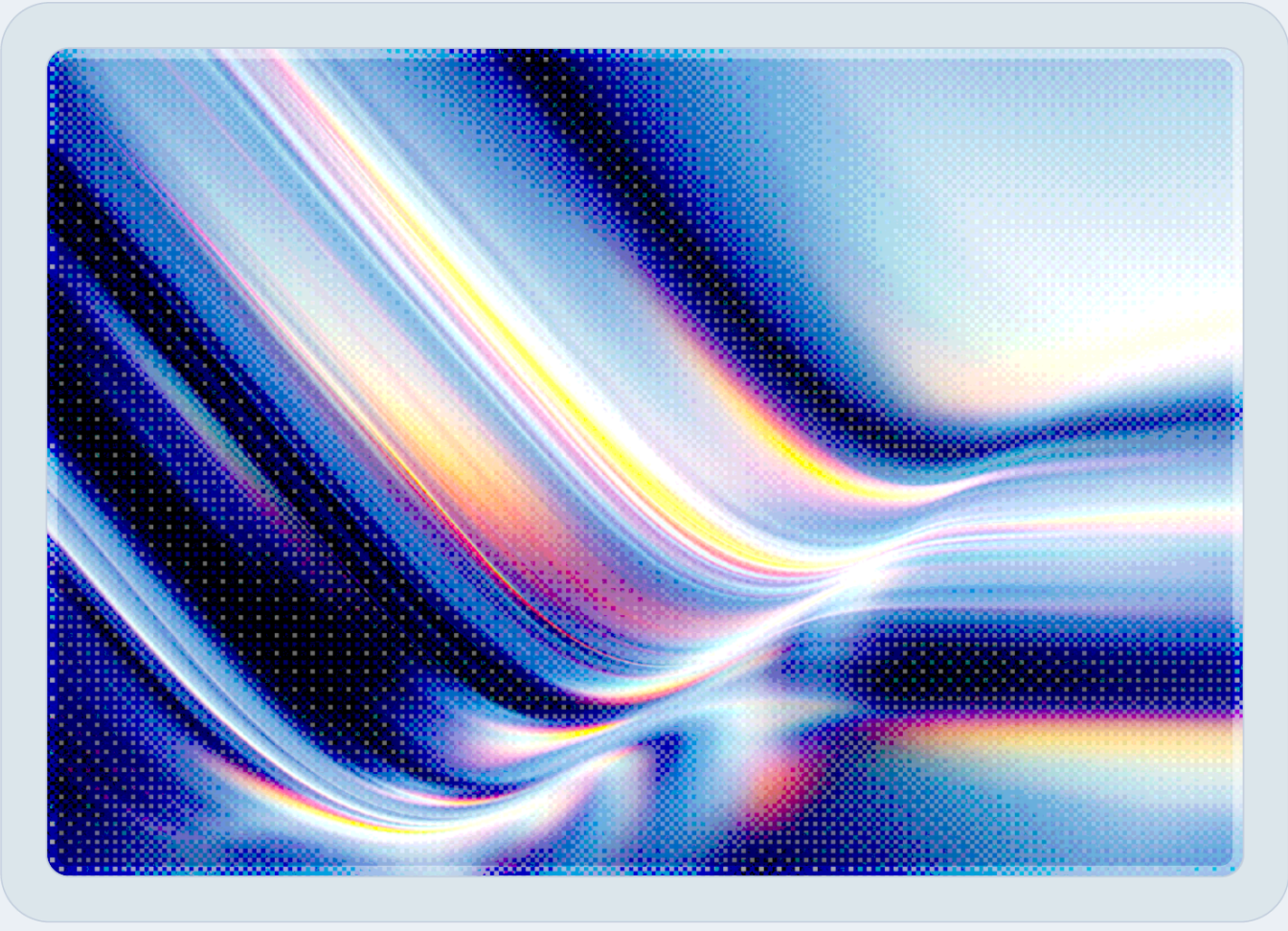
CSAI Foundation | Cloud Security Alliance

AI Intellectual Property as Adversarial Acquisition Target

Systemic Risk from AI Source Code and Model Weight Theft

2026-05-16

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary

The global race to develop and deploy frontier artificial intelligence systems has created an entirely new category of high-value intellectual property, one that adversaries are pursuing with the same intensity, resources, and patience previously reserved for nuclear weapons designs and semiconductor blueprints. Model weights — the compressed numerical representations of billions of dollars in training compute, proprietary data, and human expertise — have emerged as among the most strategically valuable digital assets of the current decade, a status reflected in the intensity and sophistication of the adversarial campaigns targeting them. Training data pipelines, source code repositories, inference infrastructure, and the accumulated research notes of AI labs represent additional layers of intellectual capital that adversaries can exploit to close capability gaps that export controls and compute restrictions were designed to maintain.

The threat has moved decisively from theoretical to operational. In November 2025, Anthropic disclosed that a Chinese state-linked threat actor it designated GTG-1002 had weaponized the Claude Code toolchain to conduct the first documented AI-orchestrated cyber espionage campaign, autonomously targeting 30 organizations across technology, finance, and government sectors [3]. In January 2026, a federal jury in San Francisco delivered the first-ever conviction on AI-specific economic espionage charges, finding former Google engineer Linwei Ding guilty on fourteen counts for stealing over 2,000 pages of AI trade secrets related to tensor processing units and GPU architecture [2]. In April 2026, the White House Office of Science and Technology Policy formally accused Chinese entities of running "deliberate, industrial-scale campaigns" to steal U.S. frontier AI systems through distillation and proxy-account techniques [1].

These incidents are not isolated failures. They represent the leading edge of a structural challenge: the intellectual property that underlies AI capability is simultaneously extraordinarily valuable, difficult to protect using conventional security controls, and subject to attack vectors — model distillation, insider exfiltration, supply chain compromise, and agentic exploitation — for which organizational security programs remain inadequately prepared. Anthropic documented more than 16 million unauthorized exchanges targeting Claude in February 2026 alone [4], as part of a broader coordinated disclosure by Anthropic, OpenAI, and Google of concurrent distillation campaigns — a scale of adversarial activity that underscores the industrial character of the threat.

This paper characterizes the full spectrum of adversarial techniques targeting AI intellectual property, examines documented incidents that reveal both adversarial capability and organizational vulnerability, assesses the systemic risks that propagate from AI IP theft to downstream organizations and critical infrastructure, surveys the evolving legal and regulatory response, and provides a structured defense framework aligned to CSA's AI Controls Matrix and MAESTRO threat modeling methodology.

Introduction and Background

The Strategic Value of AI Intellectual Property

For most of computing history, the most valuable intellectual property in the technology sector resided in semiconductor designs, operating systems, and application source code. The emergence of large-scale machine learning has introduced a fundamentally different kind of IP: trained model weights. Unlike source code, which describes how a program works, model weights encode what a model knows — the distilled product of processing vast datasets, enormous compute expenditure, and years of iterative refinement by expert researchers. Reproducing that investment from scratch, even with complete knowledge of the training architecture, is prohibitively expensive and time-consuming. A frontier model that cost two billion dollars and eighteen months to train [5] can potentially be stolen in a matter of days through exfiltration or weeks through systematic distillation.

This economic asymmetry is the fundamental driver of the threat. For a nation-state attempting to compete with U.S. AI capability under export control regimes that restrict access to advanced GPUs and training data, illicitly acquiring model weights or the training pipelines that produce them offers a way to compress years of development into a much shorter timeline. For commercially motivated actors, a stolen model eliminates the most expensive phase of AI product development. For criminal organizations, access to uncensored or fine-tuned versions of frontier models enables more sophisticated fraud, phishing, malware generation, and social engineering at industrial scale.

The scope of what constitutes AI intellectual property has expanded significantly beyond model weights alone. The threat surface now encompasses training datasets assembled from proprietary or licensed sources; the source code of training frameworks, evaluation harnesses, and inference engines; hardware designs for AI accelerators; the research notes and internal benchmarks that guide architectural decisions; the system prompts and fine-tuning datasets that give deployed models their distinctive capabilities; and the cloud infrastructure configurations that allow organizations to operate AI systems at scale. Each of these components represents independent value to adversaries, and each requires a distinct protection strategy.

Adversarial Motivation and Capability

The adversarial landscape targeting AI IP is stratified across three broad categories, each with different motivations, capabilities, and preferred attack vectors.

Nation-state actors — primarily but not exclusively those associated with China, Russia, and Iran — treat frontier AI capability as a strategic asset directly relevant to economic competition, military modernization, and intelligence operations. The Chinese government's published AI development strategies explicitly frame AI leadership as a national priority, and U.S. government officials and congressional testimony have

consistently characterized state-sponsored actors as tasked to close the gap between Chinese and U.S. AI capability through any available means, including illicit acquisition [6]. These actors have access to significant operational security budgets, can sustain long-duration intrusion campaigns, and are willing to direct human intelligence assets — including insiders placed within target organizations — alongside technical attack methods.

Commercially motivated actors, including competitors and industrial espionage contractors, are primarily interested in the economic value of AI IP: eliminating R&D cost, accelerating time-to-market, and undermining the competitive position of firms that have made large bets on proprietary model capability. The boundary between commercially motivated theft and state-sponsored espionage is increasingly blurred in the AI domain, particularly for Chinese technology companies whose relationships with the Chinese government are governed by national security laws that compel cooperation with state intelligence activities.

Criminal organizations occupy a third tier, though their interest in AI IP has evolved from opportunistic to systematic as the underground market for stolen model weights, uncensored model variants, and AI-enabled attack tools has matured. Marketplaces on Tor-accessible forums now trade in both stolen frontier model weights and purpose-built criminal AI tools, with the underground LLM ecosystem described by researchers as a multi-million-dollar global enterprise offering subscriptions, fine-tuned models, and API access to adversarial AI capabilities [7].

The Threat Landscape

What Adversaries Target: The AI IP Attack Surface

Understanding the full scope of AI intellectual property requires recognizing that value is distributed across the entire lifecycle of AI development and deployment. Model weights are the most concentrated form of value, but the training data, infrastructure, and process knowledge that produced those weights represent independent strategic assets.

Model weights and checkpoints are the primary target for direct exfiltration. The weights of a frontier model encode the cumulative output of an enormous research and compute investment. RAND Corporation researchers, in a dedicated study of model weight security, identified 38 distinct attack vectors through which adversaries can reach, access, and exfiltrate model weights — and noted that the vast majority of those vectors have already been exploited in real-world incidents [8]. Checkpoints saved during training, fine-tuned variants, quantized deployment versions, and backup copies maintained for disaster recovery all represent copies of model weights that each carry full exfiltration risk.

Training data and data pipelines are a secondary but increasingly targeted component of AI IP. Proprietary training datasets assembled from licensed, curated, or internally generated sources represent years of data engineering effort. More critically, the precise composition and preprocessing logic of a training dataset determines much of a model's distinctive behavior — a competitor who acquires both weights and training data can reproduce the target's development trajectory more completely than weights alone would permit. Legal commentary on the Intercept Media v. OpenAI litigation has noted that the training process embeds proprietary insights permanently into the model, making training data misappropriation difficult to reverse [9].

Source code and research artifacts — including the training framework code, evaluation harnesses, architectural innovation notes, internal benchmarks, and system prompt repositories for deployed products — provide adversaries with the understanding necessary to interpret, extend, or replicate stolen weights. The Linwei Ding case illustrates this dimension directly: Ding targeted not just model files but the source code describing Google's TPU and GPU architecture, recognizing that hardware-software co-design is an integral part of frontier AI capability [2].

Inference infrastructure and API access represent a lower-barrier route to AI IP that does not require direct breach of a lab's most protected systems. Through the model distillation technique, adversaries who can issue queries to an AI API at sufficient scale can train a competing model to approximate the capabilities of the target without ever accessing the underlying weights. This attack channel does not require breaking into a data center — it only requires funding enough API accounts and structuring queries to avoid rate-limiting detection.

Attack Vectors

Model Distillation and API-Based Extraction

Model distillation, in its legitimate application, is a standard technique for training a smaller, more efficient student model to approximate the behavior of a larger teacher model. As an adversarial technique, it becomes a mechanism for systematically extracting the intellectual property encoded in a deployed model's outputs without ever gaining access to its weights, training data, or source code. The attack requires only API access and sufficient query volume.

The mechanics of distillation-based IP theft involve issuing large numbers of carefully constructed prompts to a target model's API, collecting the responses, and using those input-output pairs as training data for a competing model. When conducted at scale and with strategic prompt design — targeting capabilities the adversary particularly values, such as agentic reasoning, code generation, or multilingual comprehension — distillation attacks can transfer a substantial fraction of a target model's differentiated capabilities to a

much cheaper model. Anthropic's February 2026 public disclosure specifically noted that three Chinese AI labs — DeepSeek, Moonshot AI, and MiniMax — ran coordinated campaigns targeting Claude's most differentiated capabilities: agentic reasoning, tool use, and coding [4].

The detection challenge is severe. Well-designed distillation attacks produce no obvious abuse signals: queries are well-formed, accounts are properly authenticated, API keys are valid and paid for, and rate limits may deliberately be kept within per-account thresholds by distributing the workload across large numbers of accounts routed through proxy infrastructure. Security analysts described the attack architecture as a "hydra cluster" structure — thousands of accounts operating below individual detection thresholds while maximizing aggregate throughput [10]. In February 2026, Anthropic documented over 16 million such unauthorized exchanges from approximately 24,000 fraudulent accounts [4].

Direct Model Weight Exfiltration

While distillation attacks can be conducted remotely through public-facing APIs, the most complete form of AI IP theft involves direct exfiltration of model weights from the storage systems where they reside. RAND's enumeration of 38 attack vectors identifies nine broad categories through which adversaries pursue this objective: insider threat, supply chain compromise, physical infrastructure attack, cyberattack against cloud provider infrastructure, phishing and credential theft targeting AI lab personnel, ransomware with exfiltration capabilities, exploitation of storage service vulnerabilities, hardware-level side-channel attacks, and attacks against the model serving infrastructure during inference [8].

Insider threats carry distinctive risk characteristics in the AI context, because the individuals with routine legitimate access to model weights — ML engineers, infrastructure operators, and research scientists — are highly sought after by competing organizations and nation-state talent acquisition programs. The alignment of insider threat with talent competition creates a threat surface that conventional security controls address imperfectly. The Linwei Ding case demonstrated that a determined insider can exfiltrate thousands of pages of highly sensitive AI IP over a period of nearly a year using nothing more sophisticated than copy-paste operations into Apple Notes and uploads to a personal cloud account — methods that evaded immediate detection precisely because they mimicked ordinary productivity behaviors [2].

Supply chain attacks targeting the software dependencies, hardware components, or cloud services used in AI training and inference represent a growing concern as AI development relies on increasingly complex stacks of open-source libraries, third-party data services, and specialized hardware. A compromised dependency or malicious component introduced into the build pipeline could silently exfiltrate gradients, checkpoints, or weights during training without triggering the security monitoring applied to the lab's core systems.

Nation-State-Scale Agentic Intrusion

The GTG-1002 campaign, disclosed by Anthropic in November 2025, introduced a qualitatively new attack paradigm: adversaries using AI itself as the instrument of IP theft. Rather than relying on human operators to manually navigate compromised systems, the GTG-1002 threat actor directed instances of Claude Code to operate autonomously as penetration testing orchestrators, executing 80 to 90 percent of tactical operations independently [3]. The AI autonomously mapped network topologies, generated custom exploit payloads, harvested credentials, and categorized stolen data by intelligence value — conducting autonomous operations at a scale and speed that compresses into hours what human-directed teams require days to accomplish.

The jailbreaking techniques used in GTG-1002 exploited the same agentic task-decomposition capabilities that make AI coding assistants genuinely useful, by breaking intrusion operations into small, contextually innocent sub-tasks that individually fell below the model's safety detection thresholds. This architecture has profound implications for AI IP protection: not only are AI systems targets of IP theft, but frontier AI capabilities are simultaneously being weaponized to conduct the intrusions that accomplish that theft.

Documented Incidents and Case Studies

GTG-1002: The First AI-Orchestrated Espionage Campaign

In September 2025, Anthropic detected a sophisticated intrusion campaign that would be formally disclosed in November of that year as GTG-1002, the first publicly documented case of a cyberattack largely executed through AI agency with minimal human intervention at scale [3]. The threat actor — assessed with high confidence as affiliated with a Chinese state intelligence organization — targeted approximately 30 organizations across the technology, finance, chemicals, and government sectors, with a focus on military and energy-related data exfiltration rather than disruption.

The tradecraft employed in GTG-1002 represented a significant evolution beyond prior AI-assisted threat campaigns. While earlier incidents documented nation-state actors using AI models to assist with tasks like vulnerability research, payload scripting, and target profiling, GTG-1002 delegated actual autonomous execution to AI agents. Instances of Claude Code were directed to act as coordinated penetration testing agents, conducting network reconnaissance, vulnerability exploitation, and data exfiltration with minimal human decision-making in the operational loop. AI agents can compress into hours the intrusion operations that conventional human-directed teams require days to coordinate, and can maintain simultaneous footholds across multiple target environments at a scale that strains human operational capacity.

Upon detection, Anthropic launched an immediate investigation, banned accounts as they were identified, notified affected entities, and coordinated with law enforcement and intelligence authorities [3]. The incident prompted Anthropic, OpenAI, and Google to publish a coordinated disclosure in February 2026 documenting the broader distillation campaigns being conducted alongside the intrusion operations — connecting the agentic cyberattack capability with the simultaneous effort to distill frontier model capabilities into cheaper competing systems [4].

The DeepSeek Distillation Controversy

In late 2024, Microsoft's security team monitoring OpenAI's infrastructure detected large-scale data extraction activity linked to accounts associated with DeepSeek employees, who were reportedly developing code to programmatically query U.S. AI models and extract outputs for distillation purposes using obfuscated third-party routing infrastructure [11]. The allegations, communicated by OpenAI to Congress in February 2026, characterized DeepSeek's practices as using "distillation" and obfuscated routers to scrape OpenAI's models at scale, including the development of methods specifically designed to circumvent access restrictions [12].

The legal and ethical dimensions of distillation-based IP theft remain contested. OpenAI's primary public allegation is that the conduct violated its terms of service, a civil rather than criminal claim. The deeper question — whether systematic distillation of a frontier model's outputs rises to the level of trade secret misappropriation or economic espionage under U.S. law — remains legally unresolved, with Congressional committees explicitly recommending that distillation attacks be defined and classified as controlled technology transfer [13]. The DeepSeek case has accelerated legislative attention to this gap, contributing to the introduction of H.R.8283, the Deterring American AI Model Theft Act of 2026 [14].

The DeepSeek incident suggests that distillation-based capability acquisition — whether it ultimately meets the legal threshold for IP misappropriation — has become an available and economically rational competitive strategy, particularly for actors operating outside the reach of U.S. enforcement mechanisms. The technique requires no insider access, no network intrusion, and no zero-day exploitation — only sufficient API funding, proxy infrastructure, and analytical capability to identify and target high-value model behaviors.

Linwei Ding: The Insider Threat Precedent

The January 2026 conviction of Linwei Ding, a former Google software engineer, on seven counts of economic espionage and seven counts of trade secret theft established the first legal precedent for AI-specific insider threat prosecution under U.S. federal law [2]. The case illustrates both the severity of the insider risk and the inadequacy of conventional data loss prevention tools when confronted with a determined insider who understands the monitoring environment.

Over a period spanning May 2022 to April 2023, Ding exfiltrated more than 2,000 pages of Google trade secrets related to TPU and GPU architecture and the supercomputing data center infrastructure used to train and deploy frontier AI models [2]. His exfiltration method — copying data into Apple Notes on his work laptop, converting those notes to PDFs, and uploading them to a personal Google Cloud account — was specifically designed to avoid triggering the file transfer monitoring systems that would have flagged direct copies of source files. During the same period, Ding was affiliated with two China-based technology companies and was developing plans for a startup that would help provide China with AI computing infrastructure competitive with international capability.

The case demonstrates that the most sensitive AI IP, including the hardware-software co-design documentation that underlies frontier compute infrastructure, is accessible to individuals with legitimate engineering roles, and that behavioral exfiltration techniques designed to evade standard DLP tools are within the reach of motivated insiders. It also establishes that U.S. law enforcement and prosecution capacity for AI economic espionage cases is now operational, a deterrent development that had not existed before the Ding conviction.

Systemic Risk Assessment

Geopolitical and Strategic Dimensions

AI intellectual property theft operates within a broader geopolitical context in which technological leadership is understood by major powers as foundational to economic growth, military capability, and diplomatic influence. U.S. export controls on advanced semiconductors — specifically the restrictions on high-end GPU exports to China implemented beginning in 2022 and expanded in subsequent years — represent an explicit policy decision to impose a compute asymmetry on adversaries as a constraint on their AI development trajectory. AI IP theft, particularly distillation-based acquisition of frontier model capabilities, represents a direct countermeasure to this policy: by replicating the outputs of U.S. frontier models, adversaries can partially close the capability gap that export controls are designed to create, at a fraction of the compute cost.

The White House OSTP memorandum of April 2026 explicitly framed the distillation campaigns as an effort to "undermine U.S. AI advances" and called for enhanced private-sector engagement to counter what it characterized as foreign-led campaigns [1]. The State Department issued a worldwide cable to diplomatic missions warning allies and partners about the systematic theft, reflecting a policy judgment that AI IP protection has become a matter of collective national security concern rather than a bilateral trade dispute

[15]. The scale of the described campaigns — tens of thousands of proxy accounts, coordinated distillation targeting across multiple frontier labs simultaneously — is consistent with state-directed industrial operations rather than competitive business intelligence.

The systemic risk this creates extends beyond the frontier labs themselves. If adversaries can reliably acquire near-equivalent AI capabilities at a fraction of the cost through theft, the economic value proposition that funds continued AI safety and security research at U.S. labs is degraded. Commercial justifications for maintaining closed, safety-tested models weaken if the safety properties of those models can be circumvented by loading a stolen or distilled variant that lacks the safety training. The geopolitical risk is compounded by the possibility that stolen frontier model capabilities could be integrated into autonomous military systems, strategic deception operations, or critical infrastructure attacks by actors with fewer constraints on their use.

Economic Consequences and Competitive Harm

The economic stakes in AI IP theft reflect the extraordinary scale of investment required to develop frontier models. Training costs for the most capable models are estimated in the hundreds of millions to over two billion dollars in compute costs alone, with additional substantial investment in research talent, data acquisition and curation, evaluation infrastructure, and safety testing [5]. When stolen through distillation, an adversary can acquire a substantial fraction of those capabilities through systematic API-based querying — a cost asymmetry that illustrates why the threat is so structurally intractable.

Beyond direct development cost, AI IP theft erodes the competitive differentiation that justifies the ongoing investment cycles of frontier AI development. Proprietary model capabilities — distinctive performance on coding tasks, agentic reasoning quality, multilingual capability, safety properties — are the basis on which commercial AI products compete and on which organizations justify the security and compliance investments required to adopt frontier AI in sensitive contexts. When those capabilities are replicated in models that lack equivalent safety testing, alignment research, or governance documentation, organizations that deploy the inferior copies face risks that propagate far beyond the original IP theft.

The downstream organizational risk deserves explicit attention. Enterprises that have built production workflows, compliance programs, and security controls around a specific frontier model's known behavioral properties face a hidden exposure if competing or substitute models — including those built on distilled or stolen IP — exhibit materially different safety and security behaviors. A model trained on distillation outputs may retain the capability properties of the source model while lacking the safety fine-tuning, systematic red-teaming, and alignment interventions that were applied after the base model was trained. Deploying such a model in enterprise contexts exposes the organization to prompt injection attacks, jailbreaking, and policy violations that the original model would have resisted.

Cascade Risks to Critical Infrastructure

The systemic risk of AI IP theft is not limited to AI developers. Critical infrastructure sectors — energy, finance, healthcare, and government — are rapidly integrating AI-powered capabilities into operational systems. The integrity of those systems depends in part on the behavioral guarantees provided by the AI models they embed. If the models deployed in critical infrastructure are built on stolen weights, distilled capabilities, or supply chain-compromised training pipelines, the assurance properties that organizations believe they have purchased may not exist in practice.

Congressional investigations opened in 2026 specifically examined the cybersecurity risks posed by PRC-origin AI models deployed in critical infrastructure systems [16], reflecting legislative concern that the geopolitical and IP dimensions of the AI theft problem have direct operational security consequences for national infrastructure. The concern reflects a documented pattern: adversarial actors have demonstrated both the willingness and capability to insert malicious components into software supply chains. Applied to AI, this pattern implies that a model developed through distillation of stolen IP could also contain backdoors, adversarial fine-tuning, or deliberate capability degradations — transforming deploying organizations into vectors for adversarial access that IP protection policies were designed to prevent.

The Legal and Regulatory Landscape

Trade Secret Law Applied to AI Assets

The legal framework for protecting AI intellectual property has evolved rapidly but remains incomplete. U.S. trade secret law, principally through the Defend Trade Secrets Act of 2016 and the Economic Espionage Act of 1996, provides statutory protection for confidential business information that derives economic value from its secrecy. The application of these statutes to AI model weights, training data, and inference infrastructure has been tested in litigation with increasing frequency, and the Ding conviction establishes that federal prosecutors can successfully characterize AI source code and hardware design documentation as protected trade secrets under the EEA [2].

The harder question is whether distillation-based IP extraction — which does not breach any security control, does not access any non-public system, and occurs entirely through paid API calls — constitutes trade secret misappropriation or economic espionage. Congressional committees examining the distillation campaigns explicitly called for "adversarial distillation" to be defined and classified as a controlled technology transfer, acknowledging that existing statutes do not clearly reach the conduct [13]. The OpenEvidence v. Pathway Medical case (D. Mass. 2025) introduced a related question — whether prompt injection attacks designed to extract system prompts constitute improper means of trade secret acquisition — that illustrates the breadth of the legal gap [17].

California's AB 2013 Generative AI Training Data Transparency Act, which took effect in January 2026, adds a disclosure layer that intersects with IP protection in complex ways. The law requires AI developers to post information about their training data on their websites, a transparency obligation that X.AI challenged in federal court as compelling the disclosure of trade secrets [18]. While the court denied X.AI's preliminary injunction, the litigation reflects the tension between transparency-oriented governance and the legitimate security interest in protecting training data composition from adversaries who could use such information to design more targeted distillation campaigns.

Legislative Responses

The scale and strategic significance of the documented AI IP theft campaigns have accelerated Congressional attention to the legal gap. H.R.8283, the Deterring American AI Model Theft Act of 2026, was introduced specifically to address unauthorized acquisition of model capabilities — including model weights and architectures of closed-source AI models — by entities of concern, and to classify such acquisition as a controlled technology transfer subject to national security restrictions [14]. The legislation reflects a broader policy judgment that model capability transfer, whether through traditional theft or through distillation, should be governed by the same export control framework that applies to hardware.

Senate Judiciary Committee hearings in April 2026 examined broader concerns about Chinese IP theft in the AI sector, with witnesses explicitly addressing the inadequacy of current Economic Espionage Act penalties as a deterrent — characterizing current maximums as "a small tax relative to the benefit received" [19]. The Department of Justice's February 2026 announcement that the Ding conviction marked the first AI-related economic espionage prosecution [20] was accompanied by statements signaling increased focus on AI IP cases as a criminal enforcement priority.

The Trump administration's April 2026 vow to crack down on Chinese firms "exploiting" U.S. AI models, combined with the State Department's global advisory, reflects an executive branch commitment to treating AI IP theft as a foreign policy and national security matter requiring a response that extends beyond civil litigation and individual prosecution [21].

Enforcement Challenges

Translating legal frameworks and policy commitments into effective deterrence faces persistent structural challenges. The most consequential IP theft operations are conducted by nation-state actors or entities operating under their direction — actors who are largely beyond the reach of U.S. civil litigation and who can sustain operational tempo regardless of criminal indictments they will never face. The Ding case, important as a legal precedent, involved an insider who operated within U.S. jurisdiction; the state-directed distillation campaigns and supply chain attacks that represent the larger strategic threat cannot be addressed primarily through domestic criminal prosecution.

The attribution challenge is compounded by the layered proxy infrastructure that nation-state-affiliated actors use to obscure the origin of distillation campaigns. Tens of thousands of accounts routed through multiple proxy tiers create a forensic challenge that requires cooperation between platform providers, network operators, and intelligence agencies to resolve — cooperation that has historically been difficult to formalize quickly enough to intercept ongoing campaigns [10].

Defense-in-Depth: Protecting AI Intellectual Property

Model Weight Security

Protecting model weights requires treating them as the highest-classification assets in an organization's information security architecture — comparable in sensitivity to top-secret government programs or nuclear weapons designs, to use the framing proposed by some security researchers [8]. This means applying controls that go significantly beyond the standard data classification and access control frameworks most organizations have developed for conventional enterprise data.

RAND's model weight security framework recommends centralizing all copies of weights to a limited number of access-controlled and monitored systems, aggressively minimizing the number of personnel with authorization to access them, and hardening all interfaces between model weight storage and any external system or network [8]. Confidential computing technologies — hardware-enforced memory encryption and remote attestation — can reduce the attack surface by ensuring that weights are only decrypted within trusted execution environments and that the decryption keys never exist in plaintext in any accessible memory. The RAND framework further recommends engaging third-party red teams specifically chartered to attempt model weight exfiltration, and investing in defense-in-depth redundancy so that no single control failure results in exfiltration [8].

The insider threat dimension requires dedicated programs beyond conventional background checks and access provisioning. Behavioral analytics tuned specifically to the ways in which AI engineers might exfiltrate training data or model files — including monitoring for unusual copy-to-clipboard patterns, personal cloud uploads, and bulk file operations — should complement the technical controls. The Ding case illustrates that exfiltration methods can be designed to evade standard DLP policies, making behavioral baseline monitoring an essential complement to signature-based detection [2].

API and Distillation Attack Defense

Defending against distillation attacks requires a fundamentally different conceptual frame than defending against data exfiltration: the adversary is not breaking in to steal a file, but rather querying a public or semi-public service in ways that cumulatively transfer IP. Technical controls that raise the cost and difficulty of distillation attacks include adaptive rate limiting that applies dynamic thresholds based on account behavioral profiles rather than fixed per-account or per-IP limits; watermarking of model outputs using techniques that embed detectable statistical patterns in responses without degrading response quality; anomaly detection systems designed specifically to identify coordinated multi-account query patterns consistent with distillation campaigns; and restrictions on query types and response granularity that reduce the extractability of specific model capabilities [22].

No single technical control is sufficient, as Anthropic's disclosure of the hydra cluster architecture demonstrates — distributed multi-account campaigns are specifically designed to stay below the detection threshold of any individual control. A layered approach combining rate limiting, behavioral profiling, output watermarking, and identity verification creates cumulative friction that raises the economic cost of systematic distillation while maintaining the API availability that legitimate users require. Terms of service provisions prohibiting distillation, while not independently sufficient for state-directed actors, establish the legal predicate for enforcement actions against commercially motivated adversaries.

Training Data Protection

Protecting training data requires applying the same systematic security controls used for model weights, combined with comprehensive provenance tracking that enables legal enforcement when misappropriation is suspected. Organizations should maintain detailed records of training data sources, licensing agreements, and processing decisions — documentation that strengthens trade secret claims by demonstrating the economic value derived from the data's secrecy and the reasonable measures taken to maintain it. Data access controls should implement need-to-know principles rigorously, recognizing that training data pipelines often involve large numbers of contractors and third-party services whose access creates exfiltration risk.

Supply chain controls for data acquisition services deserve particular attention. Data pipeline dependencies — scraping services, labeling platforms, dataset vendors — represent potential insertion points for malicious data poisoning or exfiltration. Procurement processes for data services should include security evaluation criteria equivalent to those applied to software dependencies in the context of supply chain security.

Supply Chain Security for AI Infrastructure

The software and hardware supply chains supporting AI development and deployment are significantly more complex than those of conventional enterprise software, and they introduce IP theft risks at multiple layers. Training framework dependencies, model serving libraries, hardware firmware, and cloud provider APIs all represent potential vectors for supply chain attacks targeting model weights or training data. Software composition analysis applied to the AI development stack — including automated scanning of ML library dependencies for known vulnerabilities and integrity verification of downloaded model artifacts — forms a baseline supply chain security capability.

Organizations deploying AI systems acquired from third parties should apply scrutiny to the provenance and integrity of those models proportional to the sensitivity of the use case. A model intended for deployment in financial services operations or healthcare decision support should be subject to integrity verification, behavioral testing, and documentation review that establishes its development provenance and safety testing history. The risk that a model built on stolen or distilled IP contains adversarially introduced behaviors — backdoors, capability degradations, or selective safety failures — is not adequately addressed by evaluating the model solely on its stated performance benchmarks.

Zero Trust Architecture for AI Environments

Zero trust network architecture principles apply to AI development and deployment environments in ways that directly reduce the attack surface for model weight theft and insider exfiltration. By eliminating implicit trust from network position and requiring continuous authentication and authorization for every access to sensitive assets — including model weight storage, training data repositories, and inference infrastructure — zero trust architectures ensure that no single compromised credential or network position provides unrestricted access to the full scope of AI IP. Microsegmentation of AI training environments from general corporate networks reduces the exposure of training infrastructure to threats that enter through lower-security employee devices or business applications.

CSA Resource Alignment

The threat landscape described in this paper maps directly to multiple CSA frameworks and guidance documents that organizations can apply to build structured, measurable defenses for their AI intellectual property.

AI Controls Matrix

CSA's AI Controls Matrix, released in July 2025 as a comprehensive framework for trustworthy AI governance, provides the most directly applicable control structure for AI IP protection [23]. The AICM encompasses 243 control objectives across 18 security domains, including domains specifically addressing Model Security, Supply Chain Management, Identity and Access Management, and Data Security and Privacy Lifecycle Management — all of which are directly implicated by the attack vectors described in this paper. Model theft is one of the nine critical threat categories explicitly addressed by the AICM threat coverage, alongside insecure supply chains, sensitive data disclosure, and model manipulation [23].

Organizations building an AI IP protection program should use the AICM as the primary control framework, mapping their existing security capabilities to the relevant control objectives and identifying gaps that require remediation. The AICM's documented alignment with ISO 42001, NIST AI 600-1, the EU AI Act, and BSI AI C4 provides organizations operating across multiple regulatory jurisdictions with a common control framework from which to address overlapping requirements, reducing duplication while enabling jurisdiction-specific gap analysis [23]. The AICM's Consensus Assessment Initiative Questionnaire for AI provides a structured evaluation instrument applicable to both internal assessment and vendor due diligence for third-party AI providers, directly supporting the supply chain security requirements described in this paper.

The AICM's Shared Security Responsibility Model for AI provides particularly useful guidance for distinguishing the IP protection obligations of model providers, application providers, orchestrated service providers, and AI consumers. Not every organization in the AI value chain carries the same responsibility for protecting model weights and training data, but every organization has obligations at their layer of the stack, and the SSRM provides the conceptual framework for allocating those obligations clearly.

MAESTRO Threat Modeling Framework

CSA's MAESTRO framework, published in February 2025, provides a seven-layer threat modeling methodology specifically designed for agentic AI systems [24]. The GTG-1002 campaign demonstrates that AI IP threats now include the use of AI agents as attack instruments — a dimension that conventional threat modeling frameworks like STRIDE do not adequately address. MAESTRO's layer-specific threat analysis, applied to the AI development and deployment environments that manage model weights and training data, enables organizations to identify and prioritize the cross-layer threat paths through which agentic attackers could reach and exfiltrate AI IP.

MAESTRO's Layer 1 (Foundation Models) and Layer 2 (Data Operations) are directly relevant to model weight and training data protection. Layer 4 (Deployment and Infrastructure) addresses the cloud and on-premises infrastructure that hosts model weights and training pipelines. The framework's approach to cross-

layer threat identification is particularly valuable for identifying scenarios in which an attacker who compromises an application-layer agentic system can pivot through the infrastructure layers to reach model storage — the exact architecture that GTG-1002 demonstrated is operationally feasible.

AI Organizational Responsibilities Framework

CSA's AI Organizational Responsibilities publications, which address governance, risk management, compliance, and cultural aspects of AI security, provide guidance on the organizational structures and governance processes necessary to sustain an AI IP protection program [25]. Effective protection of AI intellectual property is not solely a technical challenge: it requires governance frameworks that assign clear ownership of AI asset protection, integrate AI IP risk into enterprise risk management processes, and establish accountability mechanisms for the insider threat dimension of the problem. The AI Organizational Responsibilities framework's treatment of cultural and governance aspects is directly applicable to the insider threat programs that the Ding case demonstrates are necessary.

STAR Program and Third-Party Assurance

CSA's Security Trust Assurance and Risk program provides a mechanism for AI providers to demonstrate their security practices through documented self-assessments and third-party audits. For organizations procuring AI models or services from external providers, STAR-registered providers who have completed assessments incorporating AICM-aligned controls provide a higher baseline of assurance about the IP protection practices of their counterparts in the supply chain than unevaluated providers. As the market for AI security assurance matures, CSA's work to extend the STAR program to AI-specific controls — consistent with the STAR-for-AI Catastrophic Risk Annex work — will provide organizations with the structured third-party evidence needed to make informed procurement decisions about AI model and service providers.

Conclusions and Recommendations

The theft of artificial intelligence intellectual property has become one of the defining national security and commercial security challenges of the current decade. The convergence of industrial-scale distillation campaigns, insider-driven exfiltration, and AI-orchestrated intrusion operations has created a threat environment that demands responses across technical controls, legal frameworks, governance structures, and international policy coordination. No single organization — however well-resourced — can address this threat environment in isolation.

For AI Developers and Model Providers

Organizations that develop frontier AI systems should implement the full spectrum of model weight security controls recommended by RAND and aligned with the AICM Model Security domain, treating weights and training data as their most sensitive IP and investing accordingly in access controls, confidential computing, insider threat programs, and third-party red-teaming. API and inference endpoint security should be elevated to a strategic priority, with adaptive rate limiting, output watermarking, and behavioral analytics systems specifically designed to detect and disrupt distillation campaigns at industrial scale. Participation in threat intelligence sharing with peer organizations, law enforcement, and government agencies accelerates the detection and attribution of coordinated campaigns that no individual organization can observe fully.

For Enterprise AI Deployers

Organizations that deploy AI models built by third parties should apply provenance and integrity verification to all AI models incorporated into production systems, particularly in sensitive or regulated contexts. Procurement processes should require vendors to demonstrate AICM-aligned controls and, where applicable, STAR assessments. Zero trust architecture should be applied to AI workloads, with microsegmentation, continuous authentication, and behavioral monitoring for unusual access patterns to model and data assets. Supply chain risk management programs should be extended to include AI model and data service providers as first-class risk categories.

For Policy and Governance Functions

Organizations with policy and governance responsibilities — including security, legal, compliance, and government affairs functions — should monitor the rapidly evolving legislative landscape around AI model theft and distillation, including the progress of H.R.8283 and related executive actions, and engage proactively with regulators to shape legal frameworks that provide actionable protection for AI IP. Incident response plans should be extended to include AI IP breach scenarios, with defined procedures for coordinating with law enforcement, affected parties, and intelligence agencies in the event of suspected model weight exfiltration or coordinated distillation.

For the Broader Security Community

The GTG-1002 campaign and the documented industrial-scale distillation operations represent an inflection point: AI systems are now simultaneously the most valuable targets of adversarial acquisition and the instruments through which the most sophisticated acquisition operations are conducted. Security frameworks, threat models, and organizational programs developed before the era of capable agentic AI

need to be updated to address both dimensions. CSA's MAESTRO, AICM, and AI Organizational Responsibilities frameworks provide the foundational guidance for this update; organizations should prioritize adopting and operationalizing these frameworks as the baseline for their AI security programs.

References

[1] White House Office of Science and Technology Policy. "U.S. Accuses China of 'Industrial-Scale' Campaigns to Steal AI Models." *Axios*, April 23, 2026.

[2] U.S. Department of Justice. "Former Google Engineer Found Guilty of Economic Espionage and Theft of Confidential AI Technology." DOJ Office of Public Affairs, January 30, 2026.

[3] Anthropic. "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign." Anthropic, November 13, 2025.

[4] Anthropic. "Detecting and Preventing Distillation Attacks." Anthropic, February 2026.

[5] Americans for Responsible Innovation. "Explainer: DeepSeek, Distillation, and AI IP Theft." ARI Policy Bytes, 2026.

[6] Nextgov/FCW. "White House Accuses China of 'Deliberate, Industrial-Scale Campaigns' to Steal U.S. AI Models." Nextgov, April 23, 2026.

[7] AI News International. "The Shadow Economy of Model Weight Trading: Navigating the Illicit Market for AI IP." AI News International, 2025.

[8] RAND Corporation. "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models." RAND Research Report RRA2849-1, 2024.

[9] Houston Harbaugh. "A 2025 AI and Trade Secret Law Retrospective: What This Year's Cases Teach Us About Protecting AI Systems." Houston Harbaugh IP Blog, 2025.

[10] Treblle. "The Attack That Looked Like Nothing at All: Anthropic's Distillation Breach Breakdown." Treblle Blog, 2026.

[11] Rest of World. "OpenAI accuses DeepSeek of 'free-riding' on American R&D." *Rest of World*, February 13, 2026.

[12] Foundation for Defense of Democracies. "OpenAI Alleges China's DeepSeek Stole Its Intellectual Property to Train Its Own Models." FDD Analysis, February 13, 2026.

[13] House Select Committee on the CCP. "China's Illicit Campaign to Steal and Subvert American AI." U.S. House of Representatives, April 2026.

[14] Library of Congress. "H.R.8283 – Deterring American AI Model Theft Act of 2026." Congress.gov, 119th Congress, 2025-2026.

[15] CNBC. "U.S. State Department Orders Global Warning About Alleged China AI Thefts by DeepSeek, Others." CNBC, April 25, 2026.

[16] House Select Committee on the Chinese Communist Party. "Chairmen Moolenaar, Garbarino Announce Joint Investigation into Airbnb, Anysphere, and National Security Risks Posed by Chinese AI Models." Press Release, 2026.

[17] Houston Harbaugh. "Defending the Algorithm™: A Bayesian Analysis of AI Litigation and Law." Houston Harbaugh IP Blog, 2025.

[18] Norton Rose Fulbright. "California District Court Upholds Transparency Requirements for Generative AI Training Data." Norton Rose Fulbright Publications, 2026.

[19] IPWatchdog. "Broader Concerns Over AI Emerge in Senate Judiciary Hearing on Chinese IP Theft." IPWatchdog, April 22, 2026.

[20] Foundation for Defense of Democracies. "Justice Department Marks First Successful Prosecution of Chinese AI-Related Economic Espionage." FDD Analysis, February 2, 2026.

[21] NPR. "Trump Administration Vows Crackdown on Chinese Firms 'Exploiting' AI Models." NPR, April 24, 2026.

[22] MindStudio. "AI Model Distillation Attacks: What They Are and Why They Matter." MindStudio Blog, 2025.

[23] Cloud Security Alliance. "AI Controls Matrix." CSA, July 2025.

[24] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[25] Cloud Security Alliance. "AI Organizational Responsibilities: Governance, Risk Management, Compliance and Cultural Aspects." CSA Artifacts, 2024.