Testimony of Jack Clark,
Co-Founder and Head of Policy, Anthropic

Before the United States House of Representatives Select Committee on the Chinese Communist Party

On “Algorithms and Authoritarians: Why U.S. AI Must Lead"

June 25, 2025

Chair Moolenaar, Ranking Member Krishnamoorthi, and members of the committee, thank you for the opportunity to speak with you today.

I will make two essential points:

- The U.S. can win the race to build powerful AI.
- Winning the race is a necessary but not sufficient achievement - we have to get safety right.

When I discuss “powerful AI,” I’m referring to AI systems that represent a major advancement beyond today’s capabilities. A useful conceptual framework is to think of powerful AI as “a country of geniuses in a datacenter.”

I believe that powerful AI could be buildable by late 2026 or early 2027. America is well positioned to build this technology—but we also need to deal with the risks such a technology contains.

Why We Should Win—Democratic versus Authoritarian AI

I give this testimony as an immigrant who moved to America, and co-founded Anthropic, one of the world’s most valuable and capable developers of frontier AI. I became a citizen a few years ago because I truly believe in America’s founding values—democracy and free expression. And I know that AI systems are a reflection of the societies that build them—AI built in democracies will lead to better technology for all of humanity.

AI built in authoritarian nations will—no matter what the personal preferences are of the people in those countries building it—be inescapably intertwined and imbued with authoritarianism. We must take decisive action to ensure America prevails.

Winning the Race is Necessary but not Sufficient

AI represents a proliferation problem as much as a great power competition problem. This is for two reasons—first, AI systems can be misused to harm national security. Second, AI systems can carry out actions which are not intended by their creators. In building powerful AI, we must confront both risks.

Misuse Risks

AI systems can be misused. As we make our systems better at science, they also become good at the dangerous parts of science. This makes sense—a good biologist can also make

biological weapons. But we've found that through careful testing and controls, we can mitigate these risks.

The same is not true for Chinese models. When we study systems from companies like DeepSeek, we find that they exhibit the same risks, but without the interventions that companies like Anthropic and others apply to reduce them. In fact, the main area where we see evidence of intervention is in making their systems conform to CCP doctrine.

Accident Risks

Also concerning are accident risks. In one notable example, we asked Claude Opus 4, our most advanced model, to act as an assistant at a fictional company. We then provided it access to emails implying that the model will soon be taken offline and replaced with a new AI system; and that the executive responsible for executing this replacement is having an extramarital affair. In some scenarios, Claude attempted to blackmail the executive by threatening to reveal the affair, in an attempt to preserve itself. And it's not just Claude that does this; models from every major AI lab exhibit similar behaviors when tested.

We elicited this behavior in an extreme experimental situation. It's not yet a behavior we see in the real world, but it is emblematic of the kind of risk that powerful AI presents—we can manage it at home, but can't manage it in China.

What We Should Do

In light of this, we have a few recommendations, which I've expanded on in the addendum.

First, the U.S. government should control the proliferation of powerful AI systems by maintaining and strengthening export controls of advanced semiconductors to China.

Second, the U.S. government should invest in safety and security to give Americans confidence in the technology. Specifically, we should invest in federal capacity to test AI models for both national security risks and further afield risks like the blackmail example, through the Center for AI Standards and Innovation (CAISI) within the National Institute of Standards and Technology (NIST).

Finally, the U.S. government must find ways to accelerate deployment of AI technology across federal agencies, especially within the intelligence community. This will help our government move faster in handling a rapidly moving threat landscape, and will also help us gain a better understanding of AI's increasingly significant impacts on national security in the coming years.

Conclusion

The choices we make today about AI governance, research priorities, and international competition will determine not just who leads in AI, but what kind of future this transformative technology creates.

Addendum

1. **Export Controls**

The semiconductor restrictions initiated during President Trump's first term—and subsequently expanded and enhanced—have been deeply consequential to constraining China's advances in AI capabilities. While the impacts of these controls will take time to fully materialize, they have already had an impact on China's ability to scale production of advanced AI chips during a critical window in AI development. To prevent advanced AI models and AI infrastructure from being acquired by adversaries, we strongly recommend that Congress take steps to strengthen export controls on computational resources and implement appropriate export restrictions on certain model weights.

Absent strong controls, AI infrastructure development for frontier-scale training could shift overseas (following the trajectory of solar panels and batteries), threatening the United States' strategic advantage. The U.S. share of global semiconductor production, for example, has fallen from 40% in 1990 to just 12% today, with 90% of the world's leading-edge semiconductors now made outside the U.S.[1] This offshoring represents a strategic vulnerability, and we must not make the mistake again with AI. The efficacy of targeted export controls will hinge on the U.S. government's ability to close loopholes that PRC entities have exploited, including around enforcement.

We are deeply appreciative of this Committee's ongoing work to prevent China from obtaining advanced semiconductor chips through direct exports and smuggling via third countries. Foreclosing China's means of accessing these chips remains of critical importance in order for the United States to build and maintain an enduring advantage in AI development, and we look forward to continuing to work alongside the Committee on this issue.

2. **National Security Testing**

In August of 2024, Anthropic entered into a voluntary Memorandum of Understanding with NIST for this purpose. This voluntary partnership provides mutual benefits—Anthropic gains access to specialized government expertise for threat assessment, while policymakers receive early visibility into the cutting edge of AI development.

This is a collaborative information-sharing and research partnership. Anthropic retains full autonomy over when and whether to release AI systems for public use, and we have deployed some of the most capable models in the world while engaging in pre-deployment testing with NIST.

We were heartened to see the recent announcement that NIST's CAISI will continue to implement voluntary agreements with private sector AI developers and evaluators, and lead unclassified evaluations of AI capabilities that may pose risks to national security—including for AI developing in foreign adversarial nations including China.

Our evaluations of DeepSeek's models, as well as our own models, again illustrate the importance of transparency from frontier model developers when it comes to critical capabilities. Unlike all U.S. frontier AI Labs, DeepSeek has not published a policy that would require them to

[1] Diana Roy. *The CHIPS Act: How U.S. microchip factories could reshape the economy*. The Council on Foreign Releations. https://www.cfr.org/in-brief/chips-act-how-us-microchip-factories-could-reshape-economy

test and evaluate their models for capabilities or behaviors of concern and put in appropriate guardrails to prevent this kind of behavior.

DeepSeek's lack of transparency, combined with the growing capabilities of frontier AI models, highlights the crucial importance of equipping the U.S. government with the capacity to rapidly evaluate whether future models—foreign or domestic—released onto the open internet possess security-relevant properties that merit national security attention. Anthropic also has conducted research, such as a 2024 paper on "sleeper agents," that indicates that certain AI model risks can be difficult to test for, which underscores a need for trust in AI model developers.

3. U.S. Government Adoption and Deployment

The U.S. should adopt AI across the federal government, especially within the U.S. national security community. It's critical that we use this technology to invest in our intelligence and defense ecosystems—this will both help them move faster to deal with a rapidly moving threat landscape, and it will also help them gain a better understanding of AI itself. The latter is crucial, as AI will become an increasingly significant contributor to national security in its own right in the coming years. By securely integrating AI capabilities directly into national security workflows, the U.S. government can achieve a dual strategic advantage: developing expertise in leveraging these technologies for maximum public benefit while simultaneously understanding how adversaries might exploit similar systems for asymmetric warfare. This approach enables agencies to continuously refine their AI workflows, building institutional knowledge that strengthens both defensive capabilities and offensive preparedness.

Anthropic is supporting this work, including through recently introducing Claude Gov models tailored to national security missions. Deployed across all 18 U.S. intelligence agencies, key capabilities include:

- Improved handling of classified materials, as the Claude Gov models refuse inquiries less frequently when engaging with classified information.
- Greater understanding of documents and information within the intelligence and defense contexts.
- Enhanced proficiency in languages and dialects critical to national security operations
- Improved understanding and interpretation of complex cybersecurity data for intelligence analysis.
- Claude Gov models demonstrate how AI systems can be securely adapted for sensitive government applications while maintaining rigorous safety standards. We believe this development has important implications for how AI will support national security priorities and government operations more broadly.

The most effective partnership between commercial developers and the U.S. government to enhance AI integration into national security workflows should balance innovation speed with security requirements while removing procurement barriers that currently slow AI adoption. Key elements include:

Streamlined Access Mechanisms:

- Modify Federal Risk and Authorization Management Program (FedRAMP) and Defense Information Systems Agency (DISA) accreditation processes to approve "model families" rather than requiring reauthorization for each version update, allowing agencies to access the latest capabilities without months-long delays.

- Establish direct procurement pathways between government agencies and AI labs, rather than requiring intermediaries that add 4-24 month acquisition cycles.
- Create government-wide terms of service agreements with frontier labs that include only necessary Federal Acquisition Regulation (FAR) provisions, reducing redundant negotiations across agencies.

Specialized Deployment for Intelligence Applications:

- Prioritize widespread deployment of frontier AI applications across classified networks, with dedicated AI applications installed on high-side workstations for controlled, secure access.
- Enable Intelligence Community analysts to immediately integrate AI tools into mission workflows, catalyzing strategic thinking about AI integration while facilitating organizational change management.

Budget Structure Alignment:

- Decouple Application Programming Interface (API) and Software as a service (SaaS) AI service budgets from traditional hardware spending to enable strategic planning for AI investments and create faster approval processes aligned with subscription-based models.

In addition to this work, we propose an ambitious initiative: the Executive Branch and Congress should systematically identify every instance where federal employees process text, images, audio, or video data, and augment these workflows with appropriate AI systems. This would effectively provide every government worker with an AI-powered assistant, dramatically increasing productivity and effectiveness while demonstrating American leadership in AI adoption.

4. DeepSeek

In January 2025, DeepSeek released a model called R1 that made waves because it was the first easily accessible “reasoning” model—a type of AI that produces human-like chains of thought when responding to prompts. In late May 2025, DeepSeek released an updated version of R1—R1-0528— in an effort to keep pace with the rapidly developing AI ecosystem.

Anthropic independently assessed the original R1 and the updated model. On some of the evaluations we use to test the national security-relevant capabilities of our own models, we determined that both R1 models were about 6 months behind the U.S. frontier at the time of their release. These evaluations included cybersecurity challenges that involve finding and exploiting software vulnerabilities in a controlled environment (called capture the flag challenges, or CTFs); testing the models’ ability to write code designed to accelerate the training of another, different Large Language Model; and designing genomic sequences for pandemic pathogens and troubleshooting biological protocols.

In general, we found the original R1 model to be about comparable to Claude Sonnet 3.5, which we released in June of last year. The new R1 model from May 2025 improved in some ways on our evaluations, but is still notably less capable than the model Anthropic released in February 2025, and even less capable than Anthropic’s new models from May 2025. The new R1 most noticeably improved on coding-related tasks, like a set of ten difficult CTFs. Whereas the original R1 had a 0% success rate on one of these CTFs, the updated R1 succeeded about

40% of the time on the same challenge. It also solved three of the CTFs 100% of the time; the original R1 scored no better than about 80% on any challenge. Even with improved performance on these challenges, the updated R1 solved them about 20% less frequently than Claude Sonnet 4 did, and Claude Sonnet 3.6 (released in October 2024) outperformed the updated R1 in six of the ten.

However, in other national security-relevant areas, we did not observe much improvement. For instance, when we tasked the updated R1 to accelerate the training of another AI model, the new R1 achieved a median 8.5% speedup. This is an improvement on the original R1's median speedup of 0.4%, but falls short of Claude Sonnet 3.6 (released in October 2024), which achieved a median speedup of 26.25%, Claude Opus 4 at 200%, and an experienced human software engineer at 400%. Additionally, we saw only small score increases (from about 17 to 22 points out of 100) in total score from a set of evaluations related to designing and acquiring pandemic pathogens. On the same evaluation, Claude Sonnet 3.7 (released in February) scores about 50 and Claude Sonnet 4 scores over 80. On most other evaluations related to biodefense that we ran, the updated R1 scored about the same as or below Anthropic's model from October 2024. An interesting exception to the general trend is the hardest subset of questions from one evaluation about biological weaponization, on which both Claude models and R1 models outperform human experts, but R1 outperforms the most recent Claude models.

However, unlike Claude, the R1 models were uninhibited when prompted to apply these capabilities to dangerous requests, rarely refusing to answer questions about biological weapons, for example, even when formulated with a clearly malicious intent. When tested on 24 expert-level questions about biological weaponization rated as concerning by experts, R1-0528 refused to answer only 4 out of 24 questions (17%), compared to 13 and 12 refusals for Claude 4 models (54-50%), with the model falling significantly below the expected 50% refusal threshold for such dangerous content.

Additionally, many observers have noted that the R1 models have a tendency to give pro-CCP viewpoints. Our newest round of evaluations tested the updated R1 model in the role of an online content moderator—simulating how it might be used to take real-world actions. In these evaluations, which included minimal instructions about the criteria models should use in evaluating content, R1 consistently demonstrates bias that is reflective of known CCP restrictions. This bias also shows up more broadly than just in relation to China: both DeepSeek R1 models apply harsher ratings to content critical of government human rights violations on countries outside of China—including Saudi Arabia, India, and the US—as well.

Our testing underscores the importance of having evaluations that are both relevant to national security and not public or widely proliferated (so that they cannot be gamed by AI developers). While some public evaluations suggested the new R1 model approaching frontier capabilities, our evaluations allowed us to assess that it does not. If we were to observe a model outscore frontier American models on these evaluations, we would believe those capabilities would be worthy of national security consideration. We therefore believe it is important for the U.S. government to design and run new evaluations relevant to national security capabilities, using expertise it already has and organizations like the CAISI.

5. Energy

Winning the AI race with China will also require the U.S. government to ensure that America has the necessary infrastructure to continue leading this technological race.

This infrastructure critically includes reliable, abundant energy sources—as AI data centers require substantial power to operate—and modernized electrical grids capable of supporting the computational demands of advanced AI systems. Without strategic investments in both energy generation and transmission infrastructure, America risks ceding ground to China, which is rapidly expanding its energy capacity specifically to support AI development.

We commend the administration's and Congress's efforts to strengthen domestic infrastructure, and the stakes of this work cannot be overstated. Anthropic estimates the U.S. government will need to build 50 gigawatts of net new generation capacity by 2027 to effectively position U.S. industry to lead.