Testimony before the U.S. Senate Committee on the Judiciary

For the hearing titled
"Stealth Stealing: China's Ongoing Theft of U.S. Innovation"[1]
April 22, 2026

Helen Toner
Interim Executive Director
Center for Security and Emerging Technology, Georgetown University

---

Chair Grassley, Ranking Member Durbin, members of the Committee: Thank you for the opportunity to testify before you today.

I have spent the last 7 years working on AI and national security policy at Georgetown University's Center for Security and Emerging Technology (CSET), which I now lead. U.S.-China competition in artificial intelligence (AI) is a major focus of my research, as are questions of AI safety, security, and governance. I served on OpenAI's board of directors from 2021 to 2023.

Outline and Key Takeaways

Given my expertise and background, my testimony focuses on threats to U.S. IP in the AI space in particular. Recently, reports of so-called "distillation attacks" have made a splash in the world of US-China AI competition. The testimony that follows will therefore begin with a discussion of what distillation is, before situating distillation in the broader context of U.S.-China AI competition and IP.

Key takeaways:

- **There is strong evidence that Chinese AI companies are employing distillation techniques** in order to extract capabilities from American AI models in order to advance their own research and development (R&D).

[1] This written testimony is adapted from testimony submitted to a May 2025 hearing before the House Committee on the Judiciary, Subcommittee on Courts, Intellectual Property, Artificial Intelligence, titled "Protecting Our Edge: Trade Secrets and the Global AI Arms Race."

- **Distillation is far from the only factor driving Chinese AI companies' ability to fast-follow their U.S. counterparts, and far from the only threat to U.S. AI companies' IP**. Distillation clearly provides a boost to Chinese AI capabilities, but several other factors—both legitimate and illegitimate—are at least as important. Chinese AI scientists and engineers are genuinely innovative; making policy on the assumption that Chinese advances rest solely on theft would be mistaken. In addition, U.S. AI companies are vulnerable to cyber intrusions, insider threats, and other non-distillation avenues that allow Chinese companies to potentially acquire IP related to AI models, algorithmic secrets, training datasets, and AI hardware.
- **Preventing distillation should be seen as one part of protecting the full pipeline of U.S. AI IP.** Some potential approaches to preventing distillation could be counterproductive, for instance by reducing U.S. companies' commercial competitiveness or redirecting resources from other important threats. **Efforts to counter distillation should focus on broader measures that are also helpful in countering other threats,** e.g. sharing threat intelligence and assisting companies in monitoring for misuse of their models.
- Given the speed, intensity, and high stakes of advanced AI development, **IP concerns must not be used as an excuse not to share information** that could help governments and the public understand and mitigate risks.

The remainder of my written testimony elaborates on these points, and concludes with recommendations for Congress.

What is AI Distillation?

"Distillation" is a term used for a technique whereby outputs from a more advanced AI model (the "teacher") are used to improve a less advanced AI model (the "student"). This approach was originally developed as a way to extract ("distill") knowledge and capabilities from a larger model (or models) in order to create a smaller model with comparable knowledge and capabilities.

Distillation is not inherently nefarious. Since smaller models are cheaper to run and easier to use, major AI companies routinely use distillation on their own most advanced models to produce lighter-weight alternatives. Google's Gemini Flash series of models, for instance, are distilled from larger Gemini models in order to provide customers with a high-speed, low-cost option.[2] It is also commonplace for external researchers, including academics, to use distillation on models from major AI companies as part of their research process. This is largely seen as a

[2] Per Google's Chief AI Scientist Jeff Dean, from Latent Space podcast episode "Owning the AI Pareto Frontier," February 12, 2026.

legitimate practice, so long as the researchers in question are not trying to create models that will compete commercially with those of the AI company in question.

Today, one common use of distillation is as a method for creating synthetic training data. This is somewhat different from the form (originally known as "knowledge distillation") that Google researchers first developed in 2015, which was designed for image and speech recognition models rather than today's large language models.[3] Today, finding or creating training datasets that effectively improve an AI model's capabilities is a major element of cutting-edge AI research. Broadly speaking, the volume of high-quality data used to train a model is a major factor in how advanced the model ends up being, so researchers gravitate towards approaches that make it possible to automatically generate large volumes of training data. In this context, distillation simply means prompting an AI model to generate outputs that can be used as training data for a different model.

In recent months, the three leading U.S. AI companies (OpenAI, Google, and Anthropic) have released information about so-called "distillation attacks" they detected on their services, carried out by Chinese AI firms.[4] The most detailed report (from Anthropic) describes three different Chinese firms (DeepSeek, Moonshot AI, and MiniMax) using over 24,000 accounts for over 16 million interactions that Anthropic judged to be distillation attempts. Because the goal of this usage was to train models that would compete directly with the U.S. companies' offerings, **all three U.S. companies determined that it was in breach of their terms of service,** i.e. a contractual violation.

To carry out adversarial distillation of this kind, perpetrators access AI models that are offered commercially, typically via an application programming interface (API, the standard way that enterprise users access commercial AI models). This means that **distillation is just one example of the broader category of AI misuse**, which providers work to detect and prevent. Other common types of misuse include using AI to perpetrate harm or carry out illegal activity, such as planning and/or executing cyberattacks, generating child sexual abuse material, or engaging in scams. The leading U.S. AI companies take a multi-layer approach to monitoring for, detecting, preventing, and reporting misuse of their models, with mixed success. While misuse detection and prevention are improving over time, **AI companies must manage the trade-off between preventing misuse and facilitating easy access to their products.** Many of

[3] Hinton, Vinyals, and Dean 2015, "Distilling the Knowledge in a Neural Network" (arXiv).
[4] OpenAI 2026, "Letter to US House Select Committee on Strategic Competition between the United States and the Chinese Communist Party."
Google 2026, "GTIG AI Threat Tracker: Distillation, Experimentation, and (Continued) Integration of AI for Adversarial Use."
Anthropic 2026, "Detecting and preventing distillation attacks."

the approaches they might take to clamp down on adversarial distillation and other misuse are likely to also incorrectly flag legitimate activity, creating barriers to normal commercial use.

IP Vulnerabilities in the U.S. AI Pipeline

Adversarial distillation is only one of several threats to U.S. AI companies' IP. In order to design an effective response, it is important to understand the threat from distillation in the broader context of these other challenges. **An overly narrow focus on detecting and preventing distillation could be counterproductive** if it comes at the cost of reduced focus on other threats to U.S. AI companies and U.S. national security.

Within the process of training an AI model, distillation can only be used as part of a larger training pipeline—it cannot be used to "steal" an AI model wholesale. Rather, by leaning on the strengths of U.S. AI models, distillation can allow Chinese AI developers to accelerate or skip certain phases of the process of developing their own models. As described above, one of the most common uses of distillation is as a way to generate training data. On the basis of publicly available information, **it is not clear how significant the role of distillation is in boosting Chinese AI development**. The extent of Chinese distillation efforts, detailed in the previous section, suggest that it provides at least some benefit, but it is unlikely that distillation is the sole—or even primary—factor allowing Chinese AI developers to fast-follow U.S. companies.

To place distillation in context, it is useful to think about how it fits into a broader breakdown of trade secrets and IP vulnerabilities in the AI pipeline. One way of breaking down the trade secrets of concern in frontier AI development is to consider AI models themselves, plus the triad of inputs used to make them: algorithms, data, and hardware. As noted below, training data is the primary area where distillation is relevant.

- **AI models** are frontier AI companies' crown jewels. To steal a model, the attacker needs to gain access to a file containing the set of numbers known as "model weights" or "parameters." For today's leading models, this file is likely on the order of several terabytes in size, meaning exfiltration is non-trivial but possible. If an adversary is able to steal a trained model, they have access to the capabilities of that model without needing the world-class team, months of research time, and tens or hundreds of millions of dollars used to create ("train") the model. They can also flexibly modify or further build on the model. For these reasons, the weights of frontier models are broadly considered the highest priority type of intellectual property for frontier AI companies to protect. As noted above, distillation approaches are insufficient to steal a model.

- **Algorithmic secrets** is an overarching term used to describe privately held insights and techniques used to design and train models. They include model architectures (the design of the model), training "recipes" (the sequence of steps used to optimize the model), nuggets of practical know-how, and new research ideas. Algorithmic secrets often take the form of a few sentences—perhaps a single sentence—that could be extracted from internal documents or messaging platforms. They are frequently transferred between companies (and countries) inside the heads of researchers who move from one employer to another.

- **Training datasets** are the data used to create AI models. For frontier AI models, these include multi-trillion-word corpora of text and other data scraped from the internet and other sources; carefully curated human-written examples of priority use cases; collections of human-submitted or AI-generated rankings and ratings comparing different AI-generated options; AI-generated text; and other types of data. To the extent that distillation allows AI developers to recreate the training data of a competitor for their own use, it could potentially be considered to be acquisition of trade secrets by improper means.[5]

- **AI hardware** provides the computing power (often referred to as "compute") needed to train frontier models. Access to state-of-the-art AI chips, which in some cases are subject to U.S. export controls, is a major competitive advantage for U.S. frontier AI firms. The most sensitive trade secrets associated with AI hardware are chip designs, though these are of limited utility unless the thief can access leading-edge chip production facilities, which are overwhelmingly operated by the Taiwanese Semiconductor Manufacturing Corporation (TSMC).

Trade secrets in these categories have already been stolen from top firms. In 2023, an attacker gained access to internal communication channels at OpenAI and was able to extract proprietary information on how they develop their AI technology, i.e. **algorithmic secrets**.[6] In January of this year, Chinese national Linwei Ding became the first person convicted on AI-related charges under the Economic Espionage Act when he was found guilty of systematically exfiltrating large volumes of confidential information while working at Google in 2022 and 2023. Much of this information related to Google's TPU AI chips, i.e. **AI hardware**.[7] And only two weeks ago, Mercor, a major provider of **training datasets** for leading U.S. AI

[5] For a more detailed discussion of the legal questions involved, see Hrdy 2025, "Trade Secrecy Meets Generative AI" (Chicago Kent Law Review).
[6] Metz 2024, "A Hacker Stole OpenAI Secrets, Raising Fears That China Could, Too" (New York Times).
[7] Department of Justice 2024, "Chinese National Residing in California Arrested for Theft of Artificial Intelligence-Related Trade Secrets from Google."

companies, was subject to a cyber breach. Hackers seized four terabytes worth of data, much of which can likely be used to train advanced AI models.[8]

A 2024 RAND analysis found that there were many areas where frontier AI companies could immediately improve their security, and that there was also a need for significant investments over the longer term to increase their capacity to resist more sophisticated attacks.[9] As leading AI companies' technology becomes increasingly strategically sensitive, the national security implications of inadequate cybersecurity grow more acute than for a typical industry. If we believe that it's vitally important for the world's most advanced AI systems to be developed and deployed by U.S. companies, then we should strive to protect U.S. AI developers from the full pipeline of threats to their IP.

An Aside: The Importance of Visibility Into Frontier AI Development

This testimony, like many discussions of AI in 2026, focuses on so-called "**frontier AI**," which refers to cutting-edge, general-purpose AI systems such as Google's Gemini 3, OpenAI's GPT-5.4, Anthropic's Claude 4.7, and xAI's Grok 4.

Frontier AI is only one part of the larger AI ecosystem, but from a strategic perspective, it is an especially important part. The companies at the frontier are actively working to build artificial general intelligence (AGI), i.e. AI that is as capable as human experts across a wide range of fields. The CEOs of these companies claim that this goal will likely be reached within the next 2-5 years,[10] a view shared by top researchers and engineers both inside and outside of these companies. Once they reach AGI, they plan to push ahead with building "superintelligence," i.e. AI that is far smarter and more capable than humans.[11] Materials produced by frontier AI companies describe a range of severe risks that their own AI systems might soon pose, including aiding in sophisticated offensive cyber operations, enabling amateurs to carry out bioterror attacks, and potentially evading human control entirely.[12] Even if their projected timelines are overly optimistic, it is not an exaggeration to say that this level of AI would reshape the economy, upend the political system, and transform the international order.

---

[8] Kannegieter 2026, "Data Hacks and the US-China AI Race" (ChinaTalk).
[9] Nevo et al. 2024, "Securing AI Model Weights" (RAND).
[10] Anthropic CEO Dario Amodei: "Making AI that is smarter than almost all humans at almost all things [...] is most likely to happen in 2026-2027." OpenAI CEO Sam Altman: "I think AGI will probably get developed during this president's term." Google DeepMind CEO Demis Hassabis: "I think we're probably three to five years away [from AGI]."
[11] E.g. per Sam Altman: "We are beginning to turn our aim beyond AGI, to superintelligence in the true sense of the word."
[12] See Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google's Frontier Safety Framework.

Historically, technologies with such major strategic implications have been developed under the auspices of the U.S. government and closely associated firms. Today, increasingly advanced AI systems are being developed entirely within private industry, with little visibility available to Congress or the executive branch. This threatens the U.S. government's ability to appropriately manage and respond to continued developments in frontier AI, up to and including the possible development of AGI and superintelligence.

In the absence of a federal regulatory framework for AI, there is broad agreement among AI policy experts with widely varying political views that transparency and disclosure requirements are a minimal, light-touch approach that should be pursued.[13] Increasing the information flow between frontier companies and the outside world has a slew of benefits—it reduces information asymmetries, empowers government to understand and respond to advances, and equips the public to weigh in on a technology that will profoundly impact them.

Transparency requirements for AI development can take different forms, depending on the information of interest and the tradeoffs involved in sharing it. Three major types of transparency are disclosure to the *public*, disclosure to *government*, and disclosure to a *third-party auditor.* Each of these balances different pros and cons. Public disclosure goes furthest in reducing information gaps, but may be undesirable for information that is sensitive from a commercial or national security perspective. Disclosure directly to USG is well suited for information directly related to national security, such as results of tests on AI models' capabilities in areas including cyber operations, bioweapon development, and nuclear weapons. Disclosure to third-party auditors is a flexible option that can allow a neutral, independent organization to verify or assess sensitive information while keeping it largely under the AI company's control, e.g. by having the auditor work within the AI company's own facilities under a non-disclosure agreement.

Forcing U.S. companies to disclose information that would damage their competitiveness would not be a good approach. Fortunately, most of the information that is of greatest interest from a national security and public interest perspective would not damage competitiveness if disclosed appropriately. Types of information that would be valuable to share include:

- **Results of testing for AI models' capabilities and risks.** A lack of clear, up-to-date information about what the world's best AI systems are capable of puts the U.S. government at a huge disadvantage in understanding how the AI frontier is progressing and what kinds of responses might be needed. At present, information about how

[13] See, e.g., Ball and Kokotajlo 2024, "4 Ways to Advance Transparency in Frontier AI Development" (TIME).

rapidly AI is advancing is shared (or not shared) at the discretion of companies, on the timeline and in the format that they find most convenient.

- **Information about the goals or specifications AI models are being trained to pursue.** This creates visibility into how companies are making crucial, politically loaded decisions about what their models should and shouldn't do, as well as how to prioritize different values (e.g. freedom of speech vs. avoiding hate speech, supporting users vs. avoiding sycophancy, etc.). As frontier AI models are integrated into citizens' lives, businesses' software, and government's infrastructure, knowing what they are designed to optimize for will become increasingly important. OpenAI and Anthropic voluntarily share a version of this information, to their credit.

- **Analysis of why AI developers believe their current risk management practices are sufficient** (sometimes referred to as a safety case). Developing and releasing frontier AI models involves making a large number of judgment calls about how to test for risks, what results are acceptable, and how to manage uncertainty. Making at least a high-level version of that thinking transparent to a wider audience (perhaps with especially sensitive details redacted) would bring these critical decisions out from closed conference rooms into the sunlight.

- **Data on internal usage of AI and progress towards automating AI R&D.** Frontier AI companies are increasingly relying on their own (sometimes unreleased) AI models to automate their own work. It is a widely held view among AI researchers and CEOs that using AI to accelerate AI research could lead to a runaway feedback loop of increasingly advanced AI. Such a feedback loop is known as an "intelligence explosion," and is at the center of some of AI experts' greatest fears about AI catastrophe. Greater transparency about the extent to which AI is in fact accelerating research would prevent this phenomenon from transpiring in secret.[14]

- **Whistleblower reports.** If AI companies want to claim that they are building the most world-changing technology to ever exist, their employees should be able to share concerns about risks to public safety along the way without fearing the repercussions. AI whistleblowing is largely unprotected by existing whistleblower laws, which focus on outright illegal activity. Chair Grassley's AI Whistleblower Protection Act would be a significant improvement on this status quo.

[14] See Toner et al. 2026, "When AI Builds AI" (Center for Security and Emerging Technology) for more details on automating AI R&D.

Note how little these categories overlap with the trade secrets outlined in the previous section. The categories immediately above are areas where the public or governmental interest in transparency is much higher than the downside of making the information available to competitors. This cost-benefit is different for other types of information. In particular, as outlined in the previous section, information about how AI models are built (e.g. architectural details, training recipes, or full datasets) is more commercially sensitive, as are the models themselves.

In weighing the pros and cons of requiring any particular kind of information be disclosed, it is also important to recall that even commercially sensitive information is already stored inside AI companies themselves, where it is already vulnerable to espionage. This baseline should be considered alongside the sensitivity of the information, the benefit of sharing it, and the security of where it would be stored (e.g. in a classified USG environment vs. on a public website).

Recommendations

1. **Deepen and expand security-focused collaborative arrangements between the U.S. government and frontier AI companies.** The government has access to information that can be helpful to U.S. AI companies (e.g. threat intelligence on adversaries, expertise on threat models such as cyber operations and bioweapon development), and companies have access to information that can be helpful to the government and other companies (e.g. detected patterns of misuse, best practices for detecting and preventing misuse). Several small-scale, voluntary programs currently exist to facilitate this kind of collaboration, including inside the NSA's AI Security Center and at the Center for AI Standards and Innovation (CAISI) within NIST. These initiatives are highly valuable and should be strengthened. The Department of Justice (DOJ) should also ensure that delays in security clearance processing do not unduly prevent employees of leading AI companies from accessing relevant information.

2. **Clarify antitrust guidance to enable U.S. AI companies to collaborate on defensive measures.** DOJ should release guidance clarifying that U.S. AI companies may coordinate their efforts to defend against distillation and other security threats, including other misuses of their models, without facing antitrust concerns. At present, pursuant to the Cybersecurity Information Sharing Act of 2015, AI companies share some information with each other about distillation attacks and other threats.[15] However, private conversations suggest that companies' legal departments recommend

[15] Ghaffary and Eastland 2026, "OpenAI, Anthropic, Google Unite to Combat Model Copying in China" (Bloomberg).

limiting this coordination in order to prevent any antitrust liability. Guidance from DOJ on how it recommends AI companies interpret the statute could unlock stronger coordination and defense measures and thereby strengthen U.S. competitiveness.

3. **Consider removing other legal barriers preventing companies from prioritizing security.** Another place where greater legal clarity would be valuable is around AI companies' ability to engage in intensive personnel vetting and monitoring—practices that may be discouraged under state employment law but that would be essential in any serious effort against insider threats. Crucially, any personnel-focused efforts must be implemented thoughtfully in order to protect the enormous contributions of foreign nationals to U.S. AI competitiveness.[16] What is needed is a scalpel—the ability to carefully identify individuals who should not have access to highly sensitive trade secrets, and keep tabs on those who do have access—not a sledgehammer of anti-immigrant measures.

4. **Avoid interventions that would target distillation at the expense of other priorities.** As outlined above, distillation is a serious concern. However, it is only one piece of the puzzle, and it is unclear both how significant it is and how feasible it is to defend against. Accordingly, policy measures that focus solely on distillation could easily end up being harmful on net, for instance if they harm U.S. companies in other ways (e.g. by restricting how they can offer their products) or if they redirect resources away from more important issues (e.g. focusing on detecting distillation using resources that would otherwise go towards detecting Chinese use of AI for cyberattacks).

5. **Use law enforcement and intelligence collection tools to improve our understanding of Chinese efforts to acquire U.S. AI capabilities.** Are Chinese companies coordinating their distillation efforts, with or without assistance from the government? Are there privately held materials (e.g. emails, messages, memos) that outline plans for distillation campaigns or other attempts to steal U.S. intellectual property or misuse U.S. AI models, e.g. for cyberattacks? Could a "defend forward" approach detect or prevent distillation and other misuse? The FBI and the intelligence community should use their full toolkit to understand, prevent, and disrupt Chinese efforts to illegitimately target U.S. AI companies.

6. **Push frontier AI companies to improve their cybersecurity practices**, and emphasize to them that this must be a high priority. Make clear that the security of critical frontier

[16] See Zwetsloot 2021, "Winning the Tech Talent Competition" (Center for Strategic and International Studies) and Oschinski et al. 2025, "Strengthening America's AI Workforce" (Institute for Progress).

AI IP is as core to U.S. competitiveness as rapid innovation, since the United States cannot lead if adversaries have easy access to our best technology.

7. **Pass the AI Whistleblower Protection Act and seek other ways to use existing or authorities to promote transparency from frontier AI companies about their most advanced systems.** Chair Grassley's Whistleblower Protection Act would significantly improve AI company employees' ability to share information about concerning practices inside their companies. Even without new legislation, Congress can ask AI companies about their compliance with their own voluntary commitments to share safety and security information. The U.S. government can also use existing authorities, such as the Defense Production Act, to require companies to share information via secure channels.

Thank you, and I look forward to your questions.