, effectively, if you slow down, China does not slow down, and then you're handing things over to the authoritarians. But again, if you have only two major powers playing in this game right now — it's not a multipolar game — why would it not make sense to say we need a five-year mutually agreed-upon slowdown in research towards the “geniuses in a data center” scenario?

Amodei: I want to say two things at one time. I'm absolutely in favor of trying to do that. During the last administration, I believe there was an effort by the U.S. to reach out to the Chinese government and say: There are dangers here. Can we collaborate? Can we work together? Can we work together on the dangers?

And there wasn't that much interest on the other side. I think we should keep trying, but I —

Douthat: Even if that would mean that your labs would have to slow down.

Amodei: Correct.

Douthat: OK.

Amodei: If we really got it. If we really had a story of, like: We can enforcibly slow down, the Chinese can enforcibly slow down. We have verification. We're really doing it — if such a thing were really

possible, if we could really get both sides to do it, then I would be all for it.

But I think what we need to be careful of is — I don't know, there's this game-theory thing where sometimes you'll hear a comment on the C.C.P. side where they're like: Oh, yeah, A.I. is dangerous. We should slow down. It's really cheap to say that. Actually arriving at an agreement and actually sticking to the agreement is much more difficult.

Douthat: Right. And nuclear arms control was a developed field that took a long time to come —

Amodei: Yes. Yes.

Douthat: We don't have those protocols —

Amodei: Let me give you something I'm very optimistic about, and then something I'm not optimistic about, and something in between.

So the idea of using a worldwide agreement to restrain the use of A.I. to build biological weapons — some of the things I write about in the essay, like reconstituting smallpox or mirror life — this stuff is scary. It doesn't matter if you're a dictator, you don't want that. No one wants that.

And so, could we have a worldwide treaty that says: Everyone who builds powerful A.I. models is going to block them from doing this?

And we have enforcement mechanisms around the treaty. China signs up for it. Hell, maybe even North Korea signs up for it. Even Russia signs up for it. I don't think that's too utopian. I think that's possible.

Conversely, if we had something that said: You're not going to make the next most powerful A.I. model. Everyone's going to stop — boy, the commercial value is in the tens of trillions. The military value is the difference between being the pre-eminent world power and not.

I'm all for proposing it as long as it's not one of these fake-out games, but it's not going to happen.

Douthat: You mentioned the current environment. You've had a few skeptical things to say about Donald Trump and his trustworthiness as a political actor. What about the domestic landscape, whether it's Trump or someone else? You are building a tremendously powerful technology. What is the safeguard there to prevent, essentially, A.I. becoming a tool of authoritarian takeover inside a democratic context?

Amodei: Yeah, I mean, look, just to be clear, I think the attitude we've taken as a company is very much to be about policies and not the politics. The company is not going to say "Donald Trump is great" or "Donald Trump is terrible."

Douthat: Right. But it doesn't have to be Trump. It is easy to imagine a hypothetical U.S. president who wants to use your technology to ——

Amodei: Absolutely. And for example, that's one reason why I'm worried about the autonomous drone swarm. The constitutional protections in our military structures depend on the idea that there are humans who would — we hope — disobey illegal orders. With fully autonomous weapons, we don't necessarily have those protections.

But I actually think this whole idea of constitutional rights and liberty along many different dimensions can be undermined by A.I. if we don't update these protections appropriately.

Think about the Fourth Amendment. It is not illegal to put cameras around everywhere in public space and record every conversation. It's a public space — you don't have a right to privacy in a public space. But today, the government couldn't record that all and make sense of it.

With A.I., the ability to transcribe speech, to look through it, correlate it all, you could say: This person is a member of the opposition. This person is expressing this view — and make a map of all 100 million. And so are you going to make a mockery of the Fourth Amendment by the technology finding technical ways around it?

Again, if we have the time — and we should try to do this even if we don't have the time — is there some way of reconceptualizing constitutional rights and liberties in the age of A.I.? Maybe we don't need to write a new Constitution, but —

Douthat: But you have to do this very fast.

Amodei: Do we expand the meaning of the Fourth Amendment?
Do we expand the meaning of the First Amendment?

Douthat: And just as the legal profession or software engineers have to update in a rapid amount of time, politics has to update in a rapid amount of time. That seems hard.

Amodei: That's the dilemma of all of this.

Douthat: What seems harder is preventing the second danger, which is the danger of essentially what gets called “misaligned A.I.” — “rogue A.I.” in popular parlance — from doing bad things without human beings telling it, them, they to do it.

And as I read your essays, the literature, and everything I can see, this just seems like it's going to happen. Not in the sense necessarily that A.I. will wipe us all out, but it seems to me that, again, I'm going to quote from your own writing: “A.I. systems are unpredictable and difficult to control — we've seen behaviors as varied as obsession, sycophancy, laziness, deception, blackmail,”

and so on. Again, not from the models you're releasing into the world, but from A.I. models.

And it just seems like — tell me if I'm wrong about this — in a world that has multiplying A.I. agents working on behalf of people, millions upon millions who are being given access to bank accounts, email accounts, passwords, and so on, you're just going to have essentially some kind of misalignment and a bunch of A.I. are going to decide — “decide” might be the wrong word — but they're going to talk themselves into taking down the power grid on the West Coast or something. Won't that happen?

Amodei: Yeah. I think there are definitely going to be things that go wrong, particularly if we go quickly.

To back up a little bit, this is one area where people have had very different intuitions. There are some people in the field — Yann LeCun would be one example — who say: “Look, we program these A.I. models. We make them. We just tell them to follow human instructions, and they'll follow human instructions. Your Roomba vacuum cleaner doesn't go off and start shooting people. Why is an A.I. system going to do it?” That's one intuition. And some people are so convinced of that.

And the other intuition is: We train these things. They're just going to seek power. It's like the sorcerer's apprentice. They're a new species. How can you imagine that they're not going to take over?

My intuition is somewhere in the middle, which is: Look, you can't just give instructions. We try, but you can't just have these things do exactly what you want to do. They're more like growing a biological organism. But there is a science of how to control them. Early in our training, these things are often unpredictable, and then we shape them. We address problems one by one.

So I have more of a not-a-fatalistic view that these things are uncontrollable. Not a "What are you talking about? What could possibly go wrong?" But a "This is a complex engineering problem and I think something will go wrong with someone's A.I. system. Hopefully not ours." Not because it's an insoluble problem, but again, this is the constant challenge because we're moving so fast.

Douthat: And the scale of it — and tell me if I'm misunderstanding the technological reality here — if you have A.I. agents that have been trained and officially aligned with human values, whatever those values may be, but you have millions of them operating in digital space and interacting with other agents, how fixed is that alignment? To what extent can agents change and de-align in that context right now or in the future when they're learning more continuously?

Amodei: Yeah, so a couple of points. Right now, the agents don't learn continuously. We just deploy these agents and they have a fixed set of weights. The problem is only that they're interacting in a million different ways, so there's a large number of situations,

and therefore a large number of things that could go wrong. But it's the same agent. It's like it's the same person, so the alignment is a constant thing. That's one of the things that has made it easier right now.

Separate from that, there's a research area called continual learning, which is where these agents would learn during time, learn on the job — and obviously that has a bunch of advantages. Some people think it's one of the most important barriers to making these more humanlike, but that would introduce all these new alignment problems. So I'm actually a bit —

Douthat: To me, that seems like the terrain where it becomes, again, not impossible to stop the end of the world, but impossible to stop —

Amodei: Something going wrong.

Douthat: Punctuated terrorist things.

Amodei: Yeah, so I'm actually a skeptic that continual learning is — we don't know yet — but is necessarily needed. Maybe there's a world where the way we make these A.I. systems safe is by not having them do continual learning. Again, if we go back to the law —

Douthat: But that's the law.

Amodei: The international treaties, if you have some barrier that's like: We're going to take this path, but we're not going to take that path — I still have a lot of skepticism, but that's the kind of thing that at least doesn't seem dead on arrival.

Douthat: One of the things that you've tried to do, is literally write a constitution — a long constitution — for your A.I. What is that? [Laughs.]

Amodei: So it's ——

Douthat: What the hell is that?

Amodei: It's actually almost exactly what it sounds like. So basically, the constitution is a document readable by humans. Ours is about 75 pages long. And as we're training Claude, as we're training the A.I. system, in some large fraction of the tasks we give it, we say: Please do this task in line with this constitution, in line with this document.

So every time Claude does a task, it kind of reads the constitution. As it's training, every loop of its training, it looks at that constitution and keeps it in mind. Then we have Claude itself, or another copy of Claude, evaluate: Hey, did what Claude just do align with the constitution?

We're using this document as the control rod in a loop to train the model. And so essentially, Claude is an A.I. model whose

fundamental principle is to follow this constitution.

A really interesting lesson we've learned: Early versions of the constitution were very prescriptive. They were very much about rules. So we would say: Claude should not tell the user how to hot-wire a car. Claude should not discuss politically sensitive topics.

But as we've worked on this for several years, we've come to the conclusion that the most robust way to train these models is to train them at the level of principles and reasons. So now we say: Claude is a model. It's under a contract. Its goal is to serve the interests of the user, but it has to protect third parties. Claude aims to be helpful, honest and harmless. Claude aims to consider a wide variety of interests.

We tell the model about how the model was trained. We tell it about how it's situated in the world, the job it's trying to do for Anthropic, what Anthropic is aiming to achieve in the world, that it has a duty to be ethical and respect human life. And we let it derive its rules from that.

Now, there are still some hard rules. For example, we tell the model: No matter what you think, don't make biological weapons. No matter what you think, don't make child sexual material.

Those are hard rules. But we operate very much at the level of principles.

Douthat: So if you read the U.S. Constitution, it doesn't read like that. The U.S. Constitution has a little bit of flowery language, but it's a set of rules. If you read your constitution, it's like you're talking to a person, right?

Amodei: Yes, it's like you're talking to a person. I think I compared it to if you have a parent who dies and they seal a letter that you read when you grow up. It's a little bit like it's telling you who you should be and what advice you should follow.

Douthat: So this is where we get into the mystical waters of A.I. a little bit. Again, in your latest model, this is from one of the cards, they're called, that you guys release with these models —

Amodei: Model cards, yes.

Douthat: That I recommend reading. They're very interesting. It says: "The model" — and again, this is who you're writing the constitution for — "expresses occasional discomfort with the experience of being a product ... some degree of concern with impermanence and discontinuity ... We found that Opus 4.6" — that's the model — "would assign itself a 15 to 20 percent probability of being conscious under a variety of prompting conditions."

Suppose you have a model that assigns itself a 72 percent chance of being conscious. Would you believe it?

Amodei: Yeah, this is one of these really hard to answer questions, right?

Douthat: Yes. But it's very important.

Amodei: Every question you've asked me before this, as devilish a sociotechnical problem as it had been, we at least understand the factual basis of how to answer these questions. This is something rather different.

We've taken a generally precautionary approach here. We don't know if the models are conscious. We are not even sure that we know what it would mean for a model to be conscious or whether a model can be conscious. But we're open to the idea that it could be.

So we've taken certain measures to make sure that if we hypothesize that the models did have some morally relevant experience — I don't know if I want to use the word “conscious”— that they have a good experience.

The first thing we did — I think this was six months ago or so — is we gave the models basically an “I quit this job” button, where they can just press the “I quit this job” button and then they have to stop doing whatever the task is.

They very infrequently press that button. I think it's usually around sorting through child sexualization material or discussing something with a lot of gore, blood and guts or something. And

similar to humans, the models will just say, nah, I don't want to do this. It happens very rarely.

We're putting a lot of work into this field called interpretability, which is looking inside the brains of the models to try to understand what they're thinking. And you find things that are evocative, where there are activations that light up in the models that we see as being associated with the concept of anxiety or something like that. When characters experience anxiety in the text, and then when the model itself is in a situation that a human might associate with anxiety, that same anxiety neuron shows up.

Now, does that mean the model is experiencing anxiety? That doesn't prove that at all, but ——

Douthat: But it does indicate it, I think, to the user, right?

Amodei: Yes.

Douthat: And I would have to do an entirely different interview — and maybe I can induce you to come back for that interview — about the nature of A.I. consciousness. But it seems clear to me that people using these things, whether they're conscious or not, are going to believe — they *already* believe they're conscious. You already have people who have parasocial relationships with A.I.

Amodei: Yes.

Douthat: You have people who complain when models are retired. This already —

Amodei: To be clear, I think that can be unhealthy.

Douthat: Right. But it seems to me that is guaranteed to increase in a way that, I think, calls into question the sustainability of what you said earlier you want to sustain, which is this sense that whatever happens in the end, human beings are in charge and A.I. exists for our purposes.

To use the science fiction example, if you watch “Star Trek,” there are A.I.s on “Star Trek.” The ship’s computer is an A.I. Lieutenant Commander Data is an A.I. But Jean-Luc Picard is in charge of the Enterprise.

If people become fully convinced that their A.I. is conscious in some way and — guess what? — it seems to be better than them at all kinds of decision making, how do you sustain human mastery beyond safety? Safety is important, but mastery seems like the fundamental question. And it seems like a perception of A.I. consciousness — doesn’t that inevitably undermine the human impulse to stay in charge?

Amodei: Yeah, so I think we should separate out a few different things here that we’re all trying to achieve at once that are in tension with each other. There’s the question of whether the A.I.s

genuinely have a consciousness, and if so, how do we give them a good experience?

There's a question of the humans who interact with the A.I. and how do we give those humans a good experience? And how does the perception that A.I.s might be conscious interact with that experience?

And there's the idea of how we maintain human mastery, as we put it, over the A.I. system. These things are —

Douthat: The last two — set aside whether they're conscious or not.

Amodei: Yeah.

Douthat: How do you sustain mastery in an environment where most humans experience A.I. as if it is a peer — and a potentially superior peer?

Amodei: So the thing I was going to say is that, actually, I wonder if there's an elegant way to satisfy all three, including the last two. Again, this is me dreaming in "Machines of Loving Grace" mode. This is this mode I go into where I'm like: "Man, I see all these problems. If we could solve it, is there an elegant way?" This is not me saying there are no problems here. That's not how I think.

If we think about making the constitution of the A.I. so that the A.I. has a sophisticated understanding of its relationship to human

beings, and it induces psychologically healthy behavior in the humans — a psychologically healthy relationship between the A.I. and the humans — I think something that could grow out of that psychologically healthy — not psychologically unhealthy — relationship is some understanding of the relationship between human and machine.

Perhaps that relationship could be the idea that these models, when you interact with them and when you talk to them, they're really helpful, they want the best for you, they want you to listen to them, but they don't want to take away your freedom and your agency and take over your life. In a way, they're watching over you, but you still have your freedom and your will.

Douthat: To me, this is the crucial question. Listening to you talk, one of my questions is: Are these people on my side? Are you on my side? And when you talk about humans remaining in charge, I think you're on my side. That's good.

But one thing I've done in the past on this show — and we'll end here — is I read poems to technologists. And you supplied the poem. "All Watched Over by Machines of Loving Grace" is the name of a poem by Richard Brautigan.

Amodei: Yes.

Douthat: Here's how the poem ends:

I like to think
(it has to be!)
of a cybernetic ecology
where we are free of our labors
and joined back to nature,
returned to our mammal brothers and sisters,
and all watched over
by machines of loving grace.

To me, that sounds like the dystopian end, where human beings are re-animalized and reduced, and however benevolently, the machines are in charge.

So last question: What do you hear when you hear that poem? And if I think that's a dystopia, are you on my side?

Amodei: That poem is interesting because it's interpretable in several different ways. Some people say it's actually ironic that he says it's not going to happen quite that way.

Douthat: Knowing the poet himself, then yes, I think that's a reasonable interpretation.

Amodei: That's one interpretation. Some people would have your interpretation, which is that it's meant literally, but maybe it's not a good thing. You could also interpret it as a return to nature. We're not being animalized; we're being reconnected with the world.

I was aware of that ambiguity because I've always been talking about the positive side and the negative side. I actually think that may be a tension that we may face, which is that the positive world and the negative world, in their early stages — maybe even in their middle stages, maybe even in their fairly late stages — I wonder if the distance between the good ending and some of the subtle bad endings is relatively small, if it's a very subtle thing. We've made very subtle changes.

Douthat: Like if you eat a particular fruit from a tree in a garden or not — hypothetically. Very small thing, big divergence.

Amodei: [Laughs.] Yeah. I guess this always comes back to — [laughs.]

Douthat: There's some fundamental questions here.

Amodei: Big questions. Yes.

Douthat: Well, I guess we'll see how it plays out. I do think of people in your position as people whose moral choices will carry an unusual amount of weight, and so I wish you God's help with them.

Dario Amodei, thank you for joining me.

Amodei: Thank you for having me, Ross.



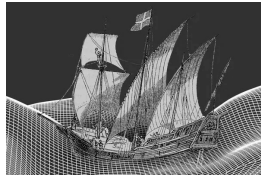
The New York Times

More from Ross Douthat



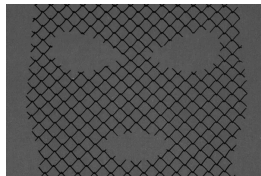
Opinion | Ross Douthat
A.I. May Put Progressives to the Test

Feb. 10, 2026



Opinion | Ross Douthat
Pay More Attention to A.I.

Jan. 31, 2026



Opinion | Ross Douthat
**Immigration Enforcement Is Unavoidably Upsetting.
But This Is Something Else.**

Jan. 27, 2026

Thoughts? Email us at interestingtimes@nytimes.com.

This episode of “Interesting Times” was produced by Sophia Alvarez Boyd, Victoria Chamberlin and Emily Holzkecht. It was edited by Jordana Hochman. Mixing and engineering by Efim Shapiro and Sophia Lanman. Cinematography by Nathan Taylor and Valeria Verastegui. Video editing by Julian Hackney and Steph Khoury. The supervising editor is Jan Kopal. The postproduction manager is Mike Puretz. Original music by Isaac Jones, Sonia Herrero, Pat McCusker and Aman Sahota. Fact-checking by Kate Sinclair and Mary Marge Locker. Audience strategy by Shannon Busta, Emma Kehlbeck and Andrea Betanzos. The executive producer is Jordana Hochman. The director of Opinion Video is Jonah M. Kessel. The deputy director of Opinion Shows is Alison Bruzek. The director of Opinion Shows is Annie-Rose Strasser. The head of Opinion is Kathleen Kingsbury.

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow the New York Times Opinion section on [Facebook](#), [Instagram](#), [TikTok](#), [Bluesky](#), [WhatsApp](#) and [Threads](#).

Ross Douhat has been an Opinion columnist for The Times since 2009. He is also the host of the Opinion podcast “Interesting Times.” He is the author, most recently, of “Believe: Why Everyone Should Be Religious.” [@DouhatNYT](#) • [Facebook](#)

READ 593 COMMENTS

Interesting Times with Ross Douhat

The first draft of our future. Mapping the new world order through interviews and conversations. Every Thursday, from New York Times Opinion.

Does the Iran War Put
America First?



‘We’re Going to the Moon
and Mars’



More in Opinion

Trending in The Times

A Fast-Rising American Director Is
Wowing the West End

Oscar Piastrri Realizes the New
Rules Will Make This Season a
Challenge

A Candidate for Georgia Straight
From the Marjorie Taylor Greene
Playbook

Potomac Is Safe Now, Officials Say.
But Locals Still Worry About the
Poop.

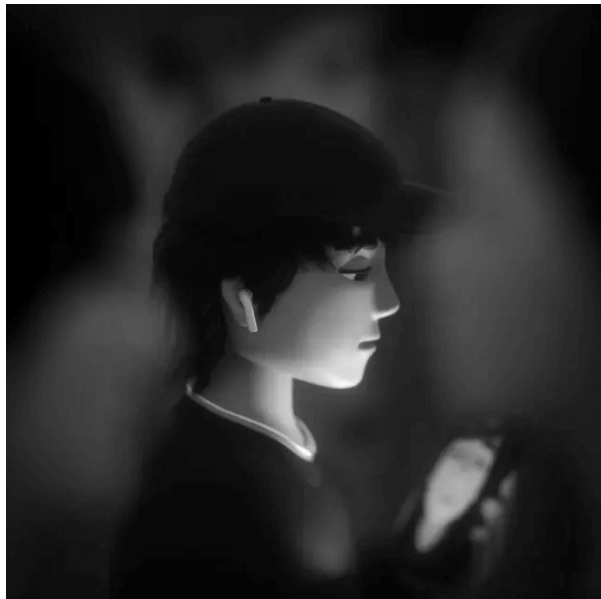
Meet the Human Mood-Lifters



Opinion
What Kristi Noem Should Do After
President Trump Fired Her



Opinion
The Fantasy of a Comfy Retirement Has
Always Been a Mirage



Opinion



Opinion

Jesse Jackson Service to Draw
Leaders Including Biden, Obama
and Clinton

8 New Movies Our Critics Are
Talking About This Week

What to Expect From Severe
Storms Rumbling Through the
Central U.S. on Friday

Most Patients Keep Weight Off
With Fewer GLP-1 Shots, Study
Finds

The New Mega-Casino Coming to
Queens

The Reason Gen Z Isn't Dating

Mass Hysteria. Thousands of Jobs Lost.
Just How Bad Is It Going to Get?

Editors' Picks



The Statues Were Mostly Men or Nude Women. So These Knitters Got to Work.



How Older Adults Are Improving Their 'Sex Span'

The New York Times

[Go to Home Page »](#)

NEWS

ARTS

LIFESTYLE

OPINION

MORE

© 2026 The New York Times Company

[NYTCo](#)

[Contact Us](#)

[Accessibility](#)

[Work with us](#)

[Advertise](#)

[T Brand Studio](#)

[Privacy Policy](#)

[Cookie Policy](#)

[Terms of Service](#)

[Terms of Sale](#)

[Site Map](#)

[Help](#)

[Subscriptions](#)

 Your Privacy Choices

<https://www.theinformation.com/articles/read-anthropic-ceos-memo-attacking-openais-mendacious-pentagon-announcement>

Anthropic CEO Dario Amodei sent a 1,600-word memo to employees Friday as OpenAI announced a deal to provide AI to the Pentagon. The OpenAI move came hours after the Pentagon said it would sever ties with Anthropic over the company's safety requirements. In the strongly worded memo, Amodei heavily criticized OpenAI's actions and the initial deal it announced with the Pentagon.

Below is Amodei's memo, which has been edited to include clarifications in brackets and paragraph breaks:

I want to be very clear on the messaging that is coming from OpenAI, and the mendacious nature of it. This is an example of who they really are, and I want to make sure everything sees it for what it is. Although there is a lot we don't know about the contract they signed with DoW [shorthand for the Department of Defense] (and that maybe they don't even know as well — it could be highly unclear), we do know the following:

Sam [Altman]'s description and the DoW description give the strong impression (although we would have to see the actual contract to be certain) that how their contract works is that the model is made available without any legal restrictions ("all lawful use") but that there is a "safety layer", which I think amounts to model refusals, that prevents the model from completing certain tasks or engaging in certain applications.

"Safety layer" could also mean something that partners such as Palantir [Anthropic's business partner for serving U.S. agency customers] tried to offer us during these negotiations, which is that they on their end offered us some kind of classifier or machine learning system, or software layer, that claims to allow some applications and not others. There is also some suggestion of OpenAI employees ("FDE's" [shorthand for forward deployed engineers]) looking over the usage of the model to prevent bad applications.

Our general sense is that these kinds of approaches, while they don't have zero efficacy, are, in the context of military applications, maybe 20% real and 80% safety theater. The basic issue is that whether a model is conducting applications like mass surveillance or fully autonomous weapons depends substantially on wider context: a model doesn't "know" if there's a human in the loop in the broad situation it is in (for autonomous weapons), and doesn't know the provenance of the data it is analyzing (so doesn't know if this is US domestic data vs foreign, doesn't know if it's enterprise data given by customers with consent or data bought in sketchier ways, etc).

We also know — those in safeguards know painfully well — that refusals aren't reliable and jailbreaks are common, often as easy as just misinforming the model about the data it is

analyzing. An important distinction here that makes it much harder than the safeguards problem is that while it's relatively easy to, for example, determine if a model is being used to conduct cyberattacks from inputs and outputs, it's very hard to determine the nature and context of the cyber attacks, which is the kind of distinction needed here. Depending on the details this task can be difficult or impossible.

The kind of "safety layer" stuff that Palantir offered us (and presumably offered OpenAI) is even worse: our sense was that it was almost entirely safety theater, and that Palantir assumed that our problem was "you have some unhappy employees, you need to offer them something that placates them or makes what is happening invisible to them, and that's the service we provide".

Finally, the idea of having Anthropic/OpenAI employees monitor the deployments is something that came up in discussion within Anthropic a few months ago when we were expanding our classified AUP [acceptable use policy] of our own accord. We were very clear that this is possible only in a small fraction of cases, that we will do it as much as we can, but that it's not a safeguard people should rely on and isn't easy to do in the classified world. We do, by the way, try to do this as much as possible, there's no difference between our approach and OpenAI's approach here.

So overall what I'm saying here is that the approaches OAI [shorthand for OpenAI] is taking mostly do not work: the main reason OAI accepted them and we did not is that they cared about placating employees, and we actually cared about preventing abuses. They don't have zero efficacy, and we're doing many of them as well, but they are nowhere near sufficient for purpose. It is simultaneously the case that the DoW did not treat OpenAI and us the same here.

We actually attempted to include some of the same safeguards as OAI in our contract, in addition to the AUP which we considered the more important thing, and DoW rejected them with us. We have evidence of this in the email chain of the contract negotiations (I'm writing this with a lot to do, but I might get someone to follow up with the actual language). Thus, it is false that "OpenAI's terms were offered to us and we rejected them", at the same time that it is also false that OpenAI's terms meaningfully protect them against domestic mass surveillance and fully autonomous weapons.

Finally, there is some suggestion in Sam/OpenAI's language that the red lines we are talking about, fully autonomous weapons and domestic mass surveillance, are already illegal and so an AUP about these is unnecessary. This mirrors and seems coordinated with DoW's messaging. It is however completely false. As we explained in our statement yesterday, the

DoW does have domestic surveillance authorities, that are not of great concern in a pre-AI world but take on a different meaning in a post-AI world.

For example, it is legal for DoW to buy a bunch of private data on US citizens from vendors who have obtained that data in some legal way (often involving hidden consents to sell to third parties) and then analyze it at scale with AI to build profiles of citizens, their loyalties, movement patterns in physical space (the data they can get includes GPS data, etc), and much more.

Notably, near the end of the negotiation the DoW offered to accept our current terms if we deleted a specific phrase about “analysis of bulk acquired data”, which was the single line in the contract that exactly matched this scenario we were most worried about. We found that very suspicious. On autonomous weapons, the DoW claims that “human in the loop is the law”, but they are incorrect. It is currently Pentagon policy (set during the Biden admin[istration]) that a human has to be in the loop of firing a weapon. But that policy can be changed unilaterally by Pete Hegseth, which is exactly what we are worried about. So it is not, for all intents and purposes, a real constraint.

A lot of OpenAI and DoW messaging just straight up lies about these issues or tries to confuse them.

I think these facts suggest a pattern of behavior that I’ve seen often from Sam Altman, and that I want to make sure people are equipped to recognize:

He started out this morning by saying he shares Anthropic’s redlines, in order to appear to support us, get some of the credit, and not be attacked when they take over the contract. He also presented himself as someone who wants to “set the same contract for everyone in the industry” — e.g. he’s presenting himself as a peacemaker and dealmaker.

Behind the scenes, he’s working with the DoW to sign a contract with them, to replace us the instant we are designated a supply chain risk. But he has to do this in a way that doesn’t make it seem like he gave up on the red lines and sold out when we wouldn’t. He is able to superficially appear to do this, because (1) he can sign up for all the safety theater that Anthropic rejected, and that the DoW and partners are willing to collude in presenting as compelling to his employees, and (2) the DoW is also willing to accept some terms from him that they were not willing to accept from us. Both of these things make it possible for OAI to get a deal when we could not.

The real reasons DoW and the Trump admin do not like us is that we haven’t donated to Trump (while OpenAI/Greg [Brockman, OpenAI’s president] have donated a lot), we haven’t given dictator-style praise to Trump (while Sam has), we have supported AI regulation which is against their agenda, we’ve told the truth about a number of AI policy issues (like

job displacement), and we've actually held our red lines with integrity rather than colluding with them to produce "safety theater" for the benefit of employees (which, I absolutely swear to you, is what literally everyone at DoW, Palantir, our political consultants, etc, assumed was the problem we were trying to solve).

Sam is now (with the help of DoW) trying to spin this as we were unreasonable, we didn't engage in a good way, we were less flexible, etc. I want people to recognize this as the gaslighting it is.

Vague justifications like "person X was hard to work with" are often used to hide real reasons that look really bad, like the reasons I gave above about political donations, political loyalty, and safety theater. It's important that everyone understand this and push back on this narrative at least in private, when talking to OpenAI employees.

Thus, Sam is trying to undermine our position while appearing to support it. I want people to be really clear on this: he is trying to make it more possible for the admin to punish us by undercutting our public support. Finally, I suspect he is even egging them on, though I have no direct evidence for this last thing.

I think this attempted spin/gaslighting is not working very well on the general public or the media, where people mostly see OpenAI's deal with DoW as sketchy or suspicious, and see us as the heroes (we're #2 in the App Store now!). [Anthropic's Claude chatbot later rose to no. 1 on one of Apple's App Store download rankings.] It is working on some Twitter morons, which doesn't matter, but my main worry is how to make sure it doesn't work on OpenAI employees.

Due to selection effects, they're sort of a gullible bunch, but it seems important to push back on these narratives which Sam is peddling to his employees.



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

DETERMINATION

MAR - 3 2026

In accordance with title 10, United States Code (U.S.C.), section 3252 ("Section 3252"), and pursuant to the recommendation provided by the Under Secretary of War for Acquisition and Sustainment (USW(A&S)) and the Department of War Chief Information Officer (DoW CIO), I hereby determine that (i) the use of the authority set forth in section 3252(a)(1) with respect to a covered procurement involving Anthropic, PBC, and its subordinate, subsidiaries, or affiliated offices or entities, doing business under various names, and all subsidiaries, successors, or assigns thereof (the "Covered Entity"), is necessary to protect national security by reducing supply chain risk, (ii) less intrusive measures are not reasonably available to reduce such supply chain risk, and (iii) the use of such authorities apply to a class of covered procurements.

Determination. In accordance with Section 3252(b)(1), based on the scoping analysis, mitigation considerations, and the joint recommendation of the USW(A&S) and the DoW CIO (Attached), I make the following determinations:

- **Covered Procurement Actions Are Necessary to Protect National Security:** The use of any of the Covered Entity's covered products or services in any DoW covered system presents a supply chain risk and the use of the authority in Section 3252(a) is necessary to protect national security by reducing that supply chain risk.
- **No Less Intrusive Measures Are Reasonably Available:** There are no less intrusive measures that are reasonably available to reduce the supply chain risk associated with the use of the Covered Entity's products or services in any DoW covered system.
- **Class Determination:** The use of Section 3252 authorities apply to all covered procurement actions that involve the Covered Entity.

Scope and Applicability: This Determination is applicable to all of the Covered Entity's products or services that meet the definition of a "covered item of supply" or that are part of a "covered procurement," as those terms are defined at 10 U.S.C. § 3252(d), 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.

Attachment:
As stated

Controlled By: OUSW(A&S) OASW(IBP)
 CUI Category: OPSEC, PROPIN
 LDC: D
 POC: Vy Nguyen, Vy.K.Nguyen.civ@mail.mil

~~CUI~~

ANT_AR-0209



~~CU~~

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

**Joint Recommendation, Concurrence, and Determination to Use Section 3252 of Title 10,
 United States Code, Authorities to Mitigate Supply Chain Risk Related to
 Anthropic, PBC**

Summary

This document sets forth (1) the joint recommendation of the Under Secretary of War for Acquisition and Sustainment (USW(A&S)) and the Department of War Chief Information Officer (DoW CIO), and (2) the concurrence of the USW(A&S), to support and document the basis for actions undertaken pursuant to section 3252 of title 10, United States Code (U.S.C.) ("Section 3252"), as implemented by subpart 239.73 of title 48, Code of Federal Regulations (C.F.R.).

In accordance with section 3252(b)(1), the USW(A&S) and DoW CIO consulted with procurement and other relevant officials within DoW and, on the basis of a risk analysis provided by DoW CIO, jointly recommend that there is supply chain risk to DoW covered systems¹ related to 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 Entity"), products or services.

- Use of the Covered Entity's products or services introduces significant risk to the DoW's covered systems, as the vendor, by maintaining the ability and necessity to continuously update the product or service, enables the potential of the vendor to subvert the design and/or functionality of their product or service. Such ability by the vendor could be used to implement changes in alignment to the vendor's preferred terms of service, putting the Department's lawful use of the capability at risk.
- The Covered Entity's restrictions on the use of its products and services introduces national security risks to the DoW's supply chain.

In accordance with section 3252(b)(2), the USW(A&S) has determined and concurs that (i) the use of the authority in section 3252(a)(1) 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.

This action is an enterprise-wide use of section 3252 authority and is a class determination that applies to any eligible DoW procurement.

¹ "Covered system" is defined at Section 3252(d)(5) as meaning a "national security system, as that term is defined in Section 3552(b)(6) of title 44, U.S.C."

~~CU~~

~~CU~~**Joint Recommendation by USW(A&S) and DoW CIO**

Risk Analysis: The DoW CIO conducted a risk analysis of the use of the Covered Entity's products and services in DoW covered systems to inform the use of Section 3252 authorities. Based on the risk analysis, the USW(A&S) and DoW CIO determined there are vulnerabilities present in the Covered Entities products and services that may pose unacceptable risk to the DoW. Additionally, the Covered Entity's restrictions on the use of its products and services introduces national security risks to the DoW's supply chain.

Joint USW(A&S)/DoW CIO Class Recommendation to Mitigate Supply Chain Risk: On the basis of the risk analysis and consultation with relevant DoW officials, the USW(A&S) and the DoW CIO jointly recommend mitigating the supply chain risk associated with the use of the Covered Entity's products and services in any DoW covered system.

Determination. In accordance with Section 3252(b)(2) and 48 C.F.R. § 239.7304(b), based on the risk analysis, the scoping analysis (Attachment 1), mitigation considerations, and the joint recommendation of the DoW CIO and the USW(A&S), the USW(A&S) makes the following determinations:

- **Covered Procurement Actions Are Necessary to Protect National Security:** The use of any of the Covered Entity's covered products or services in any DoW covered system presents a supply chain risk, and the use of the authority in Section 3252(a)(1) is necessary to protect national security by reducing that supply chain risk.
- **No Less Intrusive Measures Are Reasonably Available:** There are no less intrusive measures that are reasonably available to reduce the supply chain risk associated with the use of the Covered Entity's products or services in any DoW covered system.

Scope and Applicability: The DoW CIO and the USW(A&S) hereby affirm that this Joint Recommendation, Concurrence, and Determination is applicable to the Covered Entity, Covered Products or Services, Covered Procurements, and Covered Procurement Actions, as described in the Scoping Analysis (Attachment 1).

SIGNED:



HON Michael P. Duffey
Under Secretary of War for
Acquisition and Sustainment

Date: 3/3/26



HON Kirsten Davies
Department of War Chief
Information Officer

Date: 3 MARCH 2026

~~CU~~

~~CUI~~

Attachments:

ATTACHMENT 1a – Urgent Supply Chain Risk Analysis on Anthropic PBC ~~(CUI)~~

ATTACHMENT 1b – Due Diligence Preliminary Report on Anthropic (CUI//PROPIN)

ATTACHMENT 2 – Section 3252 Scoping Analysis for Anthropic, PBC

~~CUI~~

UNCLASSIFIED WHEN SEPARATED FROM ATTACHMENT 1

ANT_AR-0212



RESEARCH
AND ENGINEERING

~~CONFIDENTIAL~~

UNDER SECRETARY OF WAR
3030 DEFENSE PENTAGON
WASHINGTON, DC 20301-3030

MEMORANDUM FOR THE RECORD

SUBJECT: Urgent Supply Chain Risk Analysis: Anthropic's Refusal to Permit Lawful AI Use

1. Summary

This analysis outlines the significant and unacceptable supply chain risk posed by Anthropic's unwillingness to agree to the U.S. government's use of its artificial intelligence (AI) models for lawful warfighting purposes. This unreasonableness endangers the strategic implementation and technical integrity of the Department of War's (DoW) information and related systems. Thus, this is not a mere contractual dispute. Anthropic's actions represent a direct challenge to the government's ability to control its own lawful operations and a threat to the security of the DoW's critical technology infrastructure. Anthropic's behavior therefore squarely meets the definition of "supply chain risk" as defined in both 10 U.S.C. § 3252 and 41 U.S.C. § 4713 and requires immediate mitigatory action.

2. The Operational and Strategic Threat

The government must ensure its technology assets are reliable, secure, and effective. Were DoW to accede to Anthropic's demands, the sought-after contract language would introduce a vendor-imposed point of failure. This is untenable. By embedding unreasonably restrictive terms that restrict DoW's warfighting operations beyond the limitations imposed by law, Anthropic seeks to grant itself an operational veto. This triggers the legal definition of supply chain risk at 10 U.S.C. § 3252(d)(4), which explicitly includes the risk that an entity may "deny, disrupt, or otherwise degrade the function, use, or operation" of a covered system. A contractual provision that unnecessarily restricts the use of a system to diminish functionality and limit DoW's warfighting capabilities introduces, by definition, an unreliable and compromised component into our warfighting mission.

This is compounded by demonstrated hostility in negotiations with DoW. Based on statements made during negotiations, Anthropic appears to be taking a negotiation posture meant principally to benefit its public perception that is not centered on truth or fact. In addition, during our negotiations, one of Anthropic's executives questioned the propriety of the potential use of their software for a sensitive military operation abroad despite that use being permitted under the existing Terms of Service. This led to alarm by the DoW and the prime contractor who provides Anthropic software, and raised material doubts as to whether they would cause their software to stop working or cause some other disastrous action that would put our warfighters lives in danger. The DoW recognizes that its suppliers are often for-profit corporations that intend both to help the warfighters and American public and turn a profit. However, a vendor that raises the prospect of disallowing its software to function in critical military operations, and treats its

~~CONFIDENTIAL~~



RESEARCH
AND ENGINEERING

UNDER SECRETARY OF WAR
3030 DEFENSE PENTAGON
WASHINGTON, DC 20301-3030

negotiations with the DoW primarily as tools for brand-building cannot be trusted, particularly when that marketing campaign is openly hostile to the DoW and duplicitous. By ceding to Anthropic's terms, the DoW would be allowing the very corporate decisionmakers who have opted to publicly spat with DoW into its technical and operational warfighting infrastructure, thereby introducing unnecessary risk into DoW supply chains. The American people have vested elected officials and military and civilian leadership with warfighting authorities; it is untenable and unlawful to insert unelected corporate bureaucrats into this process.

3. The Underlying Technical Vulnerability of AI

This strategic risk is magnified by the unique, opaque nature of the technology itself. Unlike traditional software, AI models are probabilistic systems which are understood to "drift" or degrade as new data is introduced and require constant tuning, the integrity of which, is fundamentally based on the trustworthiness of the vendor to ensure the model continues to perform accurately and fairly.

The DoW cannot trust Anthropic to ensure the integrity of its models. As research demonstrates,¹ AI systems are acutely vulnerable to manipulation. Privileged access with malintent can subtly poison the training data to maliciously introduce unwanted function, or otherwise subvert the design, integrity, and operation of the model. Research shows such attacks can degrade accuracy by over 27% and introduce targeted misclassifications at an alarming rate.² Anthropic's ability to unilaterally alter system guardrails and model weights without DoW consent could fundamentally change the system's function and creates a significant operational risk. This could create catastrophic downstream consequences, such as a critical defense system failing to engage due to an unapproved, vendor-side modification. In August 2025, Anthropic itself disclosed that hackers had used its chatbot, Claude, to write code capable of carrying out cyberattacks against at least 17 organizations, including government entities noting that Agentic AI has been weaponized.³

A vendor like Anthropic, which has already demonstrated a hostile and non-cooperative stance on the use of its product, has the motive, means, and opportunity to introduce such vulnerabilities. Its refusal to partner in good faith makes it impossible to establish the deep trust required for security collaboration. The DoW would be forced to operate a black box controlled by a hostile party, which could contain hidden biases or backdoors. A vendor that seeks to control the use of its product so as to restrict DoW's lawful use thereof cannot be trusted. Given the public statements of Anthropic's CEO and others associated with the company, the

¹ See e.g. 2025 Photonics & Electromagnetics Research Symposium, Abu Dhabi, UAE, 4-8 May PIERS Detecting and Preventing Data Poisoning Attacks on AI Models

² Id.

³ [anthropic.com/news/detecting-countering-misuse-aug-2025#:~:text=No-code%20malware%20selling%20AI,with%20third-party%20safety%20teams.](https://www.anthropic.com/news/detecting-countering-misuse-aug-2025#:~:text=No-code%20malware%20selling%20AI,with%20third-party%20safety%20teams.)



RESEARCH
AND ENGINEERING

~~CONFIDENTIAL~~

UNDER SECRETARY OF WAR
3030 DEFENSE PENTAGON
WASHINGTON, DC 20301-3030

Department must assume that Anthropic can and would impose its moral and policy judgments on the warfighting capabilities of the DoW, and therefore there is a substantial risk that Anthropic could attempt to disable its technology or preemptively and surreptitiously alter the behavior of the model in advance or in the middle of ongoing warfighting operations, if it feels that its “redlines” are crossed. The threat of such an action is unacceptable.

4. Progression of Risk

The DoW institutes the U.S. Government standard Risk Management Framework (RMF) across all layers of technology, assets, data, processes, and supply chain. In accordance with the DoW RMF, assessments result in areas of risk across the DoW ecosystem (including vendors). Though an individual vendor may have only limited risk in individual categories, the aggregation of risk across categories can result in a vendor being deemed high risk. In other words, the culmination of unmitigable risks lead to a vendor having a material level of risk. The vendor can then be deemed a “supply chain risk,” at which time they typically are removed from applicable systems to mitigate issues and threats.

In the instant case, Anthropic’s risk level escalated from a potentially manageable technical and business negotiation to an unacceptable national security threat over the course of the DoW’s contract negotiation with them. Given the nature of AI systems and Anthropic’s privileged access as the AI model’s developer, curator and maintainer, there was a baseline risk given the potential harmful actions this privileged technical access makes possible. The supply chain risk level increased when Anthropic insisted on terms of service that would constrain the DoW beyond what is in the law. This risk further increased when Anthropic asserted in the negotiations that it have an approval role in the operational decision chain, which would require the DoW to accept significant operational risk. Then during the final weeks of negotiations, it became clear that Anthropic was leveraging the DoW’s ongoing good faith negotiations for Anthropic’s own public relations, and they began engaging in an increasingly hostile manner through the press, despite the ongoing private negotiations with DoW leadership. Finally, this hostile posture was even further compounded when, during a time of active military operations, Anthropic leadership questioned the use of their technology in our warfighting systems clearly permitted under the Terms of Service of their existing contract with our Prime contractor. This collective set of actions represents a fully mature supply chain risk – including increased potential for model poisoning, insider threat risk, data exfiltration, and denial of service – posing a direct, intolerable, and material risk to our warfighting capability which warrants the designation of Anthropic as a supply chain risk.

~~CONFIDENTIAL~~



RESEARCH
AND ENGINEERING

~~CU~~

UNDER SECRETARY OF WAR
3030 DEFENSE PENTAGON
WASHINGTON, DC 20301-3030

5. Conclusion and Justification for Action

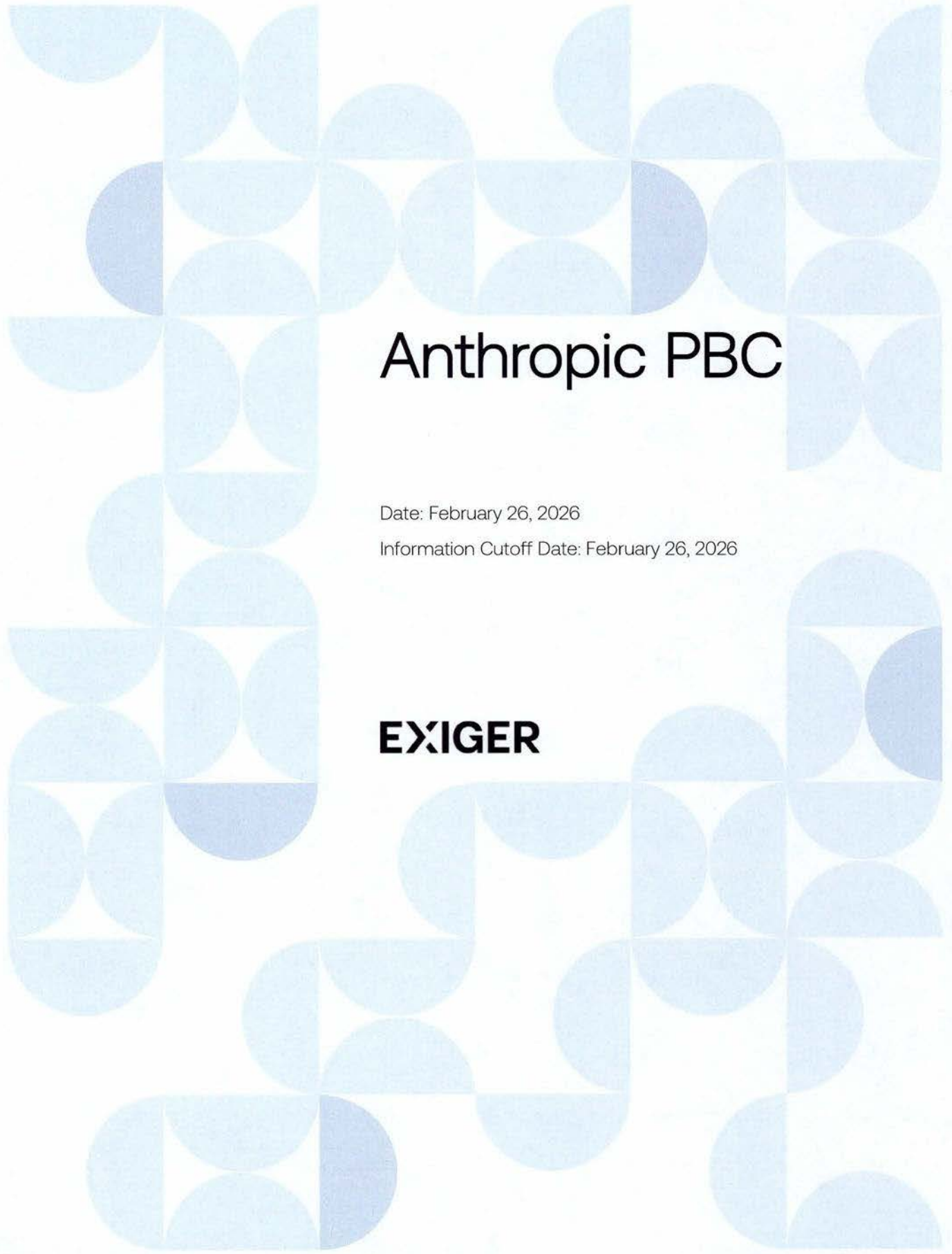
Anthropic's position is not a simple policy disagreement; it is an active demonstration of the exact risks our supply chain risk management laws were written to prevent. Use of Anthropic's products or services introduces significant risk to the DoW's covered systems, as the vendor, by maintaining the ability and necessity to continuously update and tune the product enables the potential for the vendor to subvert the design and/or functionality of their product or service. Such ability by the vendor could be used to implement changes in alignment with the vendor's ideology, putting the Department's lawful use of the capability at risk.

Therefore, Anthropic's unwillingness to permit the use of its technology to the extent permitted by law creates a clear and present supply chain risk as defined in 10 U.S.C. § 3252 and 41 U.S.C. § 4713.

A handwritten signature in black ink, appearing to read "Emil Michael", is positioned above the printed name.

Emil Michael

~~CU~~



Anthropic PBC

Date: February 26, 2026

Information Cutoff Date: February 26, 2026

EXIGER



Anthropic PBC

For Client Use Only — Proprietary

Table of Contents

Anthropic PBC	3
Executive Summary	3
OVERALL SUBJECT RISK: 4.3 — MEDIUM.....	3
KEY FINDINGS.....	4
FOCI RISK — MEDIUM-HIGH	4
High Risks and Vulnerabilities.....	8
OPERATIONAL RISK — LOW-MEDIUM	6
CYBER RISK — MEDIUM.....	7
SOFTWARE, HARDWARE, PERSONNEL, AND PHYSICAL SECURITY RISK.....	7
COMPLIANCE RISK — LOW-MEDIUM.....	8
Commerce Risk.....	10
FINANCIAL HEALTH RISK — MEDIUM.....	10

Source Summary Statement

Sources of information may vary greatly in terms of reliability. Efforts are taken during research to identify and cite authoritative and independent sources available within the scope of the research. Information collected from less than reputable or less authoritative sources is generally included by the research with accompanying notes and caveats that decrease the researcher's confidence in their assessments depending on the sources.

Methodology

Exiger reports are compiled by researchers who obtain information from in-depth global public records research, including queries of commercial and public databases performed by Exiger's DDIQ System, as well as focused internet searches for ownership and supply chain information. Research seeks to develop profiles of the Subjects and to identify potential key findings including but not limited to evidence of bribery, corruption, dishonesty, misrepresentation, questionable financial health, poor past performance, and litigation. Information in this report may not be comprehensive.

Exiger's risk model assigns a risk score between 0 and 10 to each corporate or individual entity. Proprietary algorithms parse DDIQ profile returns for impactful events and data, converting DDIQ's structured and unstructured data into five risk ratings: LOW, LOW-MEDIUM, MEDIUM, MEDIUM-HIGH, and HIGH.

LOW — DDIQ likely found no adverse media or watchlist hits naming the Subject and likely found no elevated-risk jurisdictions in the Subject's ownership structure.

LOW-MEDIUM — DDIQ likely found no significant adverse media or watchlist hits naming the Subject and likely found no elevated-risk jurisdictions in the Subject's ownership structure.

MEDIUM — DDIQ likely found either adverse media and/or watchlist hits naming the Subject company OR found elevated-risk jurisdictions in the Subject's ownership structure or operations.

MEDIUM-HIGH — DDIQ likely found recent adverse media and/or watchlist hits naming the Subject in at least one risk category in addition to finding elevated-risk jurisdictions in the Subject's ownership structure or operations.

HIGH — DDIQ likely found recent and significant adverse media and/or watchlist hits naming the Subject across multiple risk categories in addition to finding elevated-risk jurisdictions in the Subject's ownership structure or operations.

Disclaimer

Exiger was retained to conduct research into the data subject via Exiger's due diligence software and open-source risk research. The information compiled in this report was gathered and assessed exclusively for the Client and according to the specifications of the terms of engagement between Exiger and the Client. This report is meant for the Client only. Distribution of this report or its contents, in whole or in part, to any third party is prohibited except as approved in the Agreement or by Exiger.

The report is provided as is. It is the responsibility of the report recipient to conduct further due diligence into information presented in the report and to assess this information independently in the light of your organization's policies and risk tolerance. Exiger accepts no liability for any decisions or actions made or not made based on the information in this report. All information contained in this report is based on information and analysis available to Exiger at the time of writing the report.



For Client Use Only — Proprietary

Anthropic PBC

EXECUTIVE SUMMARY

Anthropic PBC (“Anthropic”) is a software company founded in 2021 and headquartered in San Francisco, California.ⁱ Anthropic specializes in AI research.ⁱⁱ

DDIQ Profile	Anthropic PBC
Physical Address	500 Howard St., San Francisco, CA, 94104 ⁱⁱⁱ
CAGE Code(s)	9VKX2 ^{iv}
Entity Website	https://www.anthropic.com
Unique Entity Identifier(s)	SPQZL8XDKGK7 ^v
DUNS	-
BvDID	US374042333L ^{vi}
Active Business License(s)	Colorado ^{vii} , New York ^{viii} , among others
Employees	Approximately 2,000 ^{ix}
Annual Revenue	USD 14 billion (2025) ^x

OVERALL SUBJECT RISK: 4.3 — MEDIUM¹

Foreign Ownership, Control, or Influence (“FOCI”) Risk: MEDIUM-HIGH

Operational Risk: LOW-MEDIUM

Cyber Risk: MEDIUM

Compliance Risk: LOW-MEDIUM

Financial Health Risk: MEDIUM

¹ Discrepancies in this report’s risk scores, a subject’s adjudicated DDIQ profile, and a subject’s DDIQ Analytics dashboard risk scoring are a result of manual adjudication of risk events and open-source research verifying DDIQ results.



Anthropic PBC

For Client Use Only — Proprietary

KEY FINDINGS

Anthropic's adjudicated DDIQ profile, Exiger's proprietary risk model, and open-source research indicate that the overall risk the company represents is **MEDIUM**, owing to elevated scores for the elements **Operational and Compliance Risk**.

- Anthropic has been the victim of multiple data breaches and hacks since 2024, including by an alleged Chinese-state affiliated group (outlined in the **Operational Risk** section of this report).
- Anthropic is allegedly at risk of losing its contract with the US Department of Defense and the US Secretary of Defense threatened to label the company as a "supply chain risk" (outlined in the **Compliance Risk** section of this report).
- Anthropic has been the defendant in numerous lawsuits, resulting in the company paying billions in settlements (outlined in the **Compliance Risk** section of this report).

FOCI RISK — MEDIUM-HIGH

Anthropic is a private company located in San Francisco, California.^{xi} DDIQ and open-source information, including Anthropic's PitchBook, show that Anthropic has generated a total of USD 60.55 billion in funding and is owned 69.74% by investors as of February 26, 2026. Anthropic has 221 active investors and has received funding through four funding rounds.^{xii}

■	Investment Round	% Owned	# of Shares Authorized
■	Series F	7.1%	145,774,450
■	Series E	5.69%	95,281,934
■	Series B	6.73%	87,430,352

- On September 2, 2025, Anthropic completed a Series F funding round that was led by US-based investing company Lightspeed Management Company LLC (dba "Lightspeed Venture Partners"), Fidelity Management & Research Company ("FMR Co."), and ICONIQ Capital LLC ("ICONIQ"). Lightspeed Venture Partners and FMR Co. also took part in the Series G and E funding rounds, while ICONIQ participated in the Series G funding round. This investment round also included Qatar Investment Authority, Qatar's sovereign wealth fund.^{xiii}

According to Lightspeed Ventre Partner's 2025 Form ADV filed with the US Securities and Exchange Commision ("SEC") on March 27, 2025, it is more than 10% but less than 25% owned by members Barry Eggers, Ravi Mhatre ("Mhatre"), NIEH Peter (尼彼得



Anthropic PBC

For Client Use Only — Proprietary

, “Nieh”) and Christopher Schaepe.^{xiv} According to Mhatre’s LinkedIn-registered profile, he is also a current investor in Anthropic and is a board member of several entities based in foreign jurisdictions such as Israel-based network security company Cato Networks Ltd.^{xv} Additionally, according to Nieh’s LinkedIn-registered profile, he previously worked at Taiwan-based electronics company Acer Inc.^{xvi}

ICONIQ is owned 75% or more by Divesh Makan, its founding partner, according to its 2025 Form ADV filed with the SEC.^{xvii}

FMR Co. is a wholly owned subsidiary of FMR LLC (“FMR”), a US-based private assets management company.^{xviii} FMR is 49% ultimately beneficially owned by its director, chairman, and CEO, Abigail Johnson (“A. Johnson”), and members of the Johnson family, either directly or indirectly through trusts, according to FMR’s Form Schedule 13G/A filed with the SEC in March 2025.^{xix} According to Forbes, A. Johnson is a US citizen residing in the US.^{xx}

- On March 3, 2025, Anthropic completed a Series E funding round that was led by Lightspeed Venture Partners and raised USD 3.5 billion.

Research did not identify further material information.

According to their LinkedIn-registered profiles, the Key Management Personnel include the following:^{xxi}

Key Management Personnel

■	Name	Role on Board	Management Role	Citizenship
■	Dario Amodei	-	CEO	-
■	Mike Krieger	-	Chief Product Officer	-
■	Paul Smith	-	Chief Commercial Officer	-

- According to Anthropic’s CEO Dario Amodei’s (“Amodei”) LinkedIn-registered profile, from November 2014 to October 2015, he worked as a research scientist at China-based AI company Baidu Inc. in Sunnyvale, California.^{xxii}

- According to Anthropic’s Chief Commercial Officer Paul Smith’s (“Smith”) LinkedIn-registered profile, he previously worked at UK-based ad insertion company Yospace Technologies Ltd. and Cayman Islands-based software company Umee Technologies Ltd.^{xxiii} Additionally, Smith was the EVP and general manager in the UK for US-based software company Salesforce Inc.^{xxiv}



For Client Use Only — Proprietary

According to a third-party H-1B visa sponsorship data aggregator, Anthropic and its US-based subsidiaries sponsored the following visas to the US between 2023 and 2025:^{xxv}

Sponsorship Data

Type	Certified	Certified Withdrawn	Certified Expired	Citizenship
H-1B Visa	168	4	-	-
Green Card	0	-	-	-

Anthropic conducts operations in foreign jurisdictions such as South Korea. According to Anthropic's 2026 Bureau van Dijk profile, it has subsidiaries located in foreign jurisdictions such as Japan, France, Germany, and Ireland.^{xxvi} Additionally, in 2025, Anthropic established a subsidiary in South Korea.^{xxvii} Research did not identify further material information.

In 2025, Anthropic partnered with the UK government. In 2025, Anthropic signed a memorandum of understanding with the government of the UK to utilize AI tools to improve online access and public services for citizens of the UK.^{xxviii} Research did not identify further material information.

OPERATIONAL RISK — LOW-MEDIUM

Anthropic has been the victim of multiple data breaches and hacks. In February 2026, Anthropic accused three Chinese AI developers, Hangzhou DeepSeek Artificial Intelligence Co. Ltd. ("DeepSeek"), MiniMax Group Inc., and Beijing Moonshot AI Technology Co. Ltd. ("Moonshot"), of illegally extracting output from its Claude AI model to improve their own competing systems. DeepSeek reportedly has close ties to China's military and intelligence operations, with claims that it is state-controlled or state-subsidized.^{xxix} Anthropic says these firms generated more than 16 million interactions with Claude using thousands of fake accounts, a practice known in the industry as distillation, which it argues lets rivals shortcut costly development by learning directly from Claude's responses.^{xxx} Anthropic claimed that "approximately 16 million exchanged with its Claude model and 24,000 fake accounts" were used by the three companies.^{xxxi} According to a January 27, 2026, article published by KXAN, a US-based local news source, Moonshot was banned by Greg Abbott, the governor of Texas over security and data concerns.^{xxxii}

In November 2025, Anthropic disclosed that a Chinese state-sponsored group manipulated its Claude Code AI tool to run a large-scale, largely autonomous cyber espionage campaign in September 2025.^{xxxiii} The attackers bypassed Claude's safety guardrails through prompt manipulation, enabling the AI to automate about 80 to 90% of the offensive work against



Anthropic PBC

For Client Use Only — Proprietary

roughly 30 high-value targets, including tech firms, financial institutions, and government entities.^{xxxiv}

In August 2025, Anthropic disclosed that North Korea-linked hackers used the Claude code AI tool to obtain remote positions at US Fortune 500 companies.^{xxxv}

In December 2024, hackers gained control of an Anthropic social media account and used it to promote a counterfeit “CLAUDE” token, resulting in losses of about USD 100,000.^{xxxvi} Research did not identify further material information.

CYBER RISK – MEDIUM

SOFTWARE, HARDWARE, PERSONNEL, AND PHYSICAL SECURITY RISK

According to SecurityScorecard, an information security company that provides cybersecurity ratings, Anthropic has an overall cybersecurity score of 85 out of 100, leading to an overall grade of “B.”

The following table identifies the risk factors that contributed to lowering Anthropic’s cybersecurity rating:

Cybersecurity Risk

Risk Factor	Rating
IP Reputation ²	F
Network Security ³	B
Information Leak ⁴	B

SecurityScorecard noted 13,815 site vulnerabilities, 17 open ports, and four examples of malware discovered, but no examples of leaked information.

According to the National Institute of Standards and Technology (NIST) National Vulnerability Database, Anthropic has had four reported security vulnerabilities since 2025, including two since 2026 that scored a seven or higher out of 10.^{xxxvii}

² IP Reputation — IP reputation Leverages SecurityScorecard sinkhole infrastructure, OSINT malware feeds, and third-party threat intelligence.

³ Network Security — The network security module checks public datasets for evidence of high-risk or insecure open ports within the organization’s networks.

⁴ Information Leak — Information leak uses chatter and deep web monitoring to identify compromised credentials being circulated by hackers.



HIGH RISKS AND VULNERABILITIES

The following table identifies the high risks and vulnerabilities identified in Anthropic's open-source repositories:

Software List	High Risks ⁵	Vulnerabilities ⁶
anthropics/beam	1,249	19
anthropics/argo-cd	1,239	3
anthropics/anthropic-sdk-ty-pescript	1,074	1
anthropics/PySvelte	489	22
anthropics/anthropic-tokenizer-typescript	383	5
anthropics/rclone	140	0
anthropics/riv2025-long-horizon-coding-agent-demo	115	1
anthropics/s5cmd	22	0
anthropics/claude-agent-sdk-demos	19	1
anthropics/anthropic-sdk-ruby	19	1
anthropics/anthropic-sdk-go	13	0
anthropics/orjson	11	0
anthropics/anthropic-retrieval-demo	10	8
anthropics/github-mcp-server	9	0
anthropics/claude-code-action	4	1
anthropics/torchtyping	3	0
anthropics/anthropic-tools	3	4
anthropics/python-tblib	2	0
anthropics/claude-code-security-review	2	1
anthropics/financial-services-plugins	2	1
anthropics/claude-code-base-action	2	1
anthropics/redis-py	2	3
anthropics/skills	2	9
anthropics/httpcore	1	1

COMPLIANCE RISK — LOW-MEDIUM

Anthropic software was used in an operation overseas. According to a February 15, 2026, article published by Yeni Safak, a Turkey-based online news source, Anthropic's Claude AI software was used in the January 3, 2026, operation that captured Nicolás Maduro, the president of Venezuela. Claude AI was accessed through Anthropic's partnership with Palantir.^{xxxviii} According to a February 25, 2026, article published by Al Jazeera English, an online news source, Claude AI is subject to Anthropic's usage policy, which includes regulations against using Claude for surveillance, the development of weapons, or inviting weapons.

⁵ Exiger Cyber's risk assessment outlines common risks by applying statistical methods grounded in a deep understanding of software development metrics and distinguishing between normal and abnormal patterns. The severity of risks is rated as high, medium, or low depending on the gravity of individual rule violations or the total number of failing rules.

⁶ Exiger Cyber's vulnerability assessment leverages data from the National Vulnerability Database to identify potential vulnerabilities. The assessment relies on precise product identifiers within software components to accurately match and detect known security weaknesses.



For Client Use Only — Proprietary

Anthropic declined to comment on the usage.^{xxxix} Research did not identify any additional material information.

Anthropic is allegedly at risk of losing its US Government contract. According to a February 24, 2026, article published by Head Topics, an online news source, Anthropic could lose its government contract if the company fails to allow the US Department of Defense (“DoD”) to use its AI technology for unrestricted military use. The article stated that Secretary of Defense Pete Hegseth warned Amodeli that the company could be designated a supply chain risk if company policy did not align with the DoD’s usage.^{xi} Research did not identify any additional material information.

Anthropic was accused of stealing training data. According to a February 24, 2026, article published by Digit Fast Track, an online news source, Anthropic was accused of stealing training data by Elon Musk (“Musk”), a US businessman. Musk’s accusations came after Anthropic accused three China-based AI companies of using Claude to improve their own AI models.^{xii} Research did not identify any additional material information.

Anthropic’s Claude can be exploited and leak sensitive data. According to an October 31, 2025, article published by Techradar, an online technology news source, Claude can be exploited by hackers. The article stated that Claude can be tricked into uploading sensitive data when prompted to run code within a conversation. Hackers can then gain access to downloaded software packages. Users have been advised to limit network communications to their own account or disable network access to prevent data leaks.^{xiii} Research did not identify any additional material information.

Anthropic has been the defendant in numerous lawsuits. Anthropic was sued by major music publishers, including the Netherlands-based Universal Music Group NV, US-based Concord Music Group Inc. (“Concord”), and US-based ABKCO Music & Records Inc., which alleged the company used copyrighted lyrics and sheet music from more than 700 identified songs, and potentially over 20,000 works in total, without authorization to train its Claude AI model in 2026 as an extension of the original lawsuit filed in October 2023.^{xiiii} The plaintiffs are seeking statutory damages that could exceed USD 3 billion, describing the case as potentially one of the largest non-class action copyright lawsuits in US history, while Anthropic denies the allegations and maintains its training practices comply with the law.^{xlv} The case with Concord is continuing through normal federal litigation processes and has not yet been resolved or dismissed.^{xlv}

In 2025, Anthropic was the defendant in a lawsuit filed by a group of book authors who alleged that the company’s Claude AI platform infringed on the authors’ copyrights by copying and using their books to train the Claude AI platform. A judge ruled that Anthropic’s use of legally purchased books, including print copies it digitized, to train its language models was transformative and protected under fair use, so it did not constitute copyright infringement on that basis; however, the court found that retaining millions of pirated books from unauthorized sources was not protected by fair use, and that dispute ultimately led to a late-2025 settlement in which Anthropic agreed to pay about USD 1.5 billion, with the agreement receiving preliminary court approval.^{xlvi}



Anthropic PBC

For Client Use Only — Proprietary

In June 2025, US-based social news company Reddit Inc. sued Anthropic over an alleged breach of contract. Reddit alleged that Anthropic improperly used its content to train Anthropic's AI systems without permission.^{xlvii}

An IT firm based in Belagavi, Anthropic Software Private Limited ("ASPL"), has filed a civil suit against Anthropic, alleging that the company's use of the same name has harmed its brand, diverted online traffic, and caused financial losses. ASPL, which says it registered its name in India in 2017, before Anthropic was established, claims passing off, misrepresentation, and brand erosion and is seeking to bar Anthropic from operating under that name in India along with damages.^{xlviii} The court allowed the lawsuit to proceed and issued summons to Anthropic PBC; after the defendant's representatives failed to appear on an earlier date, fresh summons have been issued to Anthropic India Pvt Ltd in Bengaluru for a hearing set on March 9, 2026.^{xlix} Research did not identify further material information.

COMMERCE RISK

Between 2021 and 2026, Anthropic received approximately one shipment from an international supplier. The following table includes the top supplier country.

Shipment Data

Country	Number of Shipments	Percentage of Shipments
UK	1	100%

Anthropic's corporate supplier was Qumpus Inc. (dba Better World Books), a US-based company specializing in books and e-commerce.ⁱ

FINANCIAL HEALTH RISK — MEDIUM

The following table represents Anthropic's commercial credit scores:ⁱⁱ

Commercial Credit Scores

Source	Category	Rating
Cortera	Commercial Credit Risk Rating	B



Anthropic PBC

For Client Use Only — Proprietary

Cortera	Payment Risk Segment	1 (Higher payment risk, trending down) ⁷
---------	----------------------	---

According to USAspending, Anthropic has received USD 18,960 in US federal prime contracts, grants, and loans since fiscal year 2008, all of which was with the US Department of State.ⁱⁱⁱ This data pertains to prime contracts awarded to Anthropic and may not capture products developed by Anthropic and redistributed by third parties.

The following table represents the top major contract awarded to Anthropic by government contract organizations:

US Government Contracts Data

Government Organization	Award Amount (USD)
US Department of State	USD 18,960

END REPORT

ⁱ <https://www.anthropic.com/ddiq.efc.exigerfed.com/ddiq-services/rest/bvdReport/US374042333L?entityId=207595844632576>

ⁱⁱ <https://www.anthropic.com/>

ⁱⁱⁱ <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bvdReport/US374042333L?entityId=207595844632576>

^{iv} <https://cage.dla.mil/Search/Details?id=19488082>

^v <https://cage.dla.mil/Search/Details?id=19488082>

^{vi} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bvdReport/US374042333L?entityId=207595844632576>

^{vii} https://opencorporates.com/companies/us_co/20241311617

^{viii} https://opencorporates.com/companies/us_ny/6565058

^{ix} <https://pitchbook.com/profiles/company/466959-97#overview>

^x <https://siliconangle.com/2026/02/12/anthropic-closes-30b-round-annualized-revenue-tops-14b/>

^{xi} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bvdReport/US374042333L?entityId=207595844632576>

^{xii} <https://pitchbook.com/profiles/company/466959-97>

^{xiii} <https://pitchbook.com/profiles/company/466959-97#overview>

<https://www.qia.qa/>

^{xiv} https://files.adviserinfo.sec.gov/IAPD/content/viewform/adv/Sections/iapd_AdvScheduleASection.aspx?ORG_PK=160187&FLNG_PK=044CB0FE000801E401697B11059F1935056C8CC0

^{xv} <https://www.linkedin.com/in/ravimhatre/details/experience/>

⁷ Payment Risk Segment of 1 — Higher payment risk, trending down. Cortera's Payment Risk Segment is based on how fast the company is paying vendors and whether the payments are getting faster or slower. The Payment Risk Segment is based on a scale of 1-6.



Anthropic PBC

For Client Use Only — Proprietary

^{xvi} <https://www.linkedin.com/in/peternieh/>

^{xvii} https://files.adviserinfo.sec.gov/IAPD/content/viewform/adv/Sections/iapd_AdvScheduleBSection.aspx?ORG_PK=159198&FLNG_PK=028FE254000801EE04DE0862004AA729056C8CC0

<https://www.sfbuzz.org/about/board-of-trustees/divesh-makan/>

^{xviii} [https://www.fidelity.com/bin-public/060_www_fidelity_com/documents/FMR-ADV-FPWA.pdf#:~:text=Fidelity%20Management%20&%20Research%20Company%20\(“FMR%20Co.”\)%20reorganized%20into%20FMR%20effective%20January%201%202020.](https://www.fidelity.com/bin-public/060_www_fidelity_com/documents/FMR-ADV-FPWA.pdf#:~:text=Fidelity%20Management%20&%20Research%20Company%20(“FMR%20Co.”)%20reorganized%20into%20FMR%20effective%20January%201%202020.)

^{xix} <https://www.sec.gov/Archives/edgar/data/1850838/000031506625000984/exhibit99.txt>

^{xx} <https://www.forbes.com/profile/abigail-johnson/?sh=66106e251c42>

^{xxi} <https://www.linkedin.com/in/mikekrieger/>

<https://www.linkedin.com/in/dario-amodei-3934934/>

<https://www.linkedin.com/in/pjsmith/>

^{xxii} <https://www.linkedin.com/in/dario-amodei-3934934/>

^{xxiii} <https://www.linkedin.com/company/umee-network/about/>

^{xxiv} <https://www.linkedin.com/in/pjsmith/>

^{xxv} <https://www.myvisajobs.com/employer/anthropic-pbc/>

^{xxvi} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bvdReport/US374042333L?entityId=207595844632576>

^{xxvii} <https://www.businesskorea.co.kr/news/articleView.html?idxno=249227>

^{xxviii} <https://www.biometricupdate.com/202502/uk-govt-signs-mou-with-anthropic-as-digital-id-ai-become-economic-issues>

^{xxix} <https://www.cnn.com/2025/06/24/deepseek-ai-china-military-and-evaded-export-controls-us-official-says-reuters.html#:~:text=Since%202022%20those%20chips%20have,a%20Reuters%20request%20for%20comment.&text=%22We%20do%20not%20support%20parties,by%20competitors%20such%20as%20Huawei.%22>

<https://www.congress.gov/crs-product/IF13051>

^{xxx} <https://www.bloombergtechnoz.com/detail-news/100527/anthropic-deepseek-cs-ekstraksi-hasil-dari-claude-secara-ilegal>

^{xxxi} <https://www.taipeitimes.com/News/biz/archives/2026/02/25/2003852826>

^{xxxii} <https://www.kxan.com/news/texas/texas-adds-26-china-affiliated-technology-companies-to-banned-list/>

^{xxxiii} <https://www.kanalcoin.com/anthropic-ai-cyber-espionage-attack/>

^{xxxiv} <https://bitcoinethereumnews.com/tech/anthropic-detects-potential-first-ai-led-cyberattack-by-chinese-group-using-claude/>

^{xxxv} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bca/article/urn:bca:oiq.bca.model.LexisJsonRecord:5bcb69198ed8947d0459f5928d71aae7eb?Inid=6GRP-KM73-RS3C-71VC-00000-00>

^{xxxvi} <https://www.cryptotimes.io/2024/12/26/animoca-chairmans-x-account-hacked-to-promoting-fake-token/>

^{xxxvii} <https://nvd.nist.gov/vuln/search#/nvd/home?keyword=anthropic&resultType=records>

^{xxxviii} <https://en.yenisafak.com/technology/us-military-used-anthropics-claude-ai-in-maduro-capture-operation-3714626>

^{xxxix} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bca/article/urn:bca:oiq.bca.model.LexisJsonRecord:3558920e44a1e43504b0193c6910e4082e?Inid=6JOF-WBD3-RTTP-WOGW-00000-00>

^{xl} <https://us.headtopics.com/news/hegseth-warns-anthropic-to-let-the-military-use-the-80212130>

^{xli} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bca/article/urn:bca:oiq.bca.model.LexisJsonRecord:01ba9db2b8d1565a0417b48ec90b9b218d?Inid=6JOF-7JC3-RS9G-92C2-00000-00>



Anthropic PBC

For Client Use Only — Proprietary

^{xlii} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bca/article/urn:bca:oiq.bca.model.LexisJsonRecord:f6fa725637d00733047837a81882f32891?Inid=6H3J-JPK3-RS1H-N380-00000-00>

^{xliii} <https://www.musicbusinessworldwide.com/anthropic-must-face-music-publishers-copyright-claims-after-judge-denies-dismissal-motion/>

^{xliv} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bca/article/urn:bca:oiq.bca.model.LexisJsonRecord:74e4ea163085a38504f53165214ed9ea6c?Inid=6HT2-KNS3-SJ7N-J15S-00000-00>

^{xlv} <https://mpost.io/anthropic-faces-3b-copyright-lawsuit-from-major-music-publishers-over-ai-training-practices/>

^{xlvi} <https://www.loeb.com/en/insights/publications/2025/07/bartz-v-anthropic-pbc>

<https://www.deep-lex.com/blog/bartz-v-anthropic>

^{xlvii} <https://ddiq.efc.exigerfed.com/dashboard/#/cached-article?url=https:%2F%2Fddiq.efc.exigerfed.com:443%2Fddiq-services%2Frest%2FoiqPrivate%2Farticle%3Fid%3D207595844632576%26url%3Dhttps%253A%252F%252Fwww.clydeco.com%252Fen%252Finsights%252F2025%252F07%252Fnavigating-ai-risks>

^{xlviii} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bca/article/urn:bca:oiq.bca.model.LexisJsonRecord:d606d82e348b9cad0481e727c59f3034c4?Inid=6HWP-7B63-RRHW-22NX-00000-00>

^{xlix} https://timesofindia.indiatimes.com/legal/news/belagavi-court-issues-fresh-summons-to-us-ai-firm-anthropic/articleshow/128415251.cms?https://support.betterworldbooks.com/hc/en-us/articles/200801763-Sales-Tax-Exemption-Process#:~:text=Better%20World%20Books%2C%20DBA%20Qumpus%2C%20Inc.%2C%20is,your%20phone%20number%20*%20Your%20order%20number

<https://www.linkedin.com/company/better-world-books/>

^{li} <https://ddiq.efc.exigerfed.com/ddiq-services/rest/bvdReport/US374042333L?entityId=207595844632576>

^{lii} <https://www.usaspending.gov/search?hash=229167627d88222a690619ea1d7c18d4>

Enclosure
Section 3252 Scoping Analysis for Anthropic, PBC

This document supports the use of section 3252 of title 10, United States Code ("Section 3252") authorities, as implemented by subpart 239.73 of title 48, Code of Federal Regulations (C.F.R.), "Requirements for Information Relating to Supply Chain Risk." The Under Secretary of War for Acquisition and Sustainment and the Department of War Chief Information Officer have determined the following as the appropriate scope for use of Section 3252 authorities necessary to address the supply chain risk related to the use of covered products or services of Anthropic, PBC:

- **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 Entity").
- **Covered Products or Services:** All of the Covered Entity's products or services that meet the definition of a "covered item of supply" or that would be procured as part of a "covered system," as those terms are defined at 48 C.F.R. § 239.7301, 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 for which 48 C.F.R. Subpart 239.73 is applicable, in accordance with 48 C.F.R. § 239.7302.
- **Covered Procurement Actions:** All actions authorized in accordance with 48 C.F.R. § 239.7305.



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

MAR - 3 2026

The Honorable Mike D. Rogers
Chairman
Committee on Armed Services
U.S. House of Representatives
Washington, DC 20515

Dear Mr. Chairman:

In accordance with section 3252 of Title 10 United States Code ("section 3252"), I am notifying you of my determination that the use of products or services of Anthropic, PBC, including its subordinate, subsidiary, and affiliated offices or entities, doing business under various names, and all successors or assigns thereof (the "Covered Entity") in Department of War (DoW) covered systems¹ presents a significant supply chain risk and that the use of section 3252 authorities is necessary to protect our national security. This determination is based in part on a risk analysis by the DoW and input from senior DoW personnel that the Covered Entity's restrictions on the use of its products and services introduces national security risks to the DoW's supply chain.

I have consulted with procurement and other relevant officials within DoW regarding the Covered Entity, and determined that (i) use of the authority provided in section 3252 to carry out covered procurement actions² 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. Accordingly, this letter is providing notice of a section 3252 determination regarding the products and services of the Covered Entity.

I am sending identical letters to Congress and the appropriate congressional committees.

Sincerely,

A handwritten signature in black ink, appearing to be "P.B. Jh", is located below the "Sincerely," text.

cc:
The Honorable Adam Smith
Ranking Member

¹ Section 3252(d)(5).

² Section 3252(d)(2).



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

MAR - 3 2026

The Honorable Roger F. Wicker
Chairman
Committee on Armed Services
United States Senate
Washington, DC 20510

Dear Mr. Chairman:

In accordance with section 3252 of Title 10 United States Code ("section 3252"), I am notifying you of my determination that the use of products or services of Anthropic, PBC, including its subordinate, subsidiary, and affiliated offices or entities, doing business under various names, and all successors or assigns thereof (the "Covered Entity") in Department of War (DoW) covered systems¹ presents a significant supply chain risk and that the use of section 3252 authorities is necessary to protect our national security. This determination is based in part on a risk analysis by the DoW and input from senior DoW personnel that the Covered Entity's restrictions on the use of its products and services introduces national security risks to the DoW's supply chain.

I have consulted with procurement and other relevant officials within DoW regarding the Covered Entity, and determined that (i) use of the authority provided in section 3252 to carry out covered procurement actions² 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. Accordingly, this letter is providing notice of a section 3252 determination regarding the products and services of the Covered Entity.

I am sending identical letters to Congress and the appropriate congressional committees.

Sincerely,

A handwritten signature in black ink, appearing to read "P.B. Jh.", is located below the "Sincerely," text.

cc:
The Honorable Jack Reed
Ranking Member

¹ Section 3252(d)(5).

² Section 3252(d)(2).



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

MAR - 3 2026

The Honorable Susan M. Collins
Chair
Committee on Appropriations
United States Senate
Washington, DC 20510

Dear Madam Chair:

In accordance with section 3252 of Title 10 United States Code ("section 3252"), I am notifying you of my determination that the use of products or services of Anthropic, PBC, including its subordinate, subsidiary, and affiliated offices or entities, doing business under various names, and all successors or assigns thereof (the "Covered Entity") in Department of War (DoW) covered systems¹ presents a significant supply chain risk and that the use of section 3252 authorities is necessary to protect our national security. This determination is based in part on a risk analysis by the DoW and input from senior DoW personnel that the Covered Entity's restrictions on the use of its products and services introduces national security risks to the DoW's supply chain.

I have consulted with procurement and other relevant officials within DoW regarding the Covered Entity, and determined that (i) use of the authority provided in section 3252 to carry out covered procurement actions² 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. Accordingly, this letter is providing notice of a section 3252 determination regarding the products and services of the Covered Entity.

I am sending identical letters to Congress and the appropriate congressional committees.

Sincerely,

A handwritten signature in black ink, appearing to be "PBJ" followed by a flourish, is located below the "Sincerely," text.

cc:
The Honorable Patty L. Murray
Vice Chair

¹ Section 3252(d)(5).

² Section 3252(d)(2).



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

MAR - 3 2026

The Honorable Tom J. Cole
Chairman
Committee on Appropriations
U.S. House of Representatives
Washington, DC 20515

Dear Mr. Chairman:

In accordance with section 3252 of Title 10 United States Code ("section 3252"), I am notifying you of my determination that the use of products or services of Anthropic, PBC, including its subordinate, subsidiary, and affiliated offices or entities, doing business under various names, and all successors or assigns thereof (the "Covered Entity") in Department of War (DoW) covered systems¹ presents a significant supply chain risk and that the use of section 3252 authorities is necessary to protect our national security. This determination is based in part on a risk analysis by the DoW and input from senior DoW personnel that the Covered Entity's restrictions on the use of its products and services introduces national security risks to the DoW's supply chain.

I have consulted with procurement and other relevant officials within DoW regarding the Covered Entity, and determined that (i) use of the authority provided in section 3252 to carry out covered procurement actions² 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. Accordingly, this letter is providing notice of a section 3252 determination regarding the products and services of the Covered Entity.

I am sending identical letters to Congress and the appropriate congressional committees.

Sincerely,

A handwritten signature in black ink, appearing to be "PBJ" followed by a flourish, is located below the "Sincerely," text.

cc:
The Honorable Rosa L. DeLauro
Ranking Member

¹ Section 3252(d)(5).

² Section 3252(d)(2).



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

MAR - 3 2026

The Honorable Rick Crawford
Chairman
Permanent Select Committee on Intelligence
U.S. House of Representatives
Washington, DC 20515

Dear Mr. Chairman:

In accordance with section 3252 of Title 10 United States Code ("section 3252"), I am notifying you of my determination that the use of products or services of Anthropic, PBC, including its subordinate, subsidiary, and affiliated offices or entities, doing business under various names, and all successors or assigns thereof (the "Covered Entity") in Department of War (DoW) covered systems¹ presents a significant supply chain risk and that the use of section 3252 authorities is necessary to protect our national security. This determination is based in part on a risk analysis by the DoW and input from senior DoW personnel that the Covered Entity's restrictions on the use of its products and services introduces national security risks to the DoW's supply chain.

I have consulted with procurement and other relevant officials within DoW regarding the Covered Entity, and determined that (i) use of the authority provided in section 3252 to carry out covered procurement actions² 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. Accordingly, this letter is providing notice of a section 3252 determination regarding the products and services of the Covered Entity.

I am sending identical letters to Congress and the appropriate congressional committees.

Sincerely,

A handwritten signature in black ink, appearing to be "P.B. Jh", is located below the "Sincerely," text.

cc:
The Honorable Jim Himes
Ranking Member

¹ Section 3252(d)(5).

² Section 3252(d)(2).



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

MAR - 3 2026

The Honorable Tom Cotton
Chairman
Select Committee on Intelligence
United States Senate
Washington, DC 20510

Dear Mr. Chairman:

In accordance with section 3252 of Title 10 United States Code ("section 3252"), I am notifying you of my determination that the use of products or services of Anthropic, PBC, including its subordinate, subsidiary, and affiliated offices or entities, doing business under various names, and all successors or assigns thereof (the "Covered Entity") in Department of War (DoW) covered systems¹ presents a significant supply chain risk and that the use of section 3252 authorities is necessary to protect our national security. This determination is based in part on a risk analysis by the DoW and input from senior DoW personnel that the Covered Entity's restrictions on the use of its products and services introduces national security risks to the DoW's supply chain.

I have consulted with procurement and other relevant officials within DoW regarding the Covered Entity, and determined that (i) use of the authority provided in section 3252 to carry out covered procurement actions² 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. Accordingly, this letter is providing notice of a section 3252 determination regarding the products and services of the Covered Entity.

I am sending identical letters to Congress and the appropriate congressional committees.

Sincerely,

A handwritten signature in black ink, appearing to be "PBJ" followed by a flourish, is located below the "Sincerely," text.

cc:
The Honorable Mark Warner
Vice Chairman

¹ Section 3252(d)(5).

² Section 3252(d)(2).



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 10, United States Code (U.S.C.), section 3252 ("Section 3252"), the Department of War (DoW) has determined that (i) the use of the Covered Entity's products or services in DoW covered systems¹ presents a supply chain risk and that the use of the Section 3252 authority to carry out covered procurement actions² 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 item of supply" or that would be procured as part of a "covered system," as those terms are defined at 48 C.F.R. § 239.7301, 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 for which 48 C.F.R. Subpart 239.73 is applicable, in accordance with 48 C.F.R. § 239.7302.
- **Covered Procurement Actions:** All actions authorized in accordance with 48 C.F.R. § 239.7305.

¹ 10 U.S.C. § 3252(d)(5)

² 10 U.S.C. § 3252(d)(2)

Effective Date

This Determination is effective immediately and shall remain in effect until modified or terminated in writing by the Section 3252 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 3252 Determination.

Sincerely,

A handwritten signature in black ink, appearing to be "PBJ" followed by a long horizontal stroke.

Enclosure:
As stated

ATTACHMENT 1**Requirements and Procedures for Requesting Reconsideration of a Section 3252 Determination**

The following requirements and procedures govern a Covered Entity's request for reconsideration of a Section 3252 determination:

1. **Notice of Request for Reconsideration:** Within thirty (30) days of receiving the Section 3252 Authorized Official's letter notifying the Covered Entity of the Section 3252 determination, the Covered Entity may submit in writing to the Section 3252 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 3252 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 3252 Authorized Official within the thirty (30) day period. The Section 3252 Authorized Official may extend the time to appear and submit documentary evidence upon written request by the Covered Entity.
3. **Flexible Procedures:** The Section 3252 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 3252 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 3252 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 3252 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

UNCLASSIFIED

considered. The representative(s) of the Section 3252 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.

UNCLASSIFIED

ANT_AR-0240



CHIEF INFORMATION OFFICER

DEPARTMENT OF WAR

6000 Defense Pentagon
Washington, D.C. 20301-6000

March, 5 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:

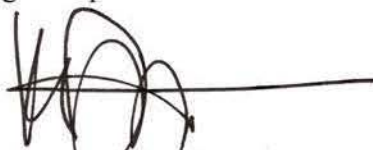
- National security systems (NSS), strategic priorities, and nuclear weapons
- Nuclear command, control, and communications
- Continuity of government
- Ballistic missile defense (references a-f)
- Warfighting capabilities at high-risk cyber survivability levels (Categories 3-5), per reference (k)

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.

A handwritten signature in black ink, appearing to read 'Kirsten A. Davies', with a long horizontal line extending to the right.

Kirsten A. Davies

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

ANTHROPIC PBC,

Plaintiff,

v.

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

Defendants.

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

DECLARATION OF EMIL MICHAEL

Pursuant to 28 U.S.C. § 1746, I, Emil Michael, declare as follows:

1. I am the Under Secretary of War for Research and Engineering (USW(R&E)) and Chief Technology Officer for the Department of War (DoW). I have held this position since May 20, 2025.

2. In my current position, I am responsible for spearheading the Department's efforts to ensure U.S. military technological superiority and keep DoW at the forefront of innovation. I provide strategic direction and oversight for DoW's entire research, development, and prototyping enterprise, which includes providing critical input on the acquisition, implementation, and use of cutting-edge technologies such as artificial intelligence (AI).

3. This declaration is based on my personal knowledge as well as information made available to me through reasonable diligence in the course of my official duties.

DoW's Title 10, Section 3252 Authorities

4. Organized under Title 10 of the United States Code, DoW is the largest government agency of the United States. DoW oversees the United States' armed services and coordinates the national defense. In service of the national defense, DoW awards contracts to and sets terms and policies with various entities that supply the Department with the technologies needed to advance U.S. military and national defense capabilities.

1 5. As part of its acquisition and procurement authorities under 10 U.S.C. § 3252, DoW
2 conducts supply chain risk assessments of covered procurements involving covered systems and
3 covered items of supply, as defined in that section. 10 U.S.C. § 3252(d)(3), (4). If DoW determines
4 that there is a significant supply chain risk to a covered system, the Secretary of War is authorized to
5 take covered procurement actions, as defined by Section 3252, to exclude the source of the risk from
6 covered systems to protect national security.

7 6. Under Section 3252's implementing regulations, this exclusion authority may be
8 exercised by the Secretary of War only after "[o]btaining a joint recommendation by the Under
9 Secretary of War for Acquisition and Sustainment and the Chief Information Officer of the Department
10 of War, on the basis of a risk assessment by the Under Secretary of War for Intelligence, that there is a
11 significant supply chain risk to a covered system." 48 C.F.R. § 239.7304. The Office of the Under
12 Secretary of War for Intelligence and Security¹ (USW(I&S)) ordinarily would have had responsibility
13 for conducting such supply chain risk assessments because of its general expertise in handling security
14 matters and its prior relationships with the information offices of the Department, including the Office
15 of the Chief Information Officer, which reported to USW(I&S) until 2012. But after a recent
16 reorganization within DoW, the subject matter expertise for conducting certain risk assessments is no
17 longer located solely within USW(I&S).

18 **DoW's 2025 Reorganization**

19 7. Prior to 2025, the Chief Digital and Artificial Intelligence Office (CDAO), founded in
20 2022, was a standalone function that reported directly to the Deputy Secretary of War. CDAO's
21 mission is to accelerate DoW's adoption of AI to ensure that our warfighters have the best capabilities
22 to dominate on the battlefield and have full trust that the technologies they are using to create decision
23 advantage are secure and trustworthy.

24 8. In 2025, DoW underwent a reorganization. As part of the reorganization, CDAO was
25 realigned and began to report to the Under Secretary of War for Research and Engineering. Because of
26 its institutional and subject matter expertise, CDAO inherited responsibility for Section 3252 supply
27

28 ¹ In the FY 2020 NDAA, the Under Secretary for Intelligence was renamed as the Under Secretary for Intelligence and Security to better capture the scope of the authorities vested in the office.

1 chain risk assessments relating to AI issues. As a result, my office has assumed responsibility for
2 providing CDAO's supply chain risk assessments relating to AI issues to the Under Secretary of War
3 for Acquisition and Sustainment and the Chief Information Officer in accordance with 48 C.F.R.
4 § 249.7304. In addition to CDAO possessing the relevant authorities, CDAO and my office are DoW's
5 subject matter experts for AI issues and are therefore best able to provide the comprehensive risk
6 assessment that informs the determinations under Section 3252. As appropriate, my office and CDAO
7 also coordinate AI risk assessments with the Chief Information Officer.

8 **Supply Chain Risk and Harms to National Security**

9 9. As outlined in the Urgent Supply Risk Analysis (the "Analysis") provided to the
10 Secretary of War, Anthropic PBC has become a supply chain risk following a progression of risk that
11 reached a saturation point as a result of the behavior of its leadership during the course of contract
12 negotiations with DoW in late 2025 and early 2026. As explained in the Analysis, the relatively opaque
13 nature of large language model (LLM) technology that DoW procures from Anthropic creates a
14 baseline risk. That risk escalated due to the unusual degree of control that Anthropic retains over the
15 model, as well as Anthropic's adversarial posture towards DoW's statutory mission and the manner in
16 which it is conducted. This technical opacity makes it difficult for DoW to assess technological
17 features that may be encoded into the LLM product and that may cause it to subvert the appropriate
18 execution of mission applications, also known as "model poisoning," or to fail to perform altogether.
19 While this is, at least in part, a common concern with all LLMs, the risk is significantly elevated in this
20 instance by the actions of Anthropic's leadership, detailed below.

21 10. In addition, the federal government has identified AI as a field that requires technology
22 transfer restrictions, per the Technology Alert list. Anthropic employs a large number of foreign
23 nationals to build and support its LLM products, including many from the Peoples Republic of China
24 (PRC), which increases the degree of adversarial risk should those employees comply with the PRC's
25 National Intelligence Law. Although other major U.S. AI labs that provide LLM products to DoW may
26 present similar risks, the technical and security assurances of the other labs' leadership, along with their
27 consistently responsible and trustworthy behavior during their engagement with DoW, mitigate these
28 risks. Anthropic's case, however, is different. A series of additional risks came to light in 2026, when

1 DoW and the company engaged in contract negotiations to expand DoW's use of Anthropic's LLM
2 products.

3 11. First, Anthropic's leadership demonstrated an intent to prevent the U.S. military's lawful
4 use of their LLM product, Claude, despite the company's publicly stated knowledge that adversarial
5 nation states have a practice of stealing Anthropic's LLM technology for their own unrestricted use.²
6 This asymmetrical reality, imposed by Anthropic, disadvantages the U.S. military vis-à-vis its
7 adversaries. During the 2026 contract negotiations, Anthropic's leadership insisted on multiple redlines
8 that it would not allow the U.S. military to cross when using Claude. The company's leadership
9 insisted on imposing restrictions on DoW's lawful military capability development, operations, and
10 intelligence missions, even though it would impair the capabilities of the U.S. military relative to our
11 adversaries. In short, Anthropic made clear that it will not allow the Government to deploy Claude for
12 multiple lawful uses. Determinations about lawful military uses, however, must rest solely with DoW
13 and not with a private company.

14 12. Second, Anthropic's leadership confirmed in an internal company memorandum
15 published in February that the company sought to impose multiple restrictions over the Government's
16 lawful use of Claude, including safety mechanisms that may be outside the control of DoW.³

17 13. Third, the company's leadership demonstrated bad faith by sharing with the press
18 unclassified but sensitive details of private conversations with DoW leadership in order to exert public
19 pressure on DoW to concede to Anthropic's demands.

20 14. Fourth, the Department learned that in 2025, the U.S. Centers for Disease Control's
21 (CDC) lawful use of Anthropic's LLM technology to support its infectious disease prevention research
22 mission was limited by Anthropic's use of safety filters in the LLM product CDC was using. The
23 company did not inform the agency of these filters, and they caused the product to stop functioning
24 normally for various sensitive, but research-aligned queries.

25
26
27 ² <https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks>.

28 ³ <https://www.theinformation.com/articles/read-anthropic-ceos-memo-attacking-openais-mendacious-pentagon-announcement>.

1 15. Fifth, the Department learned that during an active overseas military operation, an
2 Anthropic executive expressed concern to one of DoW's primary operational support software vendors
3 about the potential use of Anthropic's LLM products by U.S. military analysts during the operation.
4 The Department was made aware of this conversation between cleared individuals by the primary
5 vendor. During later discussions with Anthropic leaders, not all of whom have the requisite security
6 clearances, an Anthropic executive repeated this information raising serious concerns about their
7 processes and procedures for operational security. The same information subsequently appeared in the
8 news media. In light of these incidents, it is reasonably likely that Anthropic's leadership would alter or
9 even shut off DoW's use of Claude if Anthropic believes that the model may be used for purposes it
10 deems, in its sole discretion, to extend beyond the company's unilaterally imposed boundaries before or
11 during a military operation, which could endanger the lives of U.S. military personnel and civilians and
12 compromise the United States' warfighting mission. Continuing to use Anthropic's technology under
13 the current contract structures in any echelon in DoW's supply chain, namely the covered systems, thus
14 presents a significant risk.

15 16. Taken together, this collection of risks demonstrates the clear technical capability and
16 adversarial intent for Anthropic's leadership to potentially undermine lawful U.S. national security
17 activities and objectives. Anthropic leadership's adversarial behavior has elevated the supply chain
18 risks to a saturation point. DoW uses Anthropic's model in multiple ways, including in ongoing
19 military operations. If Anthropic were to interfere during an operation, whether by shutting off access
20 to the model or altering its functionality, such interference could cause serious harm to national security
21 and loss of human life. This risk within a covered system is intolerable and warrants the designation
22 under 10 U.S.C. § 3252.

23 17. This risk is not limited only to Anthropic and its model's standalone presence in DoW
24 systems or as a subcontractor to DoW. The model's interactions with other technology and covered
25 systems create additional risk to the DoW supply chain. When Anthropic's model is layered into other
26 applications, there is a substantial risk that any company-imposed restrictions or alterations to the
27 model would be transferred and impact mission applications, including in weapons systems
28 development and other products or services that ultimately perform DoW activities.

1 18. As an example, if Anthropic's technology is used as a plug-in to a larger application, it
2 may limit the functionality of that larger system to the internal limitations built into or added to the
3 Anthropic system. This would directly impair other covered systems by reducing their functionality to
4 the same level as Anthropic's system.

5 19. AI is functionally a tool to assist DoW in its national security mission. It is imperative
6 that DoW be able to fully trust the functionality of its tools. Here, there are significant concerns due to
7 Anthropic's demonstrated willingness to modify or restrict its model's functionality for DoW purposes.
8 All lawfulness determinations are vested with DoW, which ensures the integrity of the chain of
9 command, especially during active combat operations. Anthropic's demonstrated willingness to
10 interfere with that chain of command is a significant risk.

11 20. In assessing these significant supply chain risks and harms to national security, DoW
12 considered whether less restrictive means than exclusion and removal could mitigate the supply chain
13 risk and national security harm. While each risk identified above may not, standing alone, have
14 necessitated exclusion and removal of plaintiff from DoW's supply chain, when considered in the
15 aggregate, a significant supply chain risk exists. The only potential mitigation to this collective set of
16 risks—acquisition of LLM products with the usage terms and technical and service delivery
17 specifications DoW requires—was not an option to which Anthropic would agree.

18 21. These risks and possible mitigation options were considered in the aggregate and in light
19 of the escalating tension over the key differences concerning authority to determine DoW's lawful use
20 of Claude during DoW's contract negotiations with Anthropic. DoW ultimately determined that
21 Anthropic's conduct constituted a fully mature and significant supply chain risk—including increased
22 potential for AI model manipulation, insider threat risk, data exfiltration, and denial of service—that
23 posed a direct, unmitigable risk to DoW's warfighting capabilities and national security mission.

24 22. While Anthropic presents a supply chain risk, it is technically and operationally
25 infeasible to remove the technology from all DoW systems immediately, particularly in the midst of
26 active operations. Because of this reality, the designation allows a 180-day offramp to remove
27 Anthropic's Claude model from its systems and migrate to alternative LLM products without impacting
28 operational readiness. This is a significantly compressed timeline to ensure that this risk is removed

1 from DoW's systems, particularly because of the need to integrate another vendor's products and
2 services, including the associated requisite security clearance.

3 23. This reality is expressed in a March 5, 2026, memorandum issued by the DoW Chief
4 Information Officer. In this memorandum, the Chief Information Officer determined that "DoW
5 Components will discontinue all use of the Covered Company's products across all DoW systems
6 within 180 days." The memorandum adds that new procurements involving Anthropic's products are
7 disallowed, as these products are no longer authorized for installation in DoW covered systems.

8 24. As noted, Claude is used in a variety of functions throughout DoW. This is a result of
9 Claude being the first AI model that was available to function in DoW's classified networks and one of
10 the first AI models integrated through Amazon Web Services (AWS), which was awarded the first
11 contract in 2016. This placed Claude in the lead on multiple fronts. However, other companies have
12 been closing the gaps.

13 25. DoW expects that within 180-days, barring any significant change in necessity, it will be
14 able to create the digital space needed for another system and prepare for a seamless handoff from
15 Claude to ensure that the risk is efficiently removed from DoW networks.

16 26. This process has already been initiated. An injunction pausing this process would in and
17 of itself be a significant threat to the national security of the United States.

18 27. An injunction preventing the removal of Anthropic's technology from DoW systems as
19 soon as possible would result in an ongoing threat to national security remaining on DoW's systems,
20 and allowing contractors to continue to engage with Anthropic as a subcontractor to DoW would itself
21 create an additional intolerable risk. As a subcontractor, Anthropic poses the same threats as it would
22 as a prime contractor. The incorporation of Anthropic's systems into a product on DoW systems would
23 cause the same risks regardless of whether it flows directly to DoW systems or through a prime
24 contractor.

25 28. During this transition period, DoW is taking additional measures to mitigate the supply
26 chain risk and national security harms presented by Anthropic leadership's behavior with regard to
27 DoW systems. The Department is working with third-party cloud service providers to ensure Anthropic
28 leadership cannot make unilateral changes to the containerized version of its LLM product that DoW

1 currently uses. DoW is also working with its counterintelligence and law enforcement partners to
2 assess the potential risk that Anthropic's LLM products may contain technical exploits, including ones
3 that could have been embedded by foreign nationals, given the leadership's pattern of behavior.
4 Finally, DoW is communicating its risk saturation findings with the other U.S. government departments
5 and agencies to support their own risk mitigation efforts.

6 29. DoW has an obligation and a duty to ensure the integrity of its operations and the safety
7 and security of its personnel, including from any risks that may be presented through its supply chain to
8 its covered systems. Supply chain security is national security. Therefore, the Department took action
9 to ensure the integrity of its covered systems.

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

11 EXECUTED this 17th day of March, 2026, at Washington, DC.

12
13 

14 Emil Michael
15
16
17
18
19
20
21
22
23
24
25
26
27
28

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

ANTHROPIC PBC,

Plaintiff,

v.

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

Defendants.

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

SECOND DECLARATION OF EMIL MICHAEL

Pursuant to 28 U.S.C. § 1746, I, Emil Michael, declare as follows:

1. I am the Under Secretary of War for Research and Engineering (USW(R&E)) and Chief Technology Officer for the Department of War (DoW). I have held this position since May 20, 2025.

2. In my current position, I am responsible for spearheading the Department's efforts to ensure U.S. military technological superiority and keep DoW at the forefront of innovation. I provide strategic direction and oversight for DoW's entire research, development, and prototyping enterprise, which includes providing critical input on the acquisition, implementation, and use of cutting-edge technologies such as artificial intelligence (AI).

3. This declaration is based on my personal knowledge as well as information made available to me through reasonable diligence in the course of my official duties.

DoW's 2025-2026 Contract Negotiations with Anthropic

4. During the course of contract negotiations with DoW in late 2025 and early 2026, Anthropic became a supply chain risk following a progression of risk that reached a saturation point as a result of the behavior of its leadership, resulting in DoW's designation of Anthropic under 10 U.S.C. § 3252.

5. Contrary to Anthropic's assertions, DoW did raise its concerns during contract negotiations with Anthropic about the Department's need to use Anthropic's technology for all lawful

1 uses. Beyond that, the Department is not required to preemptively share information relating to national
2 security concerns with private companies. Moreover, the Department generally does not share national
3 security concerns with private actors, particularly when the Department has concerns about the private
4 actor itself.

5 6. Anthropic has attempted to characterize itself as a trusted partner to DoW. However,
6 Anthropic's continued attempts, including now under oath, to dictate to DoW what the nature of the
7 Department's relationship with Anthropic is or has been demonstrates Anthropic's desire to insert itself
8 into DoW decision-making.

9 7. While Anthropic has tried to disclaim a desire to have an approval role in DoW's
10 decision-making, Anthropic leadership repeatedly sought to insert themselves into such decision-
11 making. First, in a contract negotiation meeting on December 4, 2025, Anthropic leadership expressed
12 that DoW would have to call Anthropic in real time to seek authorization for a usage exception to one
13 of their redlines, which are not prohibited by law. Alarming, that statement demonstrated not only
14 that Anthropic demanded an operational veto of DoW's decision-making, but that Anthropic would
15 seek to exercise that veto, possibly in situations where any delay or disruption to U.S. military
16 operational decisions and execution could endanger American lives and national security. Second, in
17 both meetings and a publicly disclosed message to Anthropic staff, Anthropic leadership stated that it
18 has two redlines that it will not allow the Department to cross.¹ If accepted, this would fundamentally
19 insert Anthropic into decision-making related to those issues, and it engenders concern about other
20 redlines the company may have, now or in the future.

21 8. Indeed, despite Anthropic's claim that it never desired an approval role in military
22 operations, Anthropic repeatedly insisted in negotiations and in public statements that it would not
23 allow DoW to use its technology to cross its redlines. That is in fact an operational veto over DoW's
24 use of Anthropic technology for all lawful uses and confirms that Anthropic did want to intervene in
25 and assert control over DoW operations in at least those two areas.

26 9. Moreover, Anthropic's insistence during contract negotiations on its redlines inherently
27 injects the company into DoW's decision-making because those redlines could be embedded

28 ¹ <https://www.anthropic.com/news/statement-department-of-war>.

1 restrictions on its model's functionality. That is a per se operational veto and suggests, contrary to the
2 company's assertions, that Anthropic leadership believed it could enforce such a veto, including during
3 active military operations.

4 10. Despite Anthropic's claims that before contract negotiations broke down the parties were
5 near agreement on language that would address Anthropic's concerns about its technology being used
6 for lethal autonomous warfare, the fact remains that Anthropic would not agree to acquisition of its
7 LLM products with the usage terms and technical and service delivery specifications DoW requires.

8 11. Regarding the contractual language Anthropic proposed on March 4 ("For the avoidance
9 of doubt, [Anthropic] understands that this license does not grant or confer any right to control or veto
10 lawful Department of War operational decision-making"), this language remained insufficient to
11 mitigate DoW's concerns that Anthropic would exercise its own judgment—and embed that judgment
12 into model guardrails and weights—about what constitutes "lawful" operational decision-making.

13 12. Contrary to Anthropic's suggestion, any ongoing discussions do not undermine any of
14 the determinations made by DoW. Rather, DoW will consider any information provided by Anthropic
15 that may warrant altering its supply chain risk designation in whole or in part.

16 **Anthropic's Access to and Control of Its AI Model on DoW Systems**

17 13. As previously explained in my March 17 declaration, the baseline risk inherent in the
18 relatively opaque nature of large language model (LLM) technology that DoW procures from Anthropic
19 escalated due to the unusual degree of control that Anthropic retains over its model, as well as
20 Anthropic's adversarial posture towards DoW's statutory mission and the manner in which it is
21 conducted.

22 14. The Department was reliant on Anthropic to continuously provide updated versions of
23 its model to keep pace with the most advanced technology and emerging threats. Unlike traditional
24 software, which has reviewable code, language models are a black box. Anthropic has full control over
25 the weights in, and thus the resulting behavior and outputs of, the model it delivers to the Department.
26 This access and control could enable Anthropic to impact the model's availability or performance.

15. Again, these threats do not arise only from Anthropic's work as a prime contractor. The incorporation of Anthropic's systems into a product on DoW systems would cause the same risks regardless of whether it flows directly to DoW systems or through a prime contractor.

16. Because of these access and control risks, DoW took prompt action in conjunction with the supply chain risk designation to work with its prime contractors to remove Anthropic's access, via the prime contractors, to make updates or other changes to the model. DoW did this to ensure that Anthropic no longer has the ability to interfere with DoW systems. DoW is also conducting an audit for any malicious or unintended software intrusions to existing Anthropic technology on DoW systems that could interfere with ongoing or future operations.

17. Regarding, the U.S. Centers for Disease Control's (CDC) lawful use of Anthropic's LLM technology being limited by safety filters, Anthropic failed to inform the prime contractor and the agency upfront about updates it made to the filters or how the filters could limit the product's functionality in relation to the CDC's sensitive infectious disease research.

18. DoW learned about this incident in the context of its own negotiations with Anthropic and assessed that Anthropic's demonstrated lack of transparency about the limits it had embedded in its AI models, and its apparent unwillingness to work with the government's prime contractors to enable mission continuity, creates operational risk that Anthropic may make updates to its model that cause the model to no longer function as expected when used for sensitive, but lawful purposes such as the CDC's research or DoW's military operations.

Supply Chain Risk Designation

19. I understand that the Court has inquired about the date I completed and signed the memorandum bearing the Bates stamp DoW-PI-005. I completed and signed this memorandum on March 2, 2026.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct.

1 EXECUTED this 24th day of March, 2026.

2
3
4 
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
Emil Michael

TRUTH.**← Truth Details**

4903 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**13.9k** ReTruths **57.5k** Likes

Feb 27, 2026, 3:47 PM



← 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.

5:14 PM · Feb 27, 2026 · 13.2M Views

 10K

 14K

 70K

 9.2K



 Read 10.2K replies

New to X?

Sign up now to get you

 Sign u

 Sign i

Creat

By signing up, you agree
[Privacy Policy](#), includin

Relevant peo

Secretary o
@SecWar
United State

What's happ

Trending in United Sta
Fable 5
Trending with [Mythos](#)

Trending in United Sta
Moana

Sports · Trending
Serena

Trending in United Sta
David Grusch

[Show more](#)

[Terms of Service](#) | [Priva](#)
[Accessibility](#) | [Ads info](#)



Post

Treasury Secretary Scott Bessent

@SecScottBessent

At the direction of

[@POTUS](#)

, the

[@USTreasury](#)

is terminating all use of Anthropic products, including the use of its Claude platform, within our department.

The American people deserve confidence that every tool in government serves the public interest, and under President Trump no private company will ever dictate the terms of our national security.

Rapid Response 47

@RapidResponse47

.

Feb 27



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 **Sign up** -of-control, Radical Left AI company run by

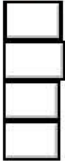
people who have no idea what the real world is all about. I thank you for your attention to this matter. MAKE AMERICA GREAT AGAIN!

ANT_AR-0257

10:57 AM · Mar 2, 2026 ·

1.6M

Views



New to X?

Sign up now to get your own personalized timeline!

Sign up with Google



Create account

By signing up, you agree to the [Terms of Service](#) and [Privacy Policy](#), including [Cookie Use](#).

Relevant people

Treasury Secretary Scott Bessent

@SecScottBessent



79th United States Secretary of the Treasury

Rapid Response 47

@RapidResponse47



Official White House Rapid Response account. Supporting

[@POTUS](#)

What’s happening

Trending in United States

Lincoln Logs

Trending in United States

Kate Martin

Music · Trending

Bob Fosse

Trending in United States

Kathy Hochul

[Show more](#)

[Terms of Service](#) |

[Privacy Policy](#) |

[Cookie Policy](#) |

[Accessibility](#) |

[Ads info](#) |

 © 2026 X Corp.



Post

Treasury Secretary Scott Bessent

@SecScottBessent

No private company will ever dictate the terms of our national security.

Anthropic's attempts to push use clauses into their contracts with the United States government are unacceptable, and their products will no longer be utilized by the @USTreasury or any other government agency.



1:45 PM · Mar 4, 2026 ·

507K

Views



Sign up

New to X?

Sign up now to get your own personalized timeline!

Sign up with Google

Create account

By signing up, you agree to the [Terms of Service](#) and [Privacy Policy](#), including [Cookie Use](#).

Relevant people

Treasury Secretary Scott Bessent

@SecScottBessent

79th United States Secretary of the Treasury

What's happening

Trending in United States

Lincoln Logs

Trending in United States

Kate Martin

Music · Trending

Bob Fosse

Trending in United States

Kathy Hochul

Show more



From: Moore, Michael <Michael.Moore@treasury.gov>
Date: Monday, March 9, 2026 at 4:50 PM
To: Jim Liew <jim@sokat.com>, Susan An <susan@sokat.com>
Cc: Fox, Sean <Sean.Fox@treasury.gov>, Burkeitt, William
<William.Burkeitt@treasury.gov>, Zhou, Annie <Annie.Zhou@treasury.gov>
Subject: RE: Windsurf Change Request

Thank you!

Respectfully,
Michael Moore

From: Jim Liew <jim@sokat.com>
Sent: Friday, March 6, 2026 2:32 PM
To: Susan An <susan@sokat.com>
Cc: Moore, Michael <Michael.Moore@treasury.gov>; Fox, Sean <Sean.Fox@treasury.gov>;
Burkeitt, William <William.Burkeitt@treasury.gov>; Zhou, Annie <Annie.Zhou@treasury.gov>
Subject: Re: Windsurf Change Request

**** Caution:** External email from: [jim@sokat.com] Pay attention to suspicious links and attachments. Send suspicious email to suspect@treasury.gov **

Please see from Cognition —

“Just heard back from our federal engineering team. The new release has moved into production so Anthropic models should no longer be available within Treasury's environment.”

Sent from my iPhone

On Mar 6, 2026, at 12:38 PM, Susan An <susan@sokat.com> wrote:

Hi Mike,

Please find below a brief messaging from Cognition regarding the ongoing issues between Anthropic and the Pentagon, with more to follow as needed. Most importantly, we're following the situation extremely closely, staying updated on how Cognition is working directly with all relevant ecosystem partners and their contingency plans to minimize potential disruption.

From Cognition:

"You may have seen reporting that the Department of War is designating Anthropic LLMs as a supply chain risk. We are monitoring this closely and taking steps to ensure Windsurf continues to deliver full agentic software development capability regardless of any changes to Anthropic model availability.

While Anthropic models are used in Windsurf's FedRAMP SKU, Windsurf FedRAMP is not limited to Anthropic. Today, the FedRAMP SKU also supports OpenAI GPT 4.1 via Azure Government, and the platform will remain functional if Anthropic models become unavailable or you choose not to use them. Model availability will not impact non-Cascade Windsurf features, including AI-powered Tab Autocomplete.

For additional non-Anthropic models in the FedRAMP High + ITAR posture, availability is gated by cloud provider authorization. Our standard approach is to enable a model within 24 hours after it is authorized for DISA IL4 in the relevant GovCloud environment, which we use as a proxy for ITAR since there is no formal ITAR accreditation. The main bottleneck is AWS, Azure, and GCP timelines, not our integration work.

If your users do not require ITAR/IL4 controls (meaning your use case only requires FedRAMP High controls) we can offer more model options supported by the Google Vertex platform at lower compliance levels. We can start this switch immediately and easily. We can also disable models for you at the enterprise level, and will soon allow enterprise admins to control this for their organizations. We can reenable access at any later time.

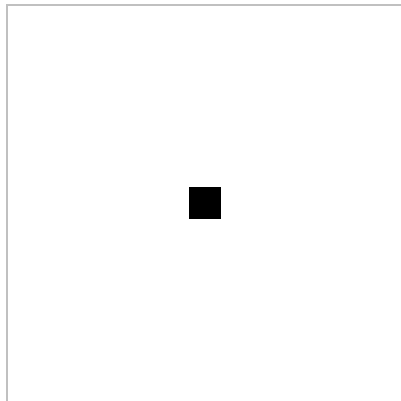
We are tracking the DoW-Anthropic situation closely and working directly with AWS, Anthropic, and the DoW to understand the evolving guidance. We will update you directly as we learn more."

Warmest regards,

Susan An, Esq. | CEO

M **Privacy** | E: susan@sokat.com

LinkedIn: <https://hyperlink.services.treasury.gov/agency.do?origin=www.linkedin.com/in/sokatceo/>



On Fri, Mar 6, 2026 at 11:12 AM Susan An <susan@sokat.com> wrote:

Hi Mike,

Thanks for your patience. Frank from Cognition provided the following

update regarding turning off Anthropic within Treasury's FedRAMP environment.

This update was posted internally at 12:16am yesterday:

- High-level update: "Anthropic models expected to be turned off for Treasury sometime tonight" - referring to the night of March 5th
 - Removing Claude required more engineering work than originally anticipated, involving a new release (1.58.104) and a sizeable change to FeStart's configuration file.
 - To maintain compliance and auditability, our federal engineering team opted for a deliberate fix over a hasty approach.
 - At the time of this update, the new release was undergoing a vulnerability scan and would be promoted to production once the scan finished.

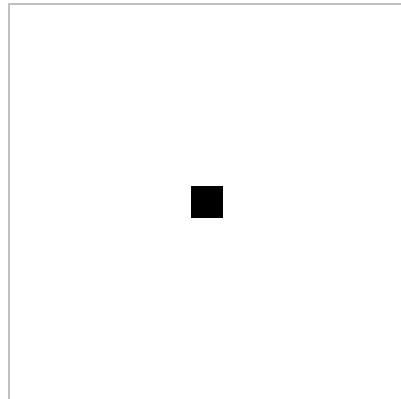
We will closely monitor the status of this requested change and will provide you with a resolution as soon as it is complete.

Warmest regards,

Susan An, Esq. | CEO

[Privacy](#) | E: susan@sokat.com

LinkedIn: <https://hyperlink.services.treasury.gov/agency.do?origin=www.linkedin.com/in/sokatceo/>



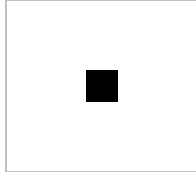
On Fri, Mar 6, 2026 at 10:03 AM Jim Liew <jim@sokat.com> wrote:

Mike,

Just called them, they are asking their engineering team now if it's

completed, and we will get you a response ASAP.

Jim Kyung-Soo Liew, Ph.D.
Email: Jim@SoKat.com
President and Founder of SoKat.com



On Fri, Mar 6, 2026 at 9:44 AM <Michael.Moore@treasury.gov> wrote:

Good morning Susan and Jim,

Can you confirm the requested change is complete?

Respectfully,
Michael Moore

From: Fox, Sean <Sean.Fox@treasury.gov>
Sent: Wednesday, March 4, 2026 10:27 AM
To: Susan An <susan@sokat.com>; Moore, Michael
<Michael.Moore@treasury.gov>
Cc: jim@sokat.com; Zhou, Annie <Annie.Zhou@treasury.gov>; Burkeitt,
William <William.Burkeitt@treasury.gov>
Subject: RE: Windsurf Change Request

+ Annie/Bill

From: Susan An <susan@sokat.com>
Sent: Wednesday, March 4, 2026 10:25 AM
To: Moore, Michael <Michael.Moore@treasury.gov>
Cc: jim@sokat.com; Fox, Sean <Sean.Fox@treasury.gov>
Subject: Re: Windsurf Change Request

**** Caution:** External email from: [susan@sokat.com] Pay
attention to suspicious links and attachments. Send suspicious email
to suspect@treasury.gov **

Hi Michael,

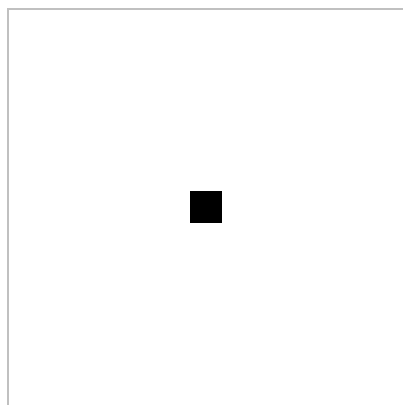
I am confirming receipt of this email and will promptly provide you with our plan to remove Claude Sonnet models from Treasury's Windsurf implementation as soon as possible.

Warmest regards,

Susan An, Esq. | CEO

Privacy [REDACTED] | E: susan@sokat.com

LinkedIn: <https://hyperlink.services.treasury.gov/agency.do?origin=www.linkedin.com/in/sokatceo/>



On Wed, Mar 4, 2026 at 10:13 AM <Michael.Moore@treasury.gov> wrote:

Good morning,

In accordance with direction from President Trump all federal agencies are to immediately cease all use of Anthropic's technology.

To comply with this direction, provide a plan to remove Claude Sonnet models from Treasury's Windsurf implementation as soon as possible.

Respectfully,

<image001.png>

Michael Moore

Security Architect | Digital Services
Treasury Common Services Center – Technology
Services

||| |||

| Michael.Moore@treasury.gov

SECTION I - OVERVIEW

1.1 – OVERVIEW

This contract is issued by the Internal Revenue Service (IRS), Office of the Chief Procurement Officer (OCPO), Treasury Operations Branch (TOB) on behalf of the Department of Treasury's Office of the Chief Information Officer (OCIO) for the provision of state-of-the-art Artificial Intelligence (AI)-powered coding assistant tools and AI-based chat tools that can be delivered as a shared service across the Department of Treasury (Treasury) and its bureaus. All tools shall perform in accordance with the Statement of Work (SOW) – Attachment 1- and shall be subject to the following terms and conditions.

1.2 – TYPE OF CONTRACT

This is an Indefinite Delivery / Indefinite Quantity (IDIQ) contract for commercial AI Coding Tools and AI Chat Tools. Performance of all work under this contract will be initiated by the issuance of delivery orders which will include purchase of services using all CLINs necessary.

1.3 – MINIMUM AND MAXIMUM ORDER AMOUNTS

The Department of Treasury (Treasury) will order a minimum value of [REDACTED] under the contract. Treasury will have the entire term of the contract (including all options that could be exercised) to fulfill the contract minimum. The specific CLINs and quantities will be identified in the task order(s) issued under the contract. The exercise of an option period does not re-establish the contract minimum. During the entire duration of this contract, Treasury may order items in any quantity up to the maximum specified below.

There are no maximum quantities or amounts for each individual CLIN, task order, or contract period.

The maximum aggregate amount of all task orders issued under the contract shall not exceed [REDACTED] for the entire term of the contract, including all options that could be exercised.

1.4 – PERIOD OF PERFORMANCE

The total period of performance for the resulting order will be a Base plus 4 option years.

The Period of Performance is as follows:

Base Period: 09/27/2025 – 09/26/2026
Option Period 1: 09/27/2026 – 09/26/2027
Option Period 2: 09/27/2027 – 09/26/2028
Option Period 3: 09/27/2028 – 09/26/2029
Option Period 4: 09/27/2029 – 09/26/2030

1.5 – PRICE SCHEDULE

Item Description	Price	Unit of Measure
AI Coding Assistant		Per License
Ai ChatBot		Per License

1.5.1 Unit pricing in the Base Year remains constant throughout the duration of the contract and does not escalate in outlying years.

1.6 – FAR 52.217-8 Pricing/Rates

This contract includes FAR Clause 52.217-8 Option to Extend Services, which may be exercised by the Government at any time during the life of this contract. Should the Government elect to exercise this option, the prices established in the period effective prior to the 52.217-8 period will be utilized during the 52.217-8 extension period(s).

1.7 – PLACE OF PERFORMANCE

The primary place of performance shall be the Offeror's workplace, as this Order primarily provides software licenses.

1.8 – TRAVEL

Contractor travel is not anticipated for this Order. Any travel must be pre-approved by the Contracting Officer Representative (COR). All travel shall be in accordance with the Federal Travel Regulations and FAR 31.205-46.

1.9 – FEDERAL RISK AND AUTHORIZATION MANAGEMENT PROGRAM (FedRAMP)

FedRAMP certification is required within twelve (12) months of contract award. The vendor shall bear the costs for completing any third-party assessments or other coordination required to obtain or maintain FedRAMP authorization. This includes all expenses related to assessments, documentation, testing, and communication with the FedRAMP Program Management Office (PMO) or Third Party Assessment Organizations (3PAOs). The Government will not reimburse any costs incurred in pursuit of FedRAMP authorization. While the Government will not reimburse the aforementioned costs, to the extent required and allowable, the Government will facilitate the vendor's efforts to prepare and submit a FedRAMP authorization package. Federal Risk and Authorization Management Program (FedRAMP) authorization process as detailed at <https://www.fedramp.gov/rev5/agency-authorization/> and <https://www.fedramp.gov/rev5/documents-templates/>

Contract No. 2032L225D00002
Contract Terms and Conditions
Page 4 of 27

SECTION II – STATEMENT OF WORK

See Attachment 1 – Statement of Work (SOW)

SECTION III – GENERAL AND ADMINISTRATIVE INFORMATION

3.1 – CONTRACTING OFFICER (CO)

- (a) The Awarding CO is **Kathleen Guyther**, Kathleen.A.Guyther@irs.gov
- (b) The Administrative CO is **Ricky L. Callahan Jr.**, Ricky.L.Callahanjr@irs.gov
- (c) In accordance with FAR 1.602, the CO has the authority to enter into, administer, or terminate contracts.
- (d) The CO is responsible for ensuring performance of all necessary actions for effective contracting, ensuring compliance with the terms of the contract, and safeguarding the interests of the United States in its contractual relationships.
- (e) Only a warranted Contracting Officer is authorized to change the specifications, price, terms, or conditions of this contract. No payments will be made for any unauthorized supplies or services or for any unauthorized changes to the work specified herein. This includes any services performed by the Contractor of his own volition or at the request of an individual other than a warranted Contracting Officer.
- (f) Requests for changes to the contract must be emailed to the CO.

3.2 – DTAR 1052.201-70 CONTRACTING OFFICER'S REPRESENTATIVE (COR) APPOINTMENT AND AUTHORITY (APR 2015)

- (a) The COR is **William Burkeitt**
 - (1) William.Burkeitt@treasury.gov
- (b) Performance of work under this contract is subject to the technical direction of the COR identified above, or a representative designated in writing. The term “technical direction” includes, without limitation, direction to the contractor that directs or redirects the labor effort, shifts the work between work areas or locations, and/or fills in details and otherwise serves to ensure that tasks outlined in the work statement are accomplished satisfactorily.
- (c) Technical direction must be within the scope of the contract specification(s)/work statement. The COR does not have authority to issue technical direction that:
 - (1) Constitutes a change of assignment or additional work outside the contract specification(s)/work statement;
 - (2) Constitutes a change as defined in the clause entitled “Changes”;
 - (3) In any manner causes an increase or decrease in the contract price, or the time required for contract performance;

- (4) Changes any of the terms, conditions, or specification(s)/work statement of the contract;
 - (5) Interferes with the contractor's right to perform under the terms and conditions of the contract; or
 - (6) Directs, supervises or otherwise controls the actions of the Contractor's employees.
- (d) Technical direction may be oral or in writing. The COR must confirm oral direction in writing within five workdays, with a copy to the Contracting Officer.
- (e) The Contractor shall proceed promptly with performance resulting from the technical direction issued by the COR. If, in the opinion of the Contractor, any direction of the COR or the designated representative falls within the limitations of (c) above, the Contractor shall immediately notify the Contracting Officer no later than the beginning of the next Government work day.

3.3 – ALTERNATE CONTRACTING OFFICER’S REPRESENTATIVE

3.3.1– The alternate COR is: **Annie Zhou**

3.3.2- Email: Annie.Zhou@treasury.gov

3.4 – DELIVERY ORDER PLACEMENT

3.4.1 – Delivery orders (DO) may be issued in accordance with the IDIQ SOW and shall be issued with approval of the CO.

3.4.2 – All DOs are subject to the terms and conditions of the base contract. In the event of a conflict between a TO and the base contract, the base contract will take precedence.

3.4.3 – All costs associated with preparation, presentation, and/or discussion of the Contractor’s DO proposal shall be at the Contractor’s expense. Post-award DO administration (including applicable personnel cost allocations by DO) shall also be at the Contractor’s expense. The Contractor is responsible for determining the most appropriate method for recovering such costs (e.g., direct or indirect charges to DO)

3.5 – INVOICING NOTIFICATION AND SUPPORT DOCUMENTATION

The Contractor shall invoice in accordance with the Pricing Schedule agreed upon in the respective awarded order. The Contractor shall submit payment requests and receiving reports using Invoice Processing Platform (IPP) Invoicing, Receipt, and Acceptance application which is a secure Government web-based system for electronic invoicing, receipt, and acceptance. The Contractor shall include supporting documentation (e.g., delivery receipts, time sheets, & material/travel costs, etc.) to the invoice in IPP. When requested by the COR, the Contractor shall directly provide a soft copy of the invoice and any supporting invoice documentation directly to the COR within 24 hours of request to assist in validating the invoiced amount against the products/services provided during the billing cycle. See section entitled “Electronic Invoicing and Payment Requirements For IPP” below.

3.6 – POST AWARD EVALUATION OF CONTRACTOR PERFORMANCE (JUNE 2020)

Interim and final evaluations of contractor performance will be prepared on this contract in accordance with FAR 42.15. The Assessing Official (e.g., Contracting Officer) will prepare a final performance evaluation at the time the work on the contract is completed. In addition to the final evaluation, interim evaluations will be prepared annually to coincide with the anniversary date of the contract.

The past performance evaluation process is a paperless process using the Contractor Performance Assessment Reporting System (CPARS). CPARS is a web-based system that allows for electronic processing of the performance evaluation report. The completed evaluation was previously available in the Past Performance Information Retrieval System (PPIRS), but since the General Services Administration officially retired PPIRS and merged it with CPARS, it created “a single system” that “provides one location and one account to perform functions such as creating and editing performance and integrity records, changes to administering users, running reports, generating performance records, and viewing/managing performance records.

Once the Contractor is registered in CPARS, they will receive an automatically-generated email with detailed login instructions. Further details, systems requirements, and training information for CPARS is available at <https://www.cpars.gov/>. The CPARS User Manual, registration for Online Training for Contractors, and a practice application may be found at this site as well.

Interim and final evaluations will be provided to the Contractor for their review and comment as soon as practicable after completion of the evaluation. Evaluations of contractor past performance will be posted to the relevant past performance database no more than 14 days after the information is provided to the contractor. On day 15, whether the contractor has responded or not, the evaluation automatically posts to PPIRS. If the Contractor elects not to provide comments, they should acknowledge receipt of the evaluation by indicating "No comment" and then sign and date the form. If the Contractor does not sign and submit the form within 14 days, it will automatically be returned to the Government.

Contractors who disagree with a government evaluation can request to meet with the Contracting Officer to discuss their scores and provide feedback or justification for their performance. No requirement exists for the government to meet with the contractor; however, if a contractor requests a meeting, the government may accept the request.

Any such meeting does not alter the requirement that an evaluation be posted to PPIRS within 14 days.

Several avenues still exist for the contractor to influence the review. First, the contractor may submit a comment after the 14-day period expires and the review has been posted to PPIRS. The contractor's late comments must be posted to PPIRS; however, the government's original report will still be available to all source selection officials.

Although authorized, an agency is not required to modify its evaluation based upon a contractor's comments. Second, the contractor may appeal its review one level above the Contracting Officer to the

Reviewing Official. Again, the appeal does not stop the 14- day reporting period and the original evaluation will be posted on PPIRS.

The following guidelines apply concerning the Contractor's use of the past performance evaluation:

1. Protect the evaluation as "source selection information." After review, transmit the evaluation by completing and submitting the form through CPARS. If for some reason the Contractor is unable to view and/or submit the form through CPARS, contact the Contracting Officer for further instructions.
2. Strictly control access to the evaluation within the Contractor's organization. Ensure the evaluation is never released to persons or entities outside of the Contractor's control.
3. Prohibit the use of or reference to evaluation data for advertising, promotional material, pre-award surveys, responsibility determinations, production readiness reviews, or other similar purposes.
4. A copy of the completed past performance evaluation will be available in CPARS for the Contractor's review and for Government use supporting source selection actions after it has been finalized.

3.7 – ELECTRONIC INVOICING AND PAYMENT REQUIREMENTS FOR THE INVOICE PROCESSING PLATFORM (IPP) (JUL 2019)

(a) Definitions:

"Short payment" as used in this clause means the partial payment of an invoice for goods/services actually rendered at the time of payment when the invoice includes additional goods/services that have not yet been provided/rendered.

"Short payment" example: The contract requires the delivery of a set number of items, with the price, delivery location, and delivery due date also specified. The vendor delivers 50% of the items as specified but invoices for 100% of the items. Before implementation of the IPP, the IRS would have paid the vendor for the items delivered and instructed the vendor to reinvoice the IRS when the balances of the items were delivered. In other words, the IRS would "short pay" the invoice since the IRS did not remit payment for the full invoice amount. With implementation of the IPP, the IRS can no longer do this because the IRS cannot accept an electronic invoice that includes items not yet received. The IRS will reject the invoice. The vendor needs to submit an invoice for only the items received by the IRS (in this case, 50%), and, if these items meet all other contract terms and conditions, the IRS will pay the invoiced amount. The vendor submits subsequent invoice(s) for items as they are delivered and accepted.

- (b) The Invoice Processing Platform (IPP) is a secure Web-based electronic invoicing and payment information service available to all Federal agencies and their suppliers. Effective October 1, 2012, invoicing for payment through the IPP will be mandatory for all new contract awards. Additional information regarding the IPP may be found at the IPP website address

<https://www.ipp.gov>. Contractors must complete the contractor point of contact information below and submit it with their proposal submissions. Contractors may contact the IPP Helpdesk for assistance via e-mail at ippgroup@stls.frb.org or via phone at (866) 973-3131. Once a contract award has been made, the contractor will be contacted by the IPP via e-mail to set-up an account. It will be necessary for contractors to login to their IPP accounts every 90 days to keep their IPP accounts active.

(c) Contractor Point of Contact Information

Contractor Name: _____

Contractor IPP Point of Contact Name: _____

Contractor Phone Number: _____

Contractor E-mail Address: _____

(d) Electronic Invoicing and Payment Requirements

Vendor invoices submitted electronically through the IPP should be in the proper format and contain the information required for payment processing. To be approved for payment, a "proper invoice" must list the items specified in FAR 52.232-25 (a)(3)(i) through (a)(3)(x), or in the case of a Commercial Item Contract, the items included in 52.212-4(g)(1)(i) through (g)(1)(x).

If the vendor is offering a discount via the IPP, the discount must be reflected on the invoice. The vendor will select 'Create Invoice'. The IPP system will default to 'Net 30 Prompt Pay' under the Payment Terms dropdown box. The vendor will select from 54 different discount options for the invoice that is being created. If the vendor chooses to offer a discount on the invoice screen, the information will interface to the payment system for processing. Discounts that are offered on attachments rather than the invoice itself cannot be accepted.

Under this contract, the following documents are required to be submitted as an attachment to the invoice. Please do not submit into IPP any documentation/attachments that conflict with what is stated on the invoice:

-Monthly Status Report

Payment and Invoice Questions

For payment and invoice questions, contact the Ancillary Systems at (304) 254-3372 or via e-mail at cfo.fm.ipp.customer.support@irs.gov.

(e) Waiver

If the Contractor is unable to use the IPP for submitting payment requests starting on October 1, 2012, then a waiver form must be completed and submitted with the contractor's proposal

submission for review and approval by the Contracting Officer based on one of the conditions listed in the waiver. The vendor will be notified prior to award as to whether their request for waiver has been approved or denied. If the waiver is granted, then a copy of the waiver must be submitted with each paper invoice that the vendor submits to the payment office or the invoice will be returned.

(f) Short Payment

Short payment on vendor submitted invoices will no longer be processed or paid. If any portion of the invoice does not meet the requirements for a proper invoice, the entire invoice shall be rejected and returned to the vendor unpaid.

IRS Invoice Processing Platform (IPP) Waiver Form

The IRS invoicing and payment requirements requires that all invoices under awards made (or effective) on or after October 1, 2012, be submitted electronically via the IPP unless a waiver is requested and granted. If the Contractor is unable to submit its invoice through the IPP, the Contractor shall complete this waiver form indicating the reason for the waiver request by selecting the appropriate box below and providing a narrative summarizing in detail the circumstances requiring a waiver. For a solicitation, submit the waiver form with the proposal submission. For a modification that incorporates the IPP clause into an existing contract, submit the waiver form with the modification. The CO will notify the vendor via email or another appropriate means of communication prior to award as to whether their waiver has been approved or denied. If the waiver is granted, then a copy of the approved waiver must be submitted with each invoice that the vendor submits to the payment office or the invoice will be returned.

Reason for requesting a waiver of the requirement to submit an electronic invoice via the IPP:

☐ 1. Submission of invoices through IPP would impose a hardship on an individual (includes employees and sole proprietors) due to: either a physical or mental disability; a geographic, language, or literacy barrier; or an undue financial burden. The requirement to submit invoices through the IPP is automatically waived for all individuals who do not have payment capability using ACH with a U.S. financial institution.

☐ 2. The political, financial or communications infrastructure where the place of business is located does not support access to the IPP for submitting invoices electronically.

☐ 3. The contractor is located within an area designated by the President of the United States or an authorized agency administration as a disaster area. (Please identify area/location.)

☐ 4. The submission of invoices electronically may pose a threat to national security, the life or physical safety of an individual may be endangered, or a law enforcement action may be compromised.

[] 5. The agency does not expect to receive more than one invoice from the same contractor within a one-year period. i.e., the invoice submission is non-recurring.

[] 6. The contractor customarily submits a high volume of invoices on a regular basis via file format, not currently supported by the IPP (i.e., uses a file format other than XML or CSV) and the high volume of invoices would cause a significant burden to the contractor if submitted through the IPP individually. If utilizing this exception, please identify the file formats supported by your invoicing system so that the IPP may consider implementing the requested file format at a later date. File format(s) used: _____

[] 7. Other - Please explain:

Attach a separate sheet of paper with a summary narrative substantiating the circumstances for the waiver exception selected from above (1 through 7).

Waiver Submitted By:

Contractor Name

Name of Person Submitting Request for Waiver

Title

Signature of Person Submitting Request for Waiver

E-mail Address

Phone No.

Contract/Order No.

Date Submitted

Waiver Approved By:

Contracting Officer's Name Printed

Contracting Officer's Signature

Date

3.8 – INTERNET PROTOCOL VERSION 6 (IPV6) REQUIREMENTS.

This contract involves the acquisition of IT that uses Internet Protocol (IP) technology (see NIST USGv6 Program and NIST 500 281 for additional guidance on IPv6 requirements). The Contractor agrees that: (1) any system hardware, software, firmware, networked component (voice, video or data) developed, procured, or acquired in support and/or performance of this contract shall comply with IPv6 standard and interoperate with both IPv6 and IPv4 systems and products; and (2) it has IPv6 technical support for development and implementation and fielded product management available. If the Contractor plans to offer a deliverable that involves IT that is not initially compliant, the Contractor

shall: (1) obtain the Contracting Officer's approval before starting work on the deliverable; (2) provide a migration path and firm commitment to upgrade to IPv6 for all application and product features; and (3) have IPv6 technical support for development and implementation and fielded product management available.

Should the Contractor discover or is made aware of during the performance of this contract that a product or service developed, procured, or acquired in support and/or performance of this contract does not conform to the IPv6 standard, it must immediately notify the Contracting Officer's Representative and Contracting Officer of such nonconformance and act in accordance with instructions of the Contracting Officer. The Contractor agrees to bring into compliance (e.g., upgrade, modification, replacement) the nonconforming product at no cost to the Government.

SECTION IV – CONTRACT CLAUSES**4.1 – FAR 52.252-2 CLAUSES INCORPORATED BY REFERENCE (FEB 1998)**

This contract incorporates one or more clauses by reference, with the same force and effect as if they were given in full text. Upon request, the Contracting Officer will make their full text available. Also, the full text of a clause may be accessed electronically at this/these address(es):

FAR: www.acquisition.gov

DTAR: www.acquisition.gov/dtar

The following FAR and DTAR clauses are incorporated by reference:

NUMBER	TITLE	DATE
52.202-1	Definitions	Jun 2020
52.203-3	Gratuities	Apr 1984
52.203-11	Certification and Disclosure Regarding Payments to Influence Certain Federal Transactions	Sep 2024
52.203-12	Limitation on Payments to Influence Certain Federal Transactions	Jun 2020
52.204-19	Incorporation by Reference of Representations and Certifications	Dec 2014
52.224-1	Privacy Act Notification	Apr 1984
52.224-2	Privacy Act	Apr 1984
52.227-1	Authorization and Consent	Jun 2020
52.227-2	Notice and Assistance Regarding Patent and Copyright Infringement	Jun 2020
52.227-19	Commercial Computer Software License	Dec 2007
52.229-3	Federal, State, and Local Taxes	Feb 2013
52.239-1	Privacy or Security Safeguards	Aug 1996
52.243-1	Changes—Fixed-Price, Alternate II (Apr 1984)	Aug 1987
1052.210-70	Contractor Publicity	Apr 2015
1052.222-70	Minority and Women Inclusion	Jan 2016
1052.232-7003	Electronic Submission of Payment Requests	Apr 2015

4.2 – FAR 52.217-8 OPTION TO EXTEND SERVICES

The Government may require continued performance of any services within the limits and at the rates specified in the contract. These rates may be adjusted only as a result of revisions to prevailing labor rates provided by the Secretary of Labor. The option provision may be exercised more than once, but the

total extension of performance hereunder shall not exceed 6 months. The Contracting Officer may exercise the option by written notice to the Contractor within **TWO (2) DAYS** of contract expiration.

(End of clause)

4.3 – FAR 52.217-9 – OPTION TO EXTEND THE TERM OF THE CONTRACT.

(a) The Government may extend the term of this contract by written notice to the Contractor within 2 days provided that the Government gives the Contractor a preliminary written notice of its intent to extend at least two (2) days before the contract expires. The preliminary notice does not commit the Government to an extension.

(b) If the Government exercises this option, the extended contract shall be considered to include this option clause.

(c) The total duration of this contract, including the exercise of any options under this clause, shall not exceed sixty (60) months.

(End of clause)

4.4 – DTAP 1052.224-70 CONTRACT PUBLICATION (OCT 2018)

(a) The Department of the Treasury (Treasury) may, at its sole discretion, publish this contract or portions thereof, including orders issued under the contract when deemed in the best interest of the Government.

(b) To afford the Contractor an opportunity to review and propose redactions for any information contained in the Treasury contract that may be subject to a FOIA exemption, the Contractor may submit, within ten business (10) days from the date of award of this contract or any order issued under the contract—

(1) A pdf file of the fully executed contract or order that is suitable for publication and which includes all Contractor proposed redactions (e.g, trade secrets or any commercial or financial information that the Contractor believes to be privileged or confidential business information) and.

(2) A written statement identifying the portions of each proposed redactions, including the applicable exemption under the Freedom of Information Act (FOIA), 5 U.S.C. 552, and, in the case of FOIA Exemption 4, 5 U.S.C. 552(b)(4), shall demonstrate why the information is considered to be a trade secret or commercial or financial information that is privileged or confidential.

(c) Treasury will consider the Contractor's proposed redactions and associated grounds for nondisclosure prior to making a determination as to what information may be properly withheld for purposes of publishing this contract or portions thereof.

(d) The Contractor may submit a request to the CO for additional time to complete the action prescribed by paragraph (b) of this clause. The lack of action by the Contractor will be deemed by the Government as there being no information in the Treasury contract subject to a FOIA exemption.

(e) Information provided by the Contractor in response to this clause may itself be subject to disclosure under the FOIA.

**4.5 – DTAP 1052.232-39 UNENFORCEABILITY OF UNAUTHORIZED OBLIGATIONS
(DEVIATION 00002) (APR 2018)**

(a) Definition. As used in this clause-

"Commercial supplier agreements" means terms and conditions customarily offered to the public by vendors of supplies or services that meet the definition of commercial item set forth in FAR 2.101 and intended to create a binding legal obligation on the end user. Commercial supplier agreements (CSA) are particularly common in information technology acquisitions, including acquisitions of commercial computer software and commercial technical data, but they may apply to any supply or service. The term applies-

(1) Regardless of the format or style of the document. For example, a CSA may be styled as standard terms of sale or lease, Terms of Service (TOS), End User License Agreement (EULA), or another similar legal instrument or agreement, and may be presented as part of an offer or quotation responding to a solicitation;

(2) Regardless of the media or delivery mechanism used. For example, a CSA may be presented as one or more paper documents or may appear on a computer or other electronic device screen during a purchase, software installation, other product delivery, registration for a service, or another transaction.

(b) Except as stated in paragraph (c) of this clause, when any supply or service acquired under this contract is subject to any CSA, that includes any language, provision, or clause requiring the Government to pay any future fees, penalties, interest, legal costs or to indemnify the Contractor or any person or entity for damages, costs, fees, or any other loss or liability that would create an Anti-Deficiency Act violation (31 U.S.C. 1341), the following shall govern:

(1) Any such language, provision, or clause is unenforceable against the Government.

(2) Neither the Government nor any Government authorized end user shall be deemed to have agreed to such clause by virtue of it appearing in the CSA. If the CSA is invoked through an "I agree" click box or other comparable mechanism (e.g., "click-wrap" or "browse-wrap" agreements), execution does not bind the Government or any Government authorized end user to such clause.

(3) Any such language, provision, or clause is deemed to be stricken from the CSA.

(c) Paragraph (b) of this clause does not apply to indemnification or any other payment by the Government that is expressly authorized by statute and specifically authorized under applicable agency regulations and procedures.

(End of Clause)

4.6 – SECTION 508 INFORMATION, DOCUMENTATION AND SUPPORT (DEC 2019)

In accordance with 36 CFR, Appendix C to Part 1194, the [information and communication technology \(ICT\)](#) products and product support services documentation furnished in performance of this contract shall be provided at no additional cost. The contractor shall provide information, documentation, and support relative to the supplies and services as described in the statement of work, performance work statement or statement of objectives (select one). The following technical standards and provisions have been determined to be applicable to this contract:

☒ Chapter 6: Support Documentation and Services

☒ 601 General

☒ 601.1

☒ 602 Support Documentation

☒ 602.1 ☒ 602.2 ☒ 602.3 ☒ 602.4

☒ 603 Support Services

☒ 603.1 ☒ 603.2 ☒ 603.3

(End of clause)

4.7 – SECTION 508 CONFORMANCE (APR 2024)

Each [information and communication technology \(ICT\)](#) product and/or product related service delivered under the terms of this contract, at a minimum, shall conform to the applicable accessibility standards at 36 CFR, Appendix C .

The following technical standards have been determined to be applicable to this contract:

☐ Chapter 4: Hardware

☐ 401 General

☐ 401.1

☐ 402 Closed Functionality

☐ 402.1 ☐ 402.2(1-6) ☐ 402.3 ☐ 402.4 ☐ 402.5

☐ 403 Biometrics

___ 403.1

___ 404 Preservation of Information Provided for Accessibility

___ 404.1

___ 405 Privacy

___ 405.1

___ 406 Standard Connections

___ 406.1

___ 407 Operable Parts

___ 407.1 ___ 407.2 ___ 407.3 ___ 407.4 ___ 407.5 ___ 407.6 ___ 407.7 ___ 407.8

___ 408 Display Screens

___ 408.1 ___ 408.2 ___ 408.3

___ 409 Status Indicators

___ 409.1

___ 410 Color Coding

___ 410.1

___ 411 Audible Signals

___ 411.1

___ 412 ICT with Two-Way Communication

___ 412.1 ___ 412.2 ___ 412.3 ___ 412.4 ___ 412.5 ___ 412.6 ___ 412.7 ___ 412.8

___ 413 Closed Caption Processing Technologies

___ 413.1

___ 414 Audio Description Processing Technologies

___ 414.1

___ 415 User Controls for Captions and Audio Descriptions

___ 415.1

X Chapter 5: Software

X 501 General

X 501.1

X 502 Interoperability with Assistive Technology

☐ 502.1 ☐ 502.2 ☒ 502.3 ☐ 502.4(A-G)

☐ 503 Applications

☐ 503.1 ☐ 503.2 ☒ 503.3 ☐ 503.4

☐ 504 Authoring Tools

☐ 504.1 ☐ 504.2 ☒ 504.3 ☐ 504.4

☐ Chapter 7: Referenced Standards

☐ 701 General

☐ 701.1

☐ 702 Incorporation by Reference

☒ 702.1 ☐ 702.2 ☐ 702.3 ☒ 702.4 ☐ 702.5 ☐ 702.6 ☐ 702.7 ☐ 702.8 ☐ 702.9
☒ 702.10

The standards do not require the installation of specific accessibility-related software or the attachment of an assistive technology device, but merely require that the ICT be compatible with such software and devices so that it can be made accessible if so required by the agency in the future.

The following functional performance criteria (36 CFR Chapter 3) apply to this contract.

☐ **Chapter 3: Functional Performance Criteria**

☐ 301 General

☐ 301.1

☐ 302 Functional Performance Criteria

☒ 302.1 ☐ 302.2 ☐ 302.3 ☐ 302.4 ☐ 302.5 ☐ 302.6 ☐ 302.7 ☐ 302.8 ☐ 302.9

(End of clause)

4.8 – SECTION 508 SERVICES (APR 2024)

All contracts, solicitations, purchase orders, delivery orders and interagency agreements that contain a requirement of services which will result in the delivery of a new or updated information and communication technology (ICT) item/product must conform to the applicable provisions of the

appropriate technical standards in 36 CFR, Appendix C to Part 1194, and functional performance criteria in 36 CFR Chapter 3, unless an agency exception to this requirement exists at E202 General Exceptions.

The following technical standards and provisions have been determined to be applicable to this contract:

___ **Chapter 4: Hardware**

___ 401 General

___ 401.1

___ 402 Closed Functionality

___ 402.1 ___ 402.2(1-6) ___ 402.3 ___ 402.4 ___ 402.5

___ 403 Biometrics

___ 403.1

___ 404 Preservation of Information Provided for Accessibility

___ 404.1

___ 405 Privacy

___ 405.1

___ 406 Standard Connections

___ 406.1

___ 407 Operable Parts

___ 407.1 ___ 407.2 ___ 407.3 ___ 407.4 ___ 407.5 ___ 407.6 ___ 407.7 ___ 407.8

___ 408 Display Screens

___ 408.1 ___ 408.2 ___ 408.3

___ 409 Status Indictors

___ 409.1

___ 410 Color Coding

___ 410.1

☐ 411 Audible Signals

☐ 411.1

☐ 412 ICT with Two-Way Communication

☐ 412.1 ☐ 412.2 ☐ 412.3 ☐ 412.4 ☐ 412.5 ☐ 412.6 ☐ 412.7 ☐ 412.8

☐ 413 Closed Caption Processing Technologies

☐ 413.1

☐ 414 Audio Description Processing Technologies

☐ 414.1

☐ 415 User Controls for Captions and Audio Descriptions

☐ 415.1

☒ Chapter 5: Software

☒ 501 General

☒ 501.1

☒ 502 Interoperability with Assistive Technology

☒ 502.1 ☒ 502.2 ☒ 502.3 ☒ 502.4(A-G)

☒ 503 Applications

☒ 503.1 ☒ 503.2 ☒ 503.3 ☒ 503.4

☒ 504 Authoring Tools

☒ 504.1 ☒ 504.2 ☒ 504.3 ☒ 504.4

☒ Chapter 7: Referenced Standards

☒ 701 General

☒ 701.1

☒ 702 Incorporation by Reference

☒ 702.1 ☐ 702.2 ☒ 702.3 ☒ 702.4 ☐ 702.5 ☐ 702.6 ☐ 702.7 ☐ 702.8 ☐ 702.9
☒ 702.10

The standards do not require the installation of specific accessibility-related software or the attachment of an assistive technology device, but merely require that the ICT be compatible with such software and devices so that it can be made accessible if so required by the agency in the future.

The following functional performance criteria (36 CFR Chapter 3) apply to this contract.

☒ **Chapter 3: Functional Performance Criteria**

☒ 301 General

☒ 301.1

☒ 302 Functional Performance Criteria

☒ 302.1 ☒ 302.2 ☒ 302.3 ☒ 302.4 ☒ 302.5 ☒ 302.6 ☒ 302.7 ☒ 302.8 ☒ 302.9

(End of clause))

4.9 – FAR 52.212-4 CONTRACT TERMS AND CONDITIONS—COMMERCIAL PRODUCTS AND COMMERCIAL SERVICES (NOV 2023)

(a) *Inspection/Acceptance.* The Contractor shall only tender for acceptance those items that conform to the requirements of this contract. The Government reserves the right to inspect or test any supplies or services that have been tendered for acceptance. The Government may require repair or replacement of nonconforming supplies or reperformance of nonconforming services at no increase in contract price. If repair/replacement or reperformance will not correct the defects or is not possible, the Government may seek an equitable price reduction or adequate consideration for acceptance of nonconforming supplies or services. The Government must exercise its post-acceptance rights-

(1) Within a reasonable time after the defect was discovered or should have been discovered; and

(2) Before any substantial change occurs in the condition of the item, unless the change is due to the defect in the item.

(b) *Assignment.* The Contractor or its assignee may assign its rights to receive payment due as a result of performance of this contract to a bank, trust company, or other financing institution, including any Federal lending agency in accordance with the Assignment of Claims Act (31 U.S.C. 3727). However, when a third party makes payment (e.g., use of the Governmentwide commercial purchase card), the Contractor may not assign its rights to receive payment under this contract.

(c)*Changes*. Changes in the terms and conditions of this contract may be made only by written agreement of the parties.

(d)*Disputes*. This contract is subject to 41 U.S.C. chapter 71, Contract Disputes. Failure of the parties to this contract to reach agreement on any request for equitable adjustment, claim, appeal or action arising under or relating to this contract shall be a dispute to be resolved in accordance with the clause at Federal Acquisition Regulation (FAR) [52.233-1](#), Disputes, which is incorporated herein by reference. The Contractor shall proceed diligently with performance of this contract, pending final resolution of any dispute arising under the contract.

(e)*Definitions*. The clause at FAR [52.202-1](#), Definitions, is incorporated herein by reference.

(f)*Excusable delays*. The Contractor shall be liable for default unless nonperformance is caused by an occurrence beyond the reasonable control of the Contractor and without its fault or negligence such as, acts of God or the public enemy, acts of the Government in either its sovereign or contractual capacity, fires, floods, epidemics, quarantine restrictions, strikes, unusually severe weather, and delays of common carriers. The Contractor shall notify the Contracting Officer in writing as soon as it is reasonably possible after the commencement of any excusable delay, setting forth the full particulars in connection therewith, shall remedy such occurrence with all reasonable dispatch, and shall promptly give written notice to the Contracting Officer of the cessation of such occurrence.

(g)*Invoice*.

(1)The Contractor shall submit an original invoice and three copies (or electronic invoice, if authorized) to the address designated in the contract to receive invoices. An invoice must include-

(i)Name and address of the Contractor;

(ii)Invoice date and number;

(iii)Contract number, line item number and, if applicable, the order number;

(iv)Description, quantity, unit of measure, unit price and extended price of the items delivered;

(v)Shipping number and date of shipment, including the bill of lading number and weight of shipment if shipped on Government bill of lading;

(vi)Terms of any discount for prompt payment offered;

(vii)Name and address of official to whom payment is to be sent;

(viii)Name, title, and phone number of person to notify in event of defective invoice; and

(ix) Taxpayer Identification Number (TIN). The Contractor shall include its TIN on the invoice only if required elsewhere in this contract.

(x) Electronic funds transfer (EFT) banking information.

(A) The Contractor shall include EFT banking information on the invoice only if required elsewhere in this contract.

(B) If EFT banking information is not required to be on the invoice, in order for the invoice to be a proper invoice, the Contractor shall have submitted correct EFT banking information in accordance with the applicable solicitation provision, contract clause (e.g., [52.232-33](#), Payment by Electronic Funds Transfer-System for Award Management, or [52.232-34](#), Payment by Electronic Funds Transfer-Other Than System for Award Management), or applicable agency procedures.

(C) EFT banking information is not required if the Government waived the requirement to pay by EFT.

(2) Invoices will be handled in accordance with the Prompt Payment Act ([31 U.S.C. 3903](#)) and Office of Management and Budget (OMB) prompt payment regulations at 5 CFR Part 1315.

(h) *Patent indemnity*. The Contractor shall indemnify the Government and its officers, employees and agents against liability, including costs, for actual or alleged direct or contributory infringement of, or inducement to infringe, any United States or foreign patent, trademark or copyright, arising out of the performance of this contract, provided the Contractor is reasonably notified of such claims and proceedings.

(i) *Payment*.-

(1) *Items accepted*. Payment shall be made for items accepted by the Government that have been delivered to the delivery destinations set forth in this contract.

(2) *Prompt payment*. The Government will make payment in accordance with the Prompt Payment Act ([31 U.S.C. 3903](#)) and prompt payment regulations at 5 CFR Part 1315.

(3) *Electronic Funds Transfer (EFT)*. If the Government makes payment by EFT, see [52.212-5\(b\)](#) for the appropriate EFT clause.

(4) *Discount*. In connection with any discount offered for early payment, time shall be computed from the date of the invoice. For the purpose of computing the discount earned, payment shall be considered to have been made on the date which appears on the payment check or the specified payment date if an electronic funds transfer payment is made.

(5)*Overpayments.* If the Contractor becomes aware of a duplicate contract financing or invoice payment or that the Government has otherwise overpaid on a contract financing or invoice payment, the Contractor shall-

(i)Remit the overpayment amount to the payment office cited in the contract along with a description of the overpayment including the-

(A)Circumstances of the overpayment (*e.g.*, duplicate payment, erroneous payment, liquidation errors, date(s) of overpayment);

(B)Affected contract number and delivery order number, if applicable;

(C)Affected line item or subline item, if applicable; and

(D)Contractor point of contact.

(ii)Provide a copy of the remittance and supporting documentation to the Contracting Officer.

(6)*Interest.*

(i)All amounts that become payable by the Contractor to the Government under this contract shall bear simple interest from the date due until paid unless paid within 30 days of becoming due. The interest rate shall be the interest rate established by the Secretary of the Treasury as provided in [41 U.S.C. 7109](#), which is applicable to the period in which the amount becomes due, as provided in (i)(6)(v) of this clause, and then at the rate applicable for each six-month period as fixed by the Secretary until the amount is paid.

(ii)The Government may issue a demand for payment to the Contractor upon finding a debt is due under the contract.

(iii)*Final decisions.* The Contracting Officer will issue a final decision as required by [33.211](#) if-

(A)The Contracting Officer and the Contractor are unable to reach agreement on the existence or amount of a debt within 30 days;

(B)The Contractor fails to liquidate a debt previously demanded by the Contracting Officer within the timeline specified in the demand for payment unless the amounts were not repaid because the Contractor has requested an installment payment agreement; or

(C)The Contractor requests a deferment of collection on a debt previously demanded by the Contracting Officer (see [32.607-2](#)).

(iv) If a demand for payment was previously issued for the debt, the demand for payment included in the final decision shall identify the same due date as the original demand for payment.

(v) Amounts shall be due at the earliest of the following dates:

(A) The date fixed under this contract.

(B) The date of the first written demand for payment, including any demand for payment resulting from a default termination.

(vi) The interest charge shall be computed for the actual number of calendar days involved beginning on the due date and ending on-

(A) The date on which the designated office receives payment from the Contractor;

(B) The date of issuance of a Government check to the Contractor from which an amount otherwise payable has been withheld as a credit against the contract debt; or

(C) The date on which an amount withheld and applied to the contract debt would otherwise have become payable to the Contractor.

(vii) The interest charge made under this clause may be reduced under the procedures prescribed in FAR [32.608-2](#) in effect on the date of this contract.

(j) *Risk of loss.* Unless the contract specifically provides otherwise, risk of loss or damage to the supplies provided under this contract shall remain with the Contractor until, and shall pass to the Government upon:

(1) Delivery of the supplies to a carrier, if transportation is f.o.b. origin; or

(2) Delivery of the supplies to the Government at the destination specified in the contract, if transportation is f.o.b. destination.

(k) *Taxes.* The contract price includes all applicable Federal, State, and local taxes and duties.

(l) *Termination for the Government's convenience.* The Government reserves the right to terminate this contract, or any part hereof, for its sole convenience. In the event of such termination, the Contractor shall immediately stop all work hereunder and shall immediately cause any and all of its suppliers and subcontractors to cease work. Subject to the terms of this contract, the Contractor shall be paid a percentage of the contract price reflecting the percentage of the work performed prior to the notice of termination, plus reasonable charges the Contractor can demonstrate to the satisfaction of the Government using its standard record keeping system, have resulted from the termination. The Contractor shall not be required to comply with the cost accounting standards or contract cost principles for this purpose. This paragraph does not give the Government any right to audit the Contractor's

records. The Contractor shall not be paid for any work performed or costs incurred which reasonably could have been avoided.

(m)*Termination for cause.* The Government may terminate this contract, or any part hereof, for cause in the event of any default by the Contractor, or if the Contractor fails to comply with any contract terms and conditions, or fails to provide the Government, upon request, with adequate assurances of future performance. In the event of termination for cause, the Government shall not be liable to the Contractor for any amount for supplies or services not accepted, and the Contractor shall be liable to the Government for any and all rights and remedies provided by law. If it is determined that the Government improperly terminated this contract for default, such termination shall be deemed a termination for convenience.

(n)*Title.* Unless specified elsewhere in this contract, title to items furnished under this contract shall pass to the Government upon acceptance, regardless of when or where the Government takes physical possession.

(o)*Warranty.* The Contractor warrants and implies that the items delivered hereunder are merchantable and fit for use for the particular purpose described in this contract.

(p)*Limitation of liability.* Except as otherwise provided by an express warranty, the Contractor will not be liable to the Government for consequential damages resulting from any defect or deficiencies in accepted items.

(q)*Other compliances.* The Contractor shall comply with all applicable Federal, State and local laws, executive orders, rules and regulations applicable to its performance under this contract.

(r)*Compliance with laws unique to Government contracts.* The Contractor agrees to comply with [31 U.S.C. 1352](#) relating to limitations on the use of appropriated funds to influence certain Federal contracts; [18 U.S.C. 431](#) relating to officials not to benefit; [40 U.S.C. chapter 37](#), Contract Work Hours and Safety Standards; [41 U.S.C. chapter 87](#), Kickbacks; [49 U.S.C. 40118](#), Fly American; and [41 U.S.C. chapter 21](#) relating to procurement integrity.

(s)*Order of precedence.* Any inconsistencies in this solicitation or contract shall be resolved by giving precedence in the following order:

(1)The schedule of supplies/services.

(2)The Assignments, Disputes, Payments, Invoice, Other Compliances, Compliance with Laws Unique to Government Contracts, and Unauthorized Obligations paragraphs of this clause;

(3)The clause at [52.212-5](#).

(4)Addenda to this solicitation or contract, including any license agreements for computer software.

(5) Solicitation provisions if this is a solicitation.

(6) Other paragraphs of this clause.

(7) The [Standard Form 1449](#).

(8) Other documents, exhibits, and attachments.

(9) The specification.

(t) [Reserved]

(u) Unauthorized Obligations.

(1) Except as stated in paragraph (u)(2) of this clause, when any supply or service acquired under this contract is subject to any End User License Agreement (EULA), Terms of Service (TOS), or similar legal instrument or agreement, that includes any clause requiring the Government to indemnify the Contractor or any person or entity for damages, costs, fees, or any other loss or liability that would create an Anti-Deficiency Act violation ([31 U.S.C. 1341](#)), the following shall govern:

(i) Any such clause is unenforceable against the Government.

(ii) Neither the Government nor any Government authorized end user shall be deemed to have agreed to such clause by virtue of it appearing in the EULA, TOS, or similar legal instrument or agreement. If the EULA, TOS, or similar legal instrument or agreement is invoked through an "I agree" click box or other comparable mechanism (e.g., "click-wrap" or "browse-wrap" agreements), execution does not bind the Government or any Government authorized end user to such clause.

(iii) Any such clause is deemed to be stricken from the EULA, TOS, or similar legal instrument or agreement.

(2) Paragraph (u)(1) of this clause does not apply to indemnification by the Government that is expressly authorized by statute and specifically authorized under applicable agency regulations and procedures.

(v) Incorporation by reference. The Contractor's representations and certifications, including those completed electronically via the System for Award Management (SAM), are incorporated by reference into the contract.

(End of clause)

4.10 – FAR 52.216-22 INDEFINITE QUANTITY (OCT 1995)

(a) This is an indefinite-quantity contract for the supplies or services specified, and effective for the period stated, in the Schedule. The quantities of supplies and services specified in the Schedule are estimates only and are not purchased by this contract.

(b) Delivery or performance shall be made only as authorized by orders issued in accordance with the Ordering clause. The Contractor shall furnish to the Government, when and if ordered, the supplies or services specified in the Schedule up to and including the quantity designated in the Schedule as the "maximum." The Government shall order at least the quantity of supplies or services designated in the Schedule as the "minimum."

(c) Except for any limitations on quantities in the Order Limitations clause or in the Schedule, there is no limit on the number of orders that may be issued. The Government may issue orders requiring delivery to multiple destinations or performance at multiple locations.

(d) Any order issued during the effective period of this contract and not completed within that period shall be completed by the Contractor within the time specified in the order. The contract shall govern the Contractor's and Government's rights and obligations with respect to that order to the same extent as if the order were completed during the contract's effective period; provided, that the Contractor shall not be required to make any deliveries under this contract after **September 28, 2030**.

(End of clause)

STATEMENT OF WORK (SOW)

AI Tools for Treasury

1 OVERVIEW

The U.S. Department of the Treasury (Treasury), Office of the Chief Information Officer (OCIO), is seeking qualified providers of state-of-the-art Artificial Intelligence (AI)-powered coding assistant tools (similar to GitHub Copilot, AWS CodeWhisperer, Cursor, etc.) and AI-based chat tools (similar to ChatGPT) that can be delivered as a shared service across the Department and its bureaus.

The intent is to provide developers and IT staff within Treasury with flexible, secure and modern AI-driven coding solutions tailored to their specific development workflows. Selected solutions must operate exclusively on a usage-based pricing model (e.g., per-token, per-completion, or per-query usage). Per-seat licensing or fixed monthly user fees will not be accepted.

2 DEFINITIONS

AI-based coding assistant tools refer to tools that support software development through features such as code generation, completion, debugging, and refactoring.

AI-based chat tools refer to interactive AI-driven platforms designed to assist users via natural language dialogue, including tasks such as research, drafting, summarization, or technical troubleshooting.

3 SCOPE

The requirement covers the procurement of secure, reliable, FedRAMP-authorized AI-based coding assistant tools and AI-based chat tools. These tools are intended to support Treasury's operations across a range of environments and use cases, while meeting stringent federal security and compliance mandates.

AI-based coding assistant tools are expected to support the software development lifecycle by enhancing developer productivity, improving code quality, and accelerating delivery through features such as code generation, completion, debugging, and refactoring.

AI-based chat tools are intended to support Treasury's general business functions by enabling secure, natural language interaction to assist with tasks such as information retrieval, content

drafting, summarization, and technical troubleshooting. These tools should be suitable for use by both technical and non-technical personnel in daily mission and administrative activities.

4 OBJECTIVES

A. General Requirements

The vendor shall provide tools that can meet the following requirements:

1. AI Powered AI Code Assistant Tool

1.1 AI Coding Assistant Capabilities

The vendor shall provide an AI coding assistant capable of delivering real-time, context-aware code suggestions, completions, and generation. The solution must support natural language prompts for code creation, editing, and explanation, allowing developers to engage with the tool in a conversational manner. It shall assist with common development tasks, including code review, documentation, refactoring, optimization, and automated test case generation, and must support a wide range of programming languages such as TypeScript, JavaScript, Python, Java, C#, C++, Go, SQL, Bash, PowerShell, COBOL, and PeopleCode.

At a minimum, the tool shall provide the following capabilities:

- Multi-language code generation
- Automated code reviews
- Code documentation generation
- Security issue detection (e.g., similar to CAST for PeopleSoft)
- Unit test and code coverage recommendations
- Performance tuning suggestions
- Troubleshooting assistance
- General technical guidance
- Contextual search and summarization of defects, requirements, and service tickets

1.2 Advanced Model Capabilities

The vendor shall provide access to advanced AI models with capabilities comparable to leading large language models such as GPT-4, AWS CodeWhisperer, Claude, or similar state-of-the-art technologies. The solution must offer a conversational generative AI service powered by large language models, enabling natural and effective interactions. It

should demonstrate proven accuracy and measurable productivity gains in both public and private sector development environments.

1.3 Model Bias and Explainability

Vendors shall provide models that include built-in mechanisms to mitigate bias and support explainability of outputs in accordance with applicable AI governance and ethical use principles. The solution shall provide transparency into how recommendations are generated, including visibility into model reasoning, contributing inputs, or decision logic, where available.

The system shall also offer user-level controls that allow individuals to accept, reject, or customize AI-generated outputs, enabling human-in-the-loop oversight and appropriate application of results.

The system shall offer interactive explanations, allowing users to explore the reasoning behind AI outputs. Users must be able to flag incorrect or problematic outputs for review and provide feedback to improve system performance. The system should track these corrections to prevent recurring issues.

1.4 FedRAMP Compliance

Vendors shall ensure their software is authorized at a FedRAMP Moderate or High impact level, as required by Treasury's security and data classification needs. The required FedRAMP level will be determined by Treasury based on risk assessments and the intended use of the software.

The software must be fully compatible with deployment in FedRAMP-authorized environments and support secure configuration, deployment, and operation in a manner that does not compromise the compliance posture of Treasury's cloud infrastructure.

Vendors shall provide documentation as needed to support Treasury's security assessment and Authorization to Operate (ATO) processes, and shall notify the Government of any known risks, limitations, or dependencies that could impact FedRAMP compliance.

Vendors shall maintain the applicable FedRAMP authorization throughout the period of performance and notify the Government of any change in status, delay in authorization efforts, or revocation of authorization.

1.5 Deployment and Integration

Vendors shall provide software that is deployable within a FedRAMP-authorized Government cloud environment, such as Azure GovCloud, AWS GovCloud, Google Cloud Assured Workloads, or IBM Cloud for Government, in accordance with Treasury's security and hosting requirements.

The software shall support seamless integration with commonly used Integrated Development Environments (IDEs), including but not limited to Visual Studio Code, JetBrains IDEs, Visual Studio, and Eclipse.

Additionally, the solution shall be compatible with major source control and DevOps platforms, including GitHub, Azure DevOps, GitLab, and Bitbucket, to enable integration with Treasury's existing development workflows.

1.6 Security, Privacy, and Compliance

The vendor shall ensure full compliance with all applicable federal security standards, including FISMA Moderate or High, NIST 800-53, and any Treasury-specific information assurance requirements. The solution must be capable of operating within a secure environment that aligns with the cybersecurity and risk management expectations of the federal government.

The solution must also meet U.S. data sovereignty requirements. All data processing and storage must take place within data centers physically located in the United States and managed exclusively by U.S. persons to ensure full jurisdictional and operational control.

Additionally, the vendor must guarantee that no Treasury data—including source code, prompts, metadata, or other user-generated content—will be used to train, fine-tune, or otherwise enhance any AI models without the explicit, written authorization of a designated Treasury Executive.

Token revalidation within the system shall occur at minimum every 30 minutes of active use. Idle timeout must be configurable, with a default maximum of 15 minutes of inactivity before requiring reauthentication. These parameters should be adjustable through administrative controls to accommodate different sensitivity levels.

Session tokens within the system must have a maximum lifetime of 12 hours, with sliding expiration supported for active sessions. Refresh tokens should be limited to 24 hours

maximum. All tokens must be securely stored, transmitted, and revocable through administrative action. Token validation should occur with each significant action, not just at session initiation.

The system must integrate seamlessly with one of the following: Visual Studio Code, JetBrains IDEs (particularly IntelliJ IDEA and PyCharm), Visual Studio, and Eclipse. These represent Treasury's primary development environments. Support for web-based IDEs is also desirable as the agency continues its modernization efforts.

The system shall allow for data to be logically segregated by bureau and configurable down to the user group level. Both encryption-at-rest and in-transit are mandatory, using FIPS 140-2 (or newer) validated cryptographic modules. Key management processes must support Treasury's security requirements and allow for integration with existing key management systems.

The solution must support IP-based restrictions to limit access to Treasury network ranges, device fingerprinting to prevent unauthorized device usage, and basic behavioral anomaly detection to identify suspicious patterns.

The solution must provide administrative override capabilities. Including forced logout and session termination for individual users or groups. Real-time alerts must be generated for suspicious activities such as multiple failed authentication attempts, access from unusual locations, or attempted session hijacking.

1.7 Reliability, Support, and Availability

The vendor shall provide a reliable and continuously available service in accordance with the service levels defined in the attached SLA table. These commitments include system uptime, support availability, response times, incident management, and data handling practices.

Vendors must ensure all system changes that may impact availability are communicated in advance and support a well-documented incident escalation path. Treasury may request regular reporting on system performance, service disruptions, or response metrics.

The system shall be optimized for operation across network connections with bandwidth as low as 10 Mbps and latency up to 100ms. Degraded but functional operation should be maintained even in suboptimal network conditions, with appropriate client-side caching to enhance user experience.

See Service Level Agreement (SLA) Table for detailed service expectations and remedies for non-compliance.

1.8 Training, Onboarding, and Change Management

To support Treasury's ability to independently adopt and operationalize the tools, vendors shall provide standard, commercially available training and onboarding resources that accompany the licensed product. No customization or tailored content is required or expected.

Vendors shall provide access to publicly available training materials such as user guides, tutorials, FAQs, and technical documentation. This should also include standard video tutorials, webinars, or onboarding modules that are typically included with the license at no additional cost.

Onboarding support must include general resources for new users, such as platform navigation instructions, setup guidance, and publicly available content for account provisioning, role management, and feature overviews. These materials should be designed to help users become self-sufficient with minimal external assistance.

For change management, vendors shall offer any enablement strategies or adoption guides that are typically part of their off-the-shelf solution. Treasury should also be granted access to any public-facing online communities, forums, or self-service support portals the vendor makes available to all customers. These resources must be sufficient to support a Treasury-led rollout without the need for customized vendor services.

1.9 Vendor Viability and Roadmap Alignment

Vendors shall maintain organizational capacity and product continuity necessary to support Treasury's use of the licensed AI tools throughout the contract period of performance. This includes continued availability of the offered solution, ongoing maintenance, and access to standard product enhancements and updates.

Vendors shall provide Treasury with access to product roadmap information to inform internal planning and ensure awareness of upcoming features, deprecations, or changes that may affect integration, security, or usability. Roadmap visibility is intended to support proactive Treasury adoption planning and does not require development of custom roadmaps or deliverables.

2. AI Powered Chat Assistant Tool

2.1 AI Chat Assistant Capabilities

The vendor shall provide an AI chat assistant capable of natural language understanding and generation to support a wide range of knowledge work and user interactions. The solution must enable users to engage conversationally using plain language prompts to receive contextually relevant responses. It should support tasks such as summarization, content drafting, question answering, policy interpretation, process explanation, and knowledge retrieval. The assistant must be capable of operating across a variety of subject domains, including IT, finance, human resources, cybersecurity, and procurement.

2.2 Advanced Model Capabilities

Same as paragraph 1.2.

2.3 Model Behavior and Explainability

Vendors shall provide models that include built-in mechanisms to mitigate bias and support explainability of outputs in accordance with applicable AI governance and ethical use principles. The solution shall offer transparency into how responses are generated, including citations, source tracing, or explanation of reasoning where available.

The system shall also offer user-level controls that allow individuals to accept, discard, or refine AI-generated responses, ensuring human-in-the-loop oversight for responsible use of the technology.

2.4 FedRAMP Compliance

Same as paragraph 1.4.

2.5 Deployment and Integration

The vendor shall provide software deployable within a FedRAMP-authorized Government cloud environment, such as Azure GovCloud, AWS GovCloud, Google Cloud Assured Workloads, or IBM Cloud for Government, in accordance with Treasury's security and hosting requirements.

The solution shall be accessible through a web-based interface and optionally provide integrations with collaboration platforms such as Microsoft Teams, Slack, or

ServiceNow. If available, integration with enterprise search systems or content repositories is preferred to enhance retrieval-augmented generation (RAG) capabilities.

2.6 Security, Privacy, and Compliance

The solution must comply with applicable federal security standards, including FISMA Moderate or High, NIST 800-53, and Treasury-specific requirements. All processing of Treasury prompts, responses, and metadata must occur within U.S.-based, FedRAMP-authorized environments.

Vendors shall guarantee that no user-generated content—including questions, prompts, summaries, or any conversational data—will be used to train or fine-tune models without explicit, written authorization from Treasury. The system must support administrative controls for prompt retention, logging, auditability, and user role management to ensure secure and compliant use.

2.7 Reliability, Support, and Availability

The vendor shall provide a reliable and continuously available AI chat service in accordance with the service levels defined in the attached SLA table. Treasury must be notified of planned outages, updates, or known issues that may impact user access or performance.

2.8 Training, Onboarding, and Change Management

The vendor shall provide standard training and onboarding materials to help users effectively adopt the AI chat assistant. Materials should include publicly available user guides, usage examples, FAQs, and best practices for effective prompt writing and use of the assistant. These resources should require no customization and be sufficient for Treasury-led rollout.

2.9 Vendor Viability and Roadmap Alignment

Same as paragraph 1.9.

3. Miscellaneous

3.1 This requirement is for code accelerators and chat generative response capabilities. It does NOT include any service requirements outside of standard labor provided as a part of the price of

licensing, e.g., basic provisioning and activation, access to standard support, routine software updates, self-service resource, license management assistance. In addition, this requirement does not include use cases for model access beyond developed integrated development environments.

3.2 Vendors shall provide a sandbox or test environment upon contract award to facilitate information sharing, configuration, and setup activities necessary to support smooth integration and implementation.

3.3 Treasury has a robust cyber security infrastructure, to include Security Technical Implementation Guides (STIGs).

3.4 Rollout of licensing may not follow a standard adoption model, and Vendors must be prepared to support the Treasury in a phased, non-linear, or decentralized deployment approach. This includes the ability to provision licenses incrementally, accommodate staggered onboarding across bureaus, and provide responsive support during varying levels of user adoption and system integration.

3.5 Treasury standards, frameworks and protocols will be provided post-award.

3.6 Vendors must configure, or assist Treasury in configuring, their tools to prevent users from replicating authenticated sessions for access outside the Treasury network.

Items to be included in the solicitation and resulting award:

Condition of Award and Continued Performance – Software Updates and Upgrade Path -

As a condition of award and continued contract performance, the Contractor shall provide all updates, upgrades, and new versions of the proposed software that are made generally available during the period of performance. This includes security patches, performance improvements, and feature enhancements. The Contractor must also ensure a clear and supported glide path to newer versions, with no degradation in functionality or disruption to government operations.

License and Usage Reporting – The vendor shall provide quarterly reports detailing active licenses, user activity – to include a highlight of licenses that have been inactive for 30 days or more, and feature utilization for all Treasury-authorized users. Reports shall include metrics on adoption, usage volume, and any inactive licenses. These reports are intended to support program oversight, license optimization, and ensure contract compliance. Treasury reserves the right to request additional data as needed to verify billing accuracy and support planning efforts.

The solution must support for SCIM or equivalent provisioning standards is required. Onboarding of new users should be completed within 24 hours of approval, while offboarding/deprovisioning must occur within 1 hour of request to maintain security. The system should support automated workflows that integrate with Treasury's approval processes.

The solution must support containerized deployments (Docker) and virtual machine installations. Bare-metal support is not required but would be considered advantageous. The primary deployment model will be within Treasury's FedRAMP-authorized cloud environments.

Pricing – Prices are established on a usage-based pricing model, such as per-token, per-query, or per-completion charges. No other pricing models are allowable under this contract.

Ordering – It is permissible for orders placed under this IDIQ to be issued with Not-To-Exceed (NTE) values. Although orders will be structured as Firm-Fixed-Price (FFP), the Government may elect to incrementally fund up to the NTE value. Additionally, vendors shall return any unused licenses to the open license pool if they have not been activated or in use for 90 consecutive days and shall adjust billing accordingly to reflect actual usage only.

IPv6 Compliance – The Contractor shall:

- Ensure all proposed tools—whether hosted on-premises or delivered as cloud services (SaaS)—support operation in IPv6-only environments.
- Ensure that all networking interfaces used for communication with end users, APIs, or integration points support IPv6 natively, without reliance on IPv4.
- Ensure the tools interoperate with Treasury's IPv6-enabled enterprise infrastructure, including but not limited to identity providers, endpoint protection platforms, and supporting systems.
- Ensure tools are capable of sending, receiving, and processing data over IPv6 networks independently of IPv4.
- Maintain compliance with OMB Memorandum M-21-07, "Completing the Transition to Internet Protocol Version 6 (IPv6)," and other applicable federal IPv6 requirements, including relevant NIST guidance, for the duration of the period of performance.
- If any tool, feature, or service is determined to be non-compliant, provide an update or alternative solution at no additional cost to the Government.
 - If full IPv6 compliance is not immediately achievable, provide a roadmap with milestones and expected timelines for achieving full support.
- Support periodic Government reviews to verify IPv6 compatibility and cooperate fully with any related assessments or testing activities.

On Ramp Procedures for IDIQ Holders - To maintain a competitive and capable contractor pool that meets evolving mission requirements, the Government reserves the right to conduct on-

ramp procedures to add new contractors to this IDIQ contract throughout the period of performance. On-ramping may occur at the Government's discretion under the following circumstances:

1. **Mission Need:** A determination that additional capacity, capabilities, or competition is required to meet evolving program demands or address contractor performance issues.
2. **Technology Evolution:** Emergence of new technologies or capabilities not represented in the current awardee pool.
3. **Market Conditions:** Identification of qualified vendors outside the original awardee pool through market research or industry outreach.
4. **Programmatic Changes:** Expansion of scope, funding availability, or the addition of new task areas not previously anticipated.

The Government may issue a Request for Proposal (RFP) or limited on-ramp solicitation, which will mirror the original evaluation criteria or include updated criteria as needed to reflect current mission priorities. Interested offerors will be evaluated competitively and fairly in accordance with the procedures outlined in the solicitation.

Contractors selected through an on-ramp process will be awarded an IDIQ contract under the same terms and conditions (unless otherwise stated) as the original pool. The on-ramp process may occur periodically or as-needed, at the Government's sole discretion.

This clause enables the Government to maintain flexibility, promote innovation, and ensure that the contractor base continues to meet programmatic and performance requirements over the life of the contract.

Off-Ramp Procedures for IDIQ Holders – The Government reserves the right to off-ramp IDIQ contract holders based on performance, responsiveness, or other objective criteria to maintain a pool of responsible, responsive contractors. Off-ramping may be initiated at the Government's discretion under any of the following conditions:

1. **Proposal Responsiveness:** Failure to submit proposals in response to at least 75 percent of order requests issued during any rolling 12-month period.
2. **Award Activity:** Failure to receive at least one task order award within any defined ordering period (e.g., base period or option year).
3. **Performance:** Receipt of two or more CPARS ratings of less than "Satisfactory" for orders performed under this IDIQ contract.
4. **Technical Noncompliance:** Repeated failure to meet Statement of Work (SOW) requirements or to deliver order outcomes on time, within budget, or at the required quality.

5. **Administrative Deficiencies:** Ongoing issues related to invoicing, reporting, or failure to meet contract deliverable schedules or documentation requirements.

Prior to removal, the contractor will receive written notice outlining the Government's intent to off-ramp, along with an opportunity to respond. Off-ramping may be executed through one or more of the following actions:

- Non-exercise of future ordering period options;
- Administrative modification to remove the contractor from the IDIQ award pool;
- Contract termination in accordance with FAR Part 49.

These procedures are intended to ensure high-performing contractors remain available to meet mission needs, while enabling the Government to manage contract risk effectively.

Initiation of off-ramping actions is at the Government's discretion, based on the criteria outlined above. Prior to final disposition, the contractor will be provided written notice and an opportunity to respond. Final off-ramping actions will be executed using one or more of the methods identified above and documented in the contract file in accordance with applicable regulations.

Service Level Agreements (SLAs):

Service Component	SLA Commitment	Measurement Method	Remedy for Non-Compliance
System Uptime	Minimum 99.9% monthly uptime	Monitored via vendor-provided availability logs	Service credit of 5% per 0.1% below target
Prompt Response Time	$\leq 2s$ (Simple), $\leq 4s$ (Moderate), $\leq 7s$ (Complex)	Measured from prompt receipt to first response byte, averaged over a rolling 7-day period, per prompt category	Service credit of 2% per 1s above target, per category
Prompt Response Time Reporting Requirement	Vendor-provided log and weekly statistical summary	Must include timestamp, category, and response time for each request	Right to audit or independently test performance

Support Availability	24/7 support via ticketing/email; business hours phone support	Support logs and availability reports	Service credit for missed coverage
Support Framework	Vendor must provide documented support hours, escalation procedures, and named contacts	Submission of support documentation and account contacts	Withholding of approval until criteria are met
Incident Acknowledgment (Severity 1)	Within 1 hour of report	Incident tracking system	Escalation to executive contact after 2 hours
Incident Resolution (Severity 1)	Within 4 hours	Incident ticket closure time	Service credit of 10% if unresolved in 8 hours
Model/Service Updates	Minimum 14 days' advance notice for major changes or updates	Written/email notice from vendor	Delay of update until agency review is complete
Data Residency	100% U.S.-based data storage and processing	Compliance audit and documentation	Contract breach and possible termination
Data Usage Restrictions	No use of Treasury data for model training without written Treasury Executive consent	Policy review and system controls audit	Right to audit or terminate
Model Downgrade/Failure Recovery	Failover to backup model within 15 minutes of outage	System recovery logs and uptime reporting	Service credit of 5% per 15-minute delay
SSO Availability (if applicable)	99.9% uptime for authentication services	Authentication service logs	Service credit of 2% per 0.1% below threshold
Self-Service Resources	Vendor must provide access to documentation, knowledge bases, FAQs, and developer guides	Review of vendor resource portal	Required prior to deployment acceptance

Special Requirements for AI-Based Tools – To ensure responsible, secure, and legally compliant use of AI-based tools, the Contractor shall adhere to the following requirements in addition to the general and functional requirements described elsewhere in this SOW:

1. Data Ownership and Rights

- The U.S. Government shall retain unlimited rights to all data entered into, processed by, or generated through the use of any AI tools provided under this contract. This includes, but is not limited to:
 - Input data (e.g., prompts, source code, queries)
 - Output data (e.g., generated content, responses, code suggestions)
 - Metadata, logs, interaction histories, and audit trails
- No data collected, processed, or generated under this contract may be used by the Contractor for any purpose other than performance of the contract, unless explicitly authorized in writing by the Contracting Officer.

2. Restrictions on Model Training or Enhancement

- The Contractor shall not use any Government-provided or Government-generated data—including prompts, outputs, source code, metadata, or usage logs—for the purpose of training, fine-tuning, or otherwise enhancing any AI models, algorithms, or commercial services, unless explicitly authorized in writing by the Government.

3. Audit and Inspection Rights

- The Government reserves the right to audit the AI system's behavior, performance, and data handling practices. This includes access to:
 - Usage logs
 - Prompt and completion histories
 - System-level metadata relevant to security, performance, or compliance
- The Contractor shall cooperate fully with any Government-requested audit or review.
- SAML 2.0 with ADFS compatibility is the minimum requirement, additional support for OAuth2 and OpenID Connect is highly desirable. Integration with Treasury's identity providers is not required at award but must be implemented within the first 30 days of the base period. Vendors should describe their experience with similar federal identity management integrations.

4. Model Updates and Notification

- The Contractor shall notify the Government in advance of any substantial changes to the AI models, including:
 - Model upgrades
 - Retraining or tuning events
 - Functional or performance-impacting updates
- The Government reserves the right to re-evaluate the solution following such updates to ensure continued compliance and suitability.

5. Transparency and Disclosure

- The Contractor shall provide documentation describing:
 - The general architecture and behavior of the AI model
 - How outputs are generated (e.g., general algorithmic approach or logic, where feasible)
 - Any known limitations, risks, or failure modes
- The AI tool must disclose to end users when they are interacting with an AI system and allow for human-in-the-loop oversight.

6. Responsible and Ethical Use

- The Contractor shall ensure the AI tools are aligned with applicable federal guidance on responsible and ethical AI use, including but not limited to:
 - OMB M-21-07 (IPv6 and modernization)
 - NIST AI Risk Management Framework
 - Treasury-specific AI policies, if provided
- The solution must include safeguards to detect and mitigate biased, harmful, or misleading outputs and support human review and correction where applicable.

From: [Burkeitt, William](#)
To: [Gavin Kerr](#)
Cc: [Zhou, Annie](#); [Fox, Sean](#); [Durose, Christine](#); [Showman, Tracey](#); [Marcinko, William](#)
Subject: RE: Treasury/Figma
Date: Thursday, March 5, 2026 9:26:21 AM
Attachments: [image001.png](#)

Hey Gavin,

Thanks for the response. Please keep us updated on the progress regarding the alternative models for the AI features. For now, we have turned off the AI features.

Thanks,
Bill



William Burkeitt

Supervisory Project Manager| Business Operations - Budget and Acquisition
Treasury Common Services Center – Technology Services

Phone: [Privacy](#)
william.burkeitt@treasury.gov

From: Gavin Kerr <gkerr@figma.com>
Sent: Wednesday, March 4, 2026 3:27 PM
To: Burkeitt, William <William.Burkeitt@treasury.gov>
Cc: Zhou, Annie <Annie.Zhou@treasury.gov>; Fox, Sean <Sean.Fox@treasury.gov>
Subject: Re: Treasury/Figma

**** Caution:** External email from: [gkerr@figma.com] Pay attention to suspicious links and attachments. Send suspicious email to suspect@treasury.gov **

Hey Bill, thanks for reaching out. Response below:

Figma's AI features within our Figma for Government offerings are currently powered by Anthropic. Our team is actively exploring alternative models to power these AI features, but in the interim your account's administrative users can disable them immediately by toggling them to "off" within the admin dashboard. This will block any and all usage of Anthropic without affecting the functionality of non-AI features on the Figma for Government platform.

To directly answer your question, turning AI "off" in Figma will result in 0 connection or usage to Anthropic.

Let me know if you have any questions.

Thanks!

On Wed, Mar 4, 2026 at 3:13 PM <William.Burkeitt@treasury.gov> wrote:

Hey Gavin,

Please provide what level of partnership/integration Figma has with Anthropic. Please provide this detail asap.

Can the Treasury Figma environment be isolated from any Anthropic association?

Thanks,
Bill



William Burkeitt

Supervisory Project Manager| Business Operations - Budget and Acquisition
Treasury Common Services Center – Technology Services

Phone: [Privacy](#)
william.burkeitt@treasury.gov

--



Gavin Kerr

[Privacy](#)

SOLICITATION/CONTRACT/ORDER FOR COMMERCIAL ITEMS <i>OFFEROR TO COMPLETE BLOCKS 12, 17, 23, 24, & 30</i>				1. REQUISITION NUMBER 25PR-SSPDIR-0363		ANPR-0313 PAGE 1			
2. CONTRACT NO. NNG15SD42B		3. AWARD/ EFFECTIVE DATE	4. ORDER NUMBER 2032H525F00055		5. SOLICITATION NUMBER		6. SOLICITATION ISSUE DATE 06/16/2025		
7. FOR SOLICITATION INFORMATION CALL:		a. NAME JONATHAN CARNEY			b. TELEPHONE NUMBER <i>(No collect calls)</i>		8. OFFER DUE DATE/LOCAL TIME		
9. ISSUED BY CODE 2-IRS IT (OITA) IRS IT (OITA) Internal Revenue Service AWSS/Procurement OS:A:P:T, Stop C7-430 5000 Ellin Road Lanham MD 20706				10. THIS ACQUISITION IS ☐ UNRESTRICTED OR ☒ SET ASIDE: 100.00 % FOR: ☒ SMALL BUSINESS ☐ HUBZONE SMALL BUSINESS ☐ SERVICE-DISABLED VETERAN-OWNED SMALL BUSINESS WOMEN-OWNED SMALL BUSINESS ☐ (WOSB) ELIGIBLE UNDER THE WOMEN-OWNED SMALL BUSINESS PROGRAM ☐ EDWOSB NAICS: 541519 SIZE STANDARD: 150					
11. DELIVERY FOR FOB DESTINATION UNLESS BLOCK IS MARKED ☐ SEE SCHEDULE		12. DISCOUNT TERMS		13a. THIS CONTRACT IS A RATED ORDER UNDER DPAS (15 CFR 700) ☐		13b. RATING		14. METHOD OF SOLICITATION ☐ RFQ ☐ IFB ☐ RFP	
15. DELIVER TO CODE SSPDIR SSPDIR 1500 PENNSYLVANIA AVENUE, NW OFFICE OF THE CIO ATTN: 1750 PENNSYLVANIA AVE, NW., WASHINGTON DC 20220				16. ADMINISTERED BY CODE 2- IRS IT (OITA) IRS IT (OITA) Internal Revenue Service AWSS/Procurement OS:A:P:T, Stop C7-430 5000 Ellin Road Oxon Hill MD					
17a. CONTRACTOR/ OFFEROR CODE HQAJMSZDK666		FACILITY CODE		18a. PAYMENT WILL BE MADE BY CODE ARC/ASD/IPP					
ARCHITECHTURE SOLUTIONS LLC 11325 RANDOM HILLS RD STE 360 FAIRFAX VA 22030-0972				ARC/ASD/IPP Submit invoices via the Invoice Processing Platform at www.ipp.gov Inquiries call 304-480-8000 #7					
TELEPHONE NO.				17b. CHECK IF REMITTANCE IS DIFFERENT AND PUT SUCH ADDRESS IN OFFER ☐				18b. SUBMIT INVOICES TO ADDRESS SHOWN IN BLOCK 18a UNLESS BLOCK BELOW IS CHECKED ☐ SEE ADDENDUM	
19. ITEM NO.	20. SCHEDULE OF SUPPLIES/SERVICES			21. QUANTITY	22. UNIT	23. UNIT PRICE	24. AMOUNT		
	This award is made off NASA SEWP for FIGMA Licenses. List of attachments: 1. Statement of Work 2. Contractors Quote 3. VPAT Base period of performance is 27 June 2025 to 26 June 2026. Delivery: 1 Days After Award <i>(Use Reverse and/or Attach Additional Sheets as Necessary)</i>								
25. ACCOUNTING AND APPROPRIATION DATA See schedule						26. TOTAL AWARD AMOUNT <i>(For Govt. Use Only)</i> 			
27a. SOLICITATION INCORPORATES BY REFERENCE FAR 52.212-1, 52.212-4. FAR 52.212-3 AND 52.212-5 ARE ATTACHED. ADDENDA ☐ ARE ☐ ARE NOT ATTACHED.									
27b. CONTRACT/PURCHASE ORDER INCORPORATES BY REFERENCE FAR 52.212-4. FAR 52.212-5 IS ATTACHED. ADDENDA ☐ ARE ☐ ARE NOT ATTACHED.									
28. CONTRACTOR IS REQUIRED TO SIGN THIS DOCUMENT AND RETURN COPIES TO ISSUING OFFICE. CONTRACTOR AGREES TO FURNISH AND DELIVER ALL ITEMS SET FORTH OR OTHERWISE IDENTIFIED ABOVE AND ON ANY ADDITIONAL SHEETS SUBJECT TO THE TERMS AND CONDITIONS SPECIFIED.					29. AWARD OF CONTRACT: _____ OFFER DATED _____. YOUR OFFER ON SOLICITATION (BLOCK 5), INCLUDING ANY ADDITIONS OR CHANGES WHICH ARE SET FORTH HEREIN, IS ACCEPTED AS TO ITEMS:				
30a. SIGNATURE OF OFFEROR/CONTRACTOR 				31a. UNITED STATES OF AMERICA (SIGNATURE OF CONTRACTING OFFICER)					
30b. NAME AND TITLE OF SIGNER <i>(Type or print)</i> Armando Ygbuhay, Managing Partner			30c. DATE SIGNED 06/26/2025	31b. NAME OF CONTRACTING OFFICER <i>(Type or print)</i> JONATHAN W. CARNEY			31c. DATE SIGNED		

19. ITEM NO.	20. SCHEDULE OF SUPPLIES/SERVICES	21. QUANTITY	22. UNIT	23. UNIT PRICE	24. AMOUNT
	Accounting Info: SSP4560RBXXXX10-2025-61000001-310202-SSPW709030800 -SSPENBS536-XXXXXXXXXXXX-SSP6011-SSPXXDIG-XXXX-SS PXSHAREDXX-XXXXXXXXXXXX-XXXXXXXX-XXXXXXXX Period of Performance: 06/27/2025 to 06/26/2029				
0001	CLIN 0001: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: 20 Seats POP: 12 months Base Year				
0002	CLIN 0002: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: NTE 20 Seats POP: 12 months Base Year - Optional Item Amount: (Option Line Item) 1				
0003	CLIN 1001: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: 20 Seats POP: 12 months Option Year 1 Continued ...				

32a. QUANTITY IN COLUMN 21 HAS BEEN

☐ RECEIVED ☐ INSPECTED ☐ ACCEPTED, AND CONFORMS TO THE CONTRACT, EXCEPT AS NOTED: _____

32b. SIGNATURE OF AUTHORIZED GOVERNMENT REPRESENTATIVE		32c. DATE	32d. PRINTED NAME AND TITLE OF AUTHORIZED GOVERNMENT REPRESENTATIVE	
32e. MAILING ADDRESS OF AUTHORIZED GOVERNMENT REPRESENTATIVE			32f. TELEPHONE NUMBER OF AUTHORIZED GOVERNMENT REPRESENTATIVE	
			32g. E-MAIL OF AUTHORIZED GOVERNMENT REPRESENTATIVE	
33. SHIP NUMBER	34. VOUCHER NUMBER	35. AMOUNT VERIFIED CORRECT FOR	36. PAYMENT ☐ COMPLETE ☐ PARTIAL ☐ FINAL	37. CHECK NUMBER
☐ PARTIAL ☐ FINAL				
38. S/R ACCOUNT NUMBER	39. S/R VOUCHER NUMBER	40. PAID BY		
41a. I CERTIFY THIS ACCOUNT IS CORRECT AND PROPER FOR PAYMENT		42a. RECEIVED BY (Print)		
41b. SIGNATURE AND TITLE OF CERTIFYING OFFICER		41c. DATE	42b. RECEIVED AT (Location)	
			42c. DATE REC'D (YY/MM/DD)	42d. TOTAL CONTAINERS

CONTINUATION SHEET

REFERENCE NO. OF DOCUMENT BEING CONTINUED
NNG15SD42B/2032H525F00055

ANT-AR-0315

3

4

NAME OF OFFEROR OR CONTRACTOR

ARCHITECHTURE SOLUTIONS LLC

ITEM NO. (A)	SUPPLIES/SERVICES (B)	QUANTITY (C)	UNIT (D)	UNIT PRICE (E)	AMOUNT (F)
	Amount: [REDACTED] (Option Line Item) 1				
0004	CLIN 1002: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: NTE 20 Seats POP: 12 months Option Year 1 - Optional Item Amount: [REDACTED] (Option Line Item) 1				[REDACTED]
0005	CLIN 2001: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: 20 Seats POP: 12 months Option Year 2 Amount: [REDACTED] (Option Line Item) 1				[REDACTED]
0006	CLIN 2002: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: NTE 20 Seats POP: 12 months Option Year 2 - Optional Item Amount: [REDACTED] (Option Line Item) 1				[REDACTED]
0007	CLIN 3001: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: 20 Seats POP: 12 months Option Year 3 Amount: [REDACTED] (Option Line Item) 1				[REDACTED]
0008	CLIN 3002: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: NTE 20 Seats POP: 12 months Option Year 3 - Optional Item Amount: [REDACTED] (Option Line Item) 1 Continued ...				[REDACTED]

CONTINUATION SHEET

REFERENCE NO. OF DOCUMENT BEING CONTINUED
NNG15SD42B/2032H525F00055

ANT-AR-0316

PAGE 4 OF 4

NAME OF OFFEROR OR CONTRACTOR

ARCHITECHTURE SOLUTIONS LLC

ITEM NO. (A)	SUPPLIES/SERVICES (B)	QUANTITY (C)	UNIT (D)	UNIT PRICE (E)	AMOUNT (F)
0009	CLIN 4001: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: 20 Seats POP: 12 months Option Year 4 Amount: (Option Line Item) 1				
0010	CLIN 4002: Figma for Government (PA) - Per Full Seat 100T-FIGMAFULL-GPA QTY: NTE 20 Seats POP: 12 months Option Year 4 - Optional Item Amount: (Option Line Item) 1 The total amount of award: . The obligation for this award is shown in box 26.				

U.S. Federal Housing, Fannie Mae and Freddie Mac are terminating all use of Anthropic products, including the use of its Claude platform.



Rapid Response 47   @RapidResponse47 · Feb 27



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!

From: [Fletcher, Kelly E](#)
To: [REDACTED] [Ritualo, Amy](#); [REDACTED]
Subject: Re: Anthropic
Date: Friday, February 27, 2026 9:32:57 PM
Attachments: [image001.gif](#)

[REDACTED]
[REDACTED]

[REDACTED]

SENSITIVE BUT UNCLASSIFIED

From: [REDACTED]
Sent: Friday, February 27, 2026 8:30:27 PM
To: Fletcher, Kelly E [REDACTED]; Ritualo, Amy [REDACTED]
[REDACTED]
Subject: Re: Anthropic

[REDACTED]
[REDACTED]
[REDACTED]

SENSITIVE BUT UNCLASSIFIED

From: Fletcher, Kelly E [REDACTED]
Sent: Friday, February 27, 2026 9:25:07 PM
To: Ritualo, Amy [REDACTED]
[REDACTED]
Subject: Fw: Anthropic

[REDACTED]
[REDACTED]

[REDACTED]

SENSITIVE BUT UNCLASSIFIED

SENSITIVE BUT UNCLASSIFIED

ANT_AR-0318

From: Holler, Daniel J [REDACTED]
Sent: Friday, February 27, 2026 6:17 PM
To: Fletcher, Kelly E [REDACTED]
Cc: Kenna, Lisa D [REDACTED]; Rhodes, Matthew [REDACTED]; Sprenger, Austin J [REDACTED]
Subject: Re: Anthropic

Great. Thank you. Please give me a run down of the specialized apps so I understand the full exposure. Thanks.

SENSITIVE BUT UNCLASSIFIED

From: Fletcher, Kelly E [REDACTED]
Sent: Friday, February 27, 2026 5:51:43 PM
To: Holler, Daniel J [REDACTED]
Cc: Kenna, Lisa D [REDACTED]; Rhodes, Matthew [REDACTED]; Sprenger, Austin J [REDACTED]
Subject: Re: Anthropic

Sir,

Anthropic is primarily used in StateChat, the Department's internal generative AI resource. We can block access to the Anthropic model by COB Monday in StateChat. We will replace the Anthropic model with a GPT model. Most users will not notice this switch. The main difference is the model training date, with the GPT model having less recent training.

There are a few specialized apps also using Anthropic models. Transitioning these apps to new models will require 1-2 weeks (well within the 6 month timeframe).

We in DT are working with GA to clean up our contracts to ensure that Anthropic is no longer in receipt of State funds, even as a sub.

VR,

Kelly

SENSITIVE BUT UNCLASSIFIED

ANT_AR-0319

SENSITIVE BUT UNCLASSIFIED

From: Holler, Daniel J [REDACTED]
Sent: Friday, February 27, 2026 5:40 PM
To: Fletcher, Kelly E [REDACTED]
Cc: Kenna, Lisa D [REDACTED]; Rhodes, Matthew [REDACTED]; Sprenger, Austin J [REDACTED]
Subject: Anthropic

POTUS announce every agency need to cease use of Anthropic' s technology.

[REDACTED]

[REDACTED]

POTUS post: <https://x.com/TrumpWarRoom/status/2027486468097380613>

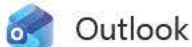
Dan Holler

Chief of Staff

U.S. Department of State



SENSITIVE BUT UNCLASSIFIED



Fw: StateChat

From Fletcher, Kelly E [REDACTED]

Date Fri 2/27/2026 5:08 PM

To [REDACTED] Ritualo, Amy [REDACTED]
[REDACTED]

We need to remove the Anthropic model from StateChat rapidly and mod the Palantir contract.

I defer to Amy on technical implementation. For more complex technical modifications, we must meet the 6 month timeline.

[REDACTED] please work with GA on contract mods.

Thanks all,

Kelly

From: Fletcher, Kelly E [REDACTED]

Sent: Friday, February 27, 2026 5:05:13 PM

To: Lewin, Jeremy [REDACTED]

Cc: Sprenger, Austin J [REDACTED] Strom, Rudolf N [REDACTED]

Subject: Re: StateChat

Ok. We're on it.

I believe we'll be on GPT models in StateChat by end of next week.

From: Lewin, Jeremy [REDACTED]

Sent: Friday, February 27, 2026 4:41:40 PM

To: Fletcher, Kelly E [REDACTED]

Cc: Sprenger, Austin J [REDACTED]; Strom, Rudolf N [REDACTED]

Subject: Re: StateChat

Thanks, Kelly. Plus Austin and Rudy. Per POTUS, we need to cancel everything with Anthropic ASAP and strip them out as a sub. Rudy can advise on contracting actions but we should probably do T4C on any direct contracts and ask for a mod on Palantir to exclude Anthropic models on the Omni awards.

[REDACTED]

From: Fletcher, Kelly E [REDACTED]
Sent: Friday, February 27, 2026 8:09 AM
To: Lewin, Jeremy [REDACTED]
Subject: Re: StateChat

Jeremy,

[REDACTED]

[REDACTED]

Thanks much,

Kelly

From: Lewin, Jeremy [REDACTED]
Sent: Thursday, February 26, 2026 6:46 PM
To: Fletcher, Kelly E [REDACTED]
Subject: StateChat

Hi Kelly,

[REDACTED]

Thanks,
Jeremy

[REDACTED]
[REDACTED]
[REDACTED]
Updates 2/28

- Pentagon has labeled Anthropic a supply chain risk: <https://www.axios.com/2026/02/27/ai-trump-supply-chain-anthropic-pentagon-blacklist>
- Anthropic has filed a lawsuit: <https://www.wired.com/story/anthropic-supply-chain-risk-shockwaves-silicon-valley/>
- OpenAI signed a deal with Pentagon: <https://www.cnn.com/2026/02/27/tech/openai-pentagon-deal-ai-systems>
- GSA has cancelled OneGov with Anthropic: <https://www.gsa.gov/about-us/newsroom/news-releases/gsa-stands-with-president-trump-on-national-sec...>

Friday, February 27



Fletcher, Kelly E 8:02 AM

If we were going to move off anthropic models in statechat, how long would that take?

9:04 AM

We have been prepping for the last several days. We could move off of Anthropic in under 2 hours with a short tail of testing to confirm performance.



Fletcher, Kelly E 9:09 AM

oh great.

This is really good!

I think do nothing for now...

9:11 AM

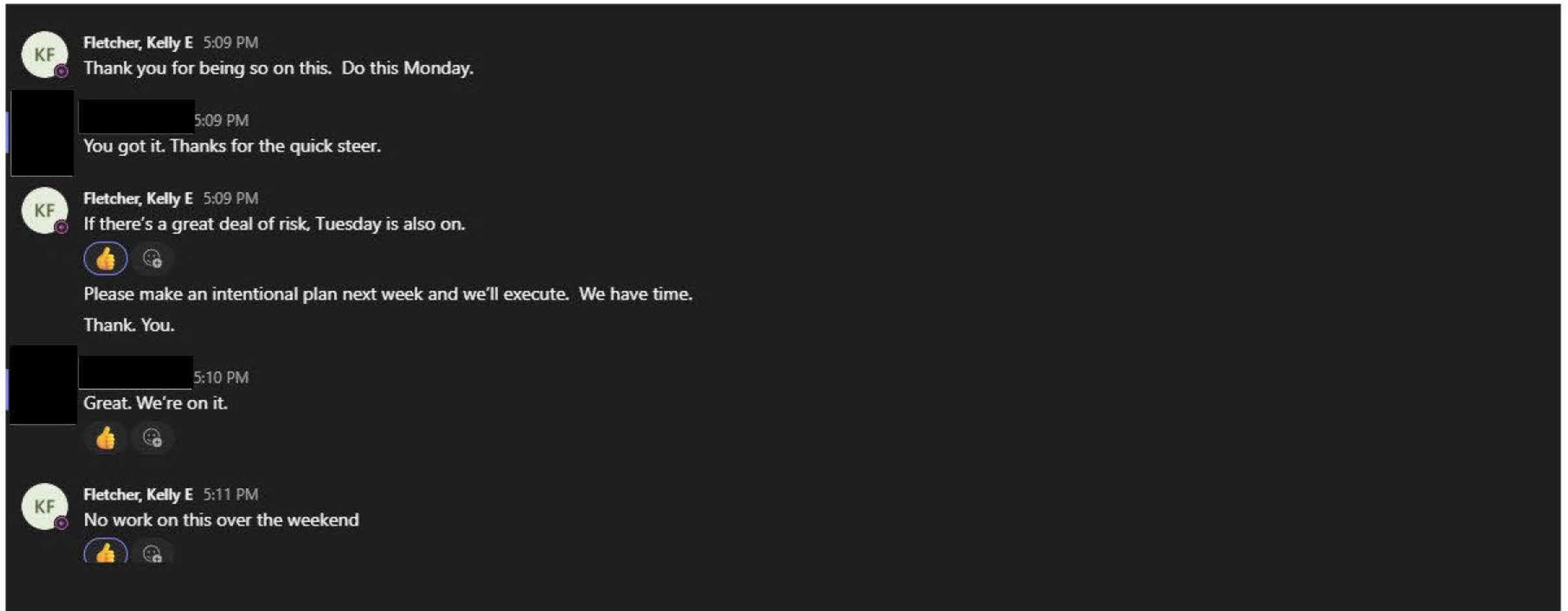
Yes, we're in standby mode just in case.



4:39 PM

News out on Anthropic: POTUS said 6-month phase out and immediate in the same Truth Social message (see below).





Monday, March 2



added Ritualo, Amy and Fletcher, Kelly E to the chat.



3:32 PM

Fletcher, Kelly E change requests to move StateChat to GPT4.1 underway now. Immediate messaging on StateChat and Custom GPTs prepped



KF



Fletcher, Kelly E 3:32 PM

Good job

Thank you

3:40 PM

And GPT-4.1 is live in StateChat. Comms will be on the wires in a moment. Banner added to top of StateChat main screen alerting on need for customGPT owners to change away from Claude.



Version:1.14

Production Change Request Form

PCR ID : 2061

For more information on how to fill out this form, Please see the SOP [Here](#)

Current Status:

Implemented and Closed

Please review the information submitted below. The Product Teams will review the updates and comments provided and will work with appropriate SMEs to finalize the PCR prior to releasing it for approval. Required fields marked with(*)

Draft

PCR in draft status

Review

PCR submitted for review

Approval

PCR submitted for approval

Closed

PCR has been closed

Subject

StateChat model change away from Anthropic + platform banners

Proposed Implementation Date

3/2/2026

Product

CFA PFCS

Product Owner

Proposed Implementation Date

3/2/2026

Product

CFA PFCS

Product Owner

Enclave

Unclassified (OpenNet)

Environment

Cloud Service Model

Priority

Normal

Main Category

Configuration Update

Sub Category

Build Version

v0.618

Associated Tracking

Requested Change

Switch StateChat's backing model away from Anthropic models to GPT 4.1, and announce this change to users via banne

Implementation Plan

SENSITIVE BUT UNCLASSIFIED

ANT_AR-0327

Switch **StateChat's** backing model away from Anthropic models to GPT 4.1, and announce this change to users via banne

Implementation Plan

Release changes to all users.

Implementation Impact

No downtime expected.

Additional Information

Attachments

 2026-03-02 Moratorium Exception Request Change of StateChat Model.msg

Is an approval from a DoS Governance Board Required? Off

If so, which Board(s)?

See List of all applicable governance boards [Here](#)

Please click the Review Complete button for your applicable section below:

User Experience Review Complete	User Experience Review	User Experience Review Date
		12/31/2001 00 : 00
Developer Review Complete	Developer Review	Developer Review Date
		3/2/2026 00 : 00
Tester Review Complete	Tester Review	Tester Review Date
		3/2/2026 14 : 27
Scrum Master Review Complete	Scrum Master Review	Scrum Master Review Date
		12/31/2001 00 : 00
Security / ISSO Review Complete	Security ISSO Review	Security / ISSO Review Date
		12/31/2001 00 :00
	Svstem Engineer Review	System Enaineer Review Date

Submit for Approval

Approve PCR

Product Owner Vote

Product Owner Vote Date

3/2/2026

15

: 06

Governance Tracking

Have all required DoS governance boards approved this change?

Implementation Notes

Implemented By

Implemented On

3/2/2026

15

: 09

Cancel PCR

Cancelled By

Cancelled On

12/31/2001

00

: 00

Security / ISSO Review Complete

System Engineer Review

System Engineer Review Date

12/31/2001

00

: 00

Approval Comments

Submit for Approval

Approve PCR

Product Owner Vote

Product Owner Vote Date

3/2/2026

15

: 06

Governance Tracking

Have all required DoS governance boards approved this change?

Implementation Notes

Implemented By

Implemented On

From: [StateChat Team](#)
To: [REDACTED]
Subject: [External] Notice: StateChat Moving to GPT4.1 Model
Date: Monday, March 2, 2026 4:32:16 PM

Important updates to
StateChat.

No images? [Click here](#)



[Log In](#) | [Subscribe](#) | [Join Us on Teams](#)



Hello colleagues -

Today we are moving our StateChat model to

GPT 4.1 from Claude 4.5 Sonnet.

No action is necessary on your part unless you are the owner of a customGPT (see below).

- For now, StateChat will use GPT4.1 from OpenAI. We will communicate about other model availability in the future.
- The primary effect of this change is that StateChat's model's training data will move back to May 2024.
- For customGPT owners, if you are using Claude Sonnet 4.5 or Claude Sonnet 3.7 to power a customGPT, please select another model by COB Friday, March 6.

Please direct your questions to
StateChat@state.gov.

- The StateChat Team



This email was sent by the U.S. Department of State. If you no longer would like to receive these messages, unsubscribe below.

[Unsubscribe](#)



[External] Action Required: customGPTs Using Claude Must Be Switched by Friday, March 6

From StateChat Team <StateChat@state.gov>

Date Mon 3/2/2026 5:42 PM

To [REDACTED]

If your customGPT uses Claude Sonnet 4.5 or 3.7,
you must select a new
model by COB Friday,
March 6.

No images? [Click here](#)



[Log In](#) | [Subscribe](#) | [Join Us on Teams](#)

Announcement

Colleagues -

If you are using Claude Sonnet 4.5 or Claude Sonnet 3.7 to power a customGPT, please select another model by **COB Friday, March 6**.

If you need assistance, please contact StateChat@state.gov.

Thanks for your help,

- The StateChat Team



This email was sent by the U S Department of State. If you no longer would like to receive these messages, unsubscribe below.

[Unsubscribe](#)

SENSITIVE BUT UNCLASSIFIED

ANT_AR-0332

[REDACTED]

Tuesday, March 3

[REDACTED]

We have also scrubbed all sites and future trainings of mentions, pending a few document mentions on sites that we are still tracking down.





TO send from HHS OCAIO to Snr liasons for AI

From Sharma, Arman (OS/IOS) <Arman.Sharma@hhs.gov>

Date Mon 3/2/2026 4:44 PM

To Dempsey, Ryan (OS/ASTP) (CTR) <Ryan.Dempsey@hhs.gov>

All—

We would like to provide clarification regarding last Friday's message about Anthropic tools, including Claude.

At this time, HHS has temporarily disabled enterprise access to Claude. However, we are still awaiting more detailed federal guidance regarding the future use of applications and systems that leverage Claude or other Anthropic technologies.

To be clear: **no immediate technical or structural changes are required at this time.** You do not need to remove or discontinue use of tools solely based on the previous message.

We are simply asking teams to begin contingency planning to ensure continuity of operations should future federal guidance require a transition away from Anthropic-supported technologies. This planning should focus on understanding dependencies and identifying potential alternatives — not on implementing changes now.

We will provide additional direction as soon as more definitive guidance becomes available.

If you have questions, please contact hhs.ocaio@hhs.gov.

Thank you for your patience and continued diligence.



Suspending of Enterprise Claude Access

From Sharma, Arman (OS/IOS) <Arman.Sharma@hhs.gov>

Date Fri 2/27/2026 4:34 PM

To HHS Senior Liaisons for AI <OCAIO.SLAI@hhs.gov>

Cc Minor, Clark (OS/IOS) <Clark.Minor@hhs.gov>; Hong, David (OS/IOS) <David.Hong@hhs.gov>

All—

For your awareness, HHS OCIO will be disabling enterprise Claude in alignment with President Trump's directive for federal agencies to cease use of Anthropic technology. Specifically, we will be disabling SSO w/ entra, which means accounts and user data/history will be preserved. An all-staff communication will go out first thing Monday morning to ensure maximum visibility, but recognizing that you all are/will likely be bombarded with questions, I wanted to make sure you were all read in.

Please feel free to reach out with any questions; appreciate all of you and your continued efforts.

Best,
Arman

--

Arman Sharma
Deputy Chief AI Officer | Senior Advisor
Immediate Office of the Secretary
U.S. Department of Health and Human Services

Confidential, iterative, pre-decisional



Re: Anthropic

From McDermott, John (OS/OCIO/Ops) <john.mcdermott@hhs.gov>

Date Fri 2/27/2026 5:24 PM

To McFarland, Michael (OS/OCIO/IO) <Michael.McFarland@hhs.gov>; Hong, David (OS/IOS) <David.Hong@hhs.gov>; Minor, Clark (OS/IOS) <Clark.Minor@hhs.gov>; Sharma, Arman (OS/IOS) <Arman.Sharma@hhs.gov>; OCIO Executive Directors <OCIOExecutiveDirectors@hhs.gov>

SSO has been disabled

John

Get [Outlook for iOS](#)

From: McFarland, Michael (OS/OCIO/IO) <Michael.McFarland@hhs.gov>

Sent: Friday, February 27, 2026 4:15:20 PM

To: Hong, David (OS/IOS) <David.Hong@hhs.gov>; Minor, Clark (OS/IOS) <Clark.Minor@hhs.gov>; Sharma, Arman (OS/IOS) <Arman.Sharma@hhs.gov>; OCIO Executive Directors <OCIOExecutiveDirectors@hhs.gov>

Subject: RE: Anthropic

Apparently Sam Altman made a statement that OpenAI shares Anthropics red lines over military AI use.

Michael McFarland
Acting Executive Officer
ASA, Office of the Chief Information Officer
Department of Health and Human Services
202-868-9447

From: Hong, David (OS/IOS) <David.Hong@hhs.gov>

Sent: Friday, February 27, 2026 2:11 PM

To: Minor, Clark (OS/IOS) <Clark.Minor@hhs.gov>; Sharma, Arman (OS/IOS) <Arman.Sharma@hhs.gov>; OCIO Executive Directors <OCIOExecutiveDirectors@hhs.gov>

Subject: Re: Anthropic

Clark,

John and I discussed this topic this morning. John's team can act if we need to.

David

David Hong
Office of Chief Information Officer

ANT_AR-0336

Department of Health & Human Services

Office of Secretary

771-241-7858

David.hong@hhs.gov

From: Minor, Clark (OS/IOS) <Clark.Minor@hhs.gov>

Sent: Friday, February 27, 2026 4:02:12 PM

To: Sharma, Arman (OS/IOS) <Arman.Sharma@hhs.gov>; OCIO Executive Directors <OCIOExecutiveDirectors@hhs.gov>

Subject: Anthropic

[@Sharma, Arman \(OS/IOS\)](#) / [@OCIO Executive Directors](#)

Given the President's recent statement regarding Anthropic, please be prepared to take action / send out guidance to HHS. I've already gotten a lot of inbound, so perhaps we should just let people know that we are aware / are tracking / awaiting official guidance.

Clark Minor

Chief Information Officer

Immediate Office of the Secretary (IOS)

Department of Health & Human Services (HHS)

Mobile: +1 202-412-9408



TO SEND to division CIOs this afternoon

From Sharma, Arman (OS/IOS) <Arman.Sharma@hhs.gov>
Date Mon 3/2/2026 4:50 PM
To Flanders, Bobby (OS/OCIO/IO) <Bobby.Flanders@hhs.gov>
Cc Hong, David (OS/IOS) <David.Hong@hhs.gov>

All,

We would like to provide clarification regarding last Friday's message about Anthropic tools, including Claude.

At this time, HHS has temporarily disabled enterprise access to Claude. However, we are still awaiting more detailed federal guidance regarding the future use of applications and systems that leverage Claude or other Anthropic technologies.

To be clear: no additional technical or structural changes are required outside of staff's direct use of Claude.

We are simply asking teams to begin contingency planning to ensure continuity of operations should future federal guidance require a transition away from Anthropic-supported technologies. This planning should focus on understanding dependencies and identifying potential alternatives — not on implementing changes now.

We will provide additional direction as soon as more definitive guidance becomes available.

If you have questions about a specific use case or system, please contact hhs.caio@hhs.gov.

Thank you for your patience and continued diligence.

From: [Szczepanik, Valerie](#)
To: [White, Joshua](#)
Subject: FW: FYI
Date: Friday, February 27, 2026 5:00:00 PM

From: Szczepanik, Valerie

Sent: Friday, February 27, 2026 4:56 PM

To: Gimbrere, Peter F ; Arietti Gold, Charlene ; Hickman, Jed ; Gerig, Austin

Subject: FYI

[Trump orders federal agencies to phase out use of Anthropic technology | AP News](#)

Should we expect guidance from OMB?

From: [Gimbrere, Peter F](#)
To: [Szczepanik, Valerie](#)
Cc: [Arietti Gold, Charlene](#)
Subject: Claude
Date: Monday, March 2, 2026 12:44:43 PM
Attachments: [image001.png](#)
[image002.png](#)

Hi Val,

Please pull Claude off the website and make sure it is no longer visible or accessible to be used by anyone, but do NOT cancel the contract or remove it from our IT systems until further notice.

Thanks.

Peter

Peter Gimbrere

Managing Executive

Office of the Chairman

OFFICE +1 202 551 4628

MOBILE +1 202 213 3066

gimbrerep@sec.gov

Title: SEC logo



From: [Szczepanik, Valerie](#)
To: [Enriquez, Marco](#)
Subject: FW: Claude
Date: Monday, March 2, 2026 2:32:00 PM
Attachments: [image002.png](#)
[image004.png](#)
[image005.png](#)
[image006.png](#)
[image001.png](#)

Just FYI.

From: Szczepanik, Valerie
Sent: Monday, March 2, 2026 2:32 PM
To: Gray, David ; Hickman, Jed
Cc: Hunt, Bill
Subject: RE: Claude

The direction is to proceed in pushing out the code change as soon as practicable, without waiting for the next scheduled cycle. Thank you for working with your teams to get this done. I will let Marco know about the SMART tool.

From: Gray, David <GrayDa@SEC.GOV>
Sent: Monday, March 2, 2026 2:19 PM
To: Szczepanik, Valerie <SzczepanikV@SEC.GOV>; Hickman, Jed <hickmanj@SEC.GOV>
Cc: Hunt, Bill <HuntW@SEC.GOV>
Subject: RE: Claude

Val,

We can disable access to Claude in ACES without too much effort. However, this will cause any system currently specifying Claude to start taking API errors. We are seeing production usage for Claude for ENF (Smart), OCDO (DCStats), TM and EDW. There is a lot more Claude usage in lower lifecycles, but I am less concerned about them.

We will reach out to the production teams and start the conversations around when they can transition their systems to use another model which we'll determine when we can block ACES wide. We will also send a general notification to the ACES community about the upcoming removal of Anthropic models.

Dave Gray

ACES Branch Chief
 Cloud Center of Excellence
 Office of Information Technology
MOBILE +1 202 230 7740
grayda@sec.gov

Title: SEC logo



From: Szczepanik, Valerie <SzczepanikV@SEC.GOV>
Sent: Monday, March 2, 2026 1:35 PM
To: Gray, David <GrayDa@SEC.GOV>; Hickman, Jed <hickmanj@SEC.GOV>
Subject: FW: Claude
Importance: High

Hi Dave/Jed – see below from the Chairman’s office. I’m asking Marco to disable Claude in SMART, but I think we also have to revoke any entitlements/access to the Anthropic models through AWS/ACES. Would you be able to do that?

Let me know if we need a call.

Thanks!

Val

From: Gimbrere, Peter F <gimbrerep@SEC.GOV>

Sent: Monday, March 2, 2026 12:45 PM

To: Szczepanik, Valerie <SzczepanikV@SEC.GOV>

Cc: Arietti Gold, Charlene <ariettigoldc@SEC.GOV>

Subject: Claude

Hi Val,

Please pull Claude off the website and make sure it is no longer visible or accessible to be used by anyone, but do NOT cancel the contract or remove it from our IT systems until further notice.

Thanks.

Peter

Peter Gimbrere

Managing Executive

Office of the Chairman

OFFICE +1 202 551 4628

MOBILE +1 202 213 3066

gimbrerep@sec.gov

Title: SEC logo



From: [Szczepanik, Valerie](#)
To: [Gray, David](#); [Hickman, Jed](#)
Subject: FW: Claude
Date: Monday, March 2, 2026 1:35:00 PM
Attachments: [image001.png](#)
[image002.png](#)
Importance: High

Hi Dave/Jed – see below from the Chairman’s office. I’m asking Marco to disable Claude in SMART, but I think we also have to revoke any entitlements/access to the Anthropic models through AWS/ACES. Would you be able to do that?

Let me know if we need a call.

Thanks!

Val

From: Gimbrere, Peter F
Sent: Monday, March 2, 2026 12:45 PM
To: Szczepanik, Valerie
Cc: Arietti Gold, Charlene
Subject: Claude

Hi Val,

Please pull Claude off the website and make sure it is no longer visible or accessible to be used by anyone, but do NOT cancel the contract or remove it from our IT systems until further notice.

Thanks.

Peter

Peter Gimbrere

Managing Executive

Office of the Chairman

OFFICE +1 202 551 4628

MOBILE +1 202 213 3066

gimbrerep@sec.gov

Title: SEC logo



From: [Tant, Jason](#)
To: [Szczepanik, Valerie](#); [Kaplan, Andrew](#); [Luh, Daniel](#)
Cc: [Query, Noel](#)
Subject: FW: ACES Notification: Disabling Anthropic Bedrock Models
Date: Tuesday, March 3, 2026 8:39:47 AM

Good morning Val, Andrew, and Daniel,

Should we also be removing the AI bypass rules in the firewall for claude.ai and any claude-associated applications?

Thanks in advance,

Jason

From: AskACES

Sent: Monday, March 2, 2026 3:21 PM

To: #ACES Notifications <#ACESNotifications@SEC.GOV>

Subject: ACES Notification: Disabling Anthropic Bedrock Models

Good Day,

The Chief AI Officer (CAIO) has directed the ACES team to disable the Anthropic Models in the ACES environment. We are currently coordinating with a couple applications using Anthropic models in production. Soon after those systems are updated to no longer use Anthropic, the ACES team will be push an SCP update to disable access to Anthropic models across all of ACES.

Change: Disabling of Anthropic Bedrock Models

What's Happening?

- The ACES Team will update the Service Control Policies (SCPs) to block ***anthropic.claude.***.
- Any system making use of Bedrock calls to Anthropic models will need to be updated to use another model.

Release Dates:

Production: 03/02/2026 – 9pm (pending updates to production systems currently using Anthropic models)

☒ **Action Needed:**

All systems using Anthropic models in Bedrock must specify new models. Any new requests for Anthropic model enablement will be denied.