Is AI distillation a chink in the armour of American innovation?

6 min read | 24 April 2026

Maia Biermann

Shuterstock/Mark Van Scyoc

For much of the past year, concern about adversarial “distillation” circulated through the country’s leading AI labs, cyber security and IP community, but until this week it failed to translate into sustained policy action. A 23 April memorandum from the White House Office of Science and Technology Policy instructing federal agencies to treat "**Adversarial Distillation of American AI Models**” as a strategic threat to US leadership in frontier AI has now pushed the issue beyond industry anxiety into the realm of coordinated government response.

In industry jargon, “model distillation” refers to techniques that use the outputs of a more capable “teacher” model to enhance a weaker “student” model – a practice that is both common and often perfectly legitimate within AI development. Companies use it internally to reduce costs, while researchers rely on it to study model behaviour. As Helen Toner, interim executive director of Georgetown University’s Center for Security and Emerging Technology, told lawmakers this week, distillation is “not inherently nefarious”.

The risk emerges when distillation is used adversarially – at scale, systematically and with the aim of replicating the differentiated capabilities of frontier systems. Speaking a day earlier before the Senate Judiciary Committee at a hearing titled “Stealth Stealing: China’s Ongoing Theft of U.S. Innovation”, Toner **testified** that there is strong evidence Chinese AI firms are using distillation techniques to accelerate their own research by extracting behavioural insights from US models via commercial APIs – paid interfaces that allow developers to access and integrate proprietary models.

Against a backdrop of mounting **political concern** over technology theft, the hearing lent urgency to warnings that access which is contractually authorised can nonetheless be weaponised in practice.

Industry **has not hidden** the fact that much of the innovation behind generative AI models is protected as trade secrets. Yet unlike source code or model weights, AI-generated outputs are intentionally exposed. When those outputs can be harvested at scale, targeted at high-value capabilities, and repurposed as synthetic training data, they may enable rivals to recreate sensitive elements of a proprietary development pipeline without ever accessing its core crown jewels.

Nor have companies been shy about what they are seeing first-hand. Across disclosures and congressional briefings, Anthropic, Google and OpenAI have reported adversarial distillation occurring at industrial scale, evolving rapidly and targeting the most valuable parts of their systems.

Anthropic has **provided** the most granular public detail. It disclosed that three Chinese AI labs – DeepSeek, Moonshot AI and MiniMax – generated more than 16 million exchanges with its Claude model using roughly 24,000 fraudulent accounts. What distinguished this activity, the company said, was not any single interaction but the coordinated pattern aimed at extracting agentic reasoning, advanced coding and tool-use capabilities.

Google has also **identified** a rise in model extraction attempts during 2025, describing distillation as a fundamental shift in IP risk – one in which large models offered as services can be exploited without direct intrusion into internal systems.

OpenAI has **characterised** the threat as persistent and adaptive – and tied closely to geopolitical rivalry. In a memorandum to the US House Select Committee on Strategic Competition between the United States and the Chinese Communist Party, the company said most adversarial distillation activity observed on its platform appears to originate from China, with occasional activity linked to Russia. It singled out DeepSeek as continuing to pursue activity consistent with adversarial distillation.

Indeed, the intensity of the 22 April hearing revived memories of the DeepSeek episode in early 2025, when a Chinese-developed chatbot briefly overtook ChatGPT in US app store rankings, triggering a sharp market reaction. Distillation was **floated** at the time as one possible explanation, but did not translate into formal legal allegations.

One reason distillation has proved so vexing is that it resists traditional IP categories, sitting uncomfortably outside copyright, patent and trademark doctrines and only imperfectly within trade secret law. As Toner observed, today's IP regimes were not designed for systems whose commercial value is distributed across data, algorithms, training processes and hardware optimisation.

Distillation rarely enables wholesale copying. Instead, it allows competitors to bypass stages of development, shaving months or even years off costly R&D pathways. That nuance complicates both legal enforcement and policy response. Heavy restrictions on API access risk harming the US firms whose business models depend on openness, while weaker controls leave valuable capabilities exposed.

Trade secret laws under strain

For IP lawyers, distillation raises uncomfortable questions about the boundaries of trade secret protection. **Michael Ng**, Kobre & Kim's lead trial counsel with experience in AI-related trade secret disputes, argues that providers are "not powerless" even when outputs are widely accessible. Distillation, he tells IAM, implicates at least three aspects of trade secret protection: whether reasonable measures were taken to preserve secrecy, whether the information was readily ascertainable, and whether the means of acquisition were proper.

Contextual flexibility is a defining strength of trade secret law, particularly in how courts assess secrecy in real-world commercial settings. Recent US case law, according to Ng, confirms that scale alone does not defeat secrecy, provided disclosures occur within a framework of confidence supported by clear contractual and operational safeguards.

Companies, he says, can reinforce their "reasonable measures" by explicitly prohibiting distillation in their terms of use, even where outputs are widely accessible. As the Virginia Supreme Court put it in *Appian v Pegasystems*, purported trade secrets can be disclosed to a million people and remain protected if those disclosures are made in confidence.

Ng recommends: "Governments can also act, whether through export controls, regulatory frameworks, or other tools, to restrict or prohibit distillation. That can create independent deterrence, but also reinforce the improper means analysis, making clear that technical feasibility alone does not make distillation lawful."

Matthew Kohel, a partner at Saul Ewing whose practice focuses on trade secret and AI governance, is more sceptical about existing doctrine's reach. He argues that improper means analysis struggles when distillation relies on authorised access via standard APIs. Plaintiffs can show contractual breach, but linking that breach to trade secret misappropriation is harder when what is extracted is behavioural insight rather than a discrete artefact.

Kohel sees a "significant structural gap", explaining that trade secret law evolved to prevent theft of existing information, not the recreation of value through synthetic data

generated at scale. As he tells IAM: “If the access is authorised and the method resembles a high-volume version of legitimate research practices, plaintiffs may struggle to demonstrate it as improper under existing trade secret statutes.”

He adds, however, that where an adversary deploys thousands of fraudulent accounts to evade rate limits or usage restrictions, the conduct begins to resemble “the type of improper means that courts have traditionally found in other digital contexts”. Where the objective is to extract proprietary training data or behavioural insights for a competing product, the use of automated and deceptive account creation may be “viewed as circumvention of the ‘reasonable measures’ companies take to protect their trade secrets,” notes Kohel.

Yet both experts converge on one point: companies cannot address the challenge alone, whether through litigation or tighter internal controls as much as distillation exposes weaknesses in the entire AI IP pipeline. Kohel recommended doctrinal evolution supported by legislative reform, including clearer rules that allow companies to share threat intelligence collaboratively without antitrust risk, and protections for whistleblowers who surface internal misuse early.

“The response must be a broader strengthening of the entire IP pipeline against cyber intrusions and insider threats while maintaining the openness that fuels US innovation”, he said, adding that an overly narrow focus on distillation alone could be counterproductive.

The government’s willingness to intervene leaves American AI innovators on firmer ground.

The 23 April memorandum, issued by Michael Kratsios, Assistant to the President for Science and Technology, warns that foreign actors, “principally based in China”, are conducting industrial-scale campaigns to distil US frontier AI systems, using tens of thousands of proxy accounts, jailbreaking techniques and coordinated evasion methods. Read alongside the 22 April Senate Judiciary Committee hearing, where distillation was repeatedly cited as a new vector of IP loss, the memo signals a decisive shift in Washington’s thinking.

Policymakers have begun to view distillation not as an isolated cybersecurity concern, but as a national security challenge that exploits the openness underpinning American innovation. As Anthropic has observed in its 23 February **announcement**, “no company can solve this alone”.

Maia Biermann

Senior Reporter

IAM

maia.biermann@lbresearch.com