

Industrial-Scale AI Distillation

Engineering U.S. Frontier AI Systems to Constrain Capability Extraction

Dr. Ron Ross

CEO, RONROSSECURE, LLC
ISTS FELLOW, DARTMOUTH COLLEGE

Abstract

A September 8, 2026 joint cybersecurity advisory from the National Security Agency (NSA), Cybersecurity and Infrastructure Security Agency (CISA), and Federal Bureau of Investigation (FBI) describes several industrial-scale campaigns by China-based artificial intelligence companies to extract proprietary capabilities from U.S. frontier models. The reported activity uses millions of requests, distributed accounts, cloud providers, third-party aggregators, proxy transfer stations, automated failover, and capability-focused prompting to transform ordinary inference access into a strategic collection mechanism. This article responds from a systems security engineering perspective. It argues that malicious distillation is not merely account abuse, fraud, or excessive API consumption, but an unauthorized transfer of computational capability assembled from individually plausible interactions. Effective defense requires more than stronger authentication and static rate limits. Providers must define unacceptable losses, establish security domains, mediate every consequential interaction, limit persistence and sharing, detect coordinated anomalies across organizational boundaries, adapt responses, preserve mission utility, and produce evidence that the architecture constrains extraction under realistic adversary behavior. NIST Special Publication 800-160, Volumes 1 and 2, together with Engineering for Compromise, provide the principles for building that defense.

1. The Advisory Changes the Engineering Problem

A September 8, 2026, joint cybersecurity advisory from the National Security Agency (NSA), Cybersecurity and Infrastructure Security Agency (CISA), and Federal Bureau of Investigation (FBI) reports that China-based AI companies have systematically extracted proprietary functionality and capabilities from U.S. frontier AI models at industrial scale [1]. The campaigns allegedly span millions of exchanges and billions of tokens and target reasoning behavior, specialized optimizations, agentic functions, domain expertise, and other capabilities that would otherwise require substantial research, development, data, time, and compute to reproduce. The activity is strategically significant because it can compress a competitor's development cycle while externalizing much of the cost to the U.S. companies whose models are queried.

The defensive challenge is not a single stolen checkpoint or compromised administrator. The provider intentionally exposes a powerful inference service; the adversary uses that surface repeatedly, distributes requests across multiple identities and pathways, blends malicious activity with legitimate traffic, evaluates defensive changes, and recombines outputs into a new training asset. Each request can appear permissible in isolation while the campaign as a whole produces an unauthorized transfer of capability.

“The security problem is not one prohibited query. It is the prohibited capability assembled from millions of permitted-looking interactions.”

This distinction changes the *unit of analysis*: a provider can authenticate every account, authorize every call, and encrypt every connection, yet still lose strategically valuable capability. Traditional controls answer

whether a request may reach the service. Industrial-scale distillation requires the system to determine whether the cumulative pattern, purpose, and effect of requests remain within authorized use. The *security boundary* must therefore extend across accounts, sessions, providers, aggregators, time, model versions, and organizations.

The advisory recommends comprehensive detection and mitigation, targeted response changes, and cross-organization intelligence sharing [1]. Those actions are *necessary* but not *sufficient*. Systems security engineering adds the architectural question: what properties must the AI service possess so those actions remain effective when individual accounts, pathways, and detection mechanisms fail?

2. Define the Loss Before Selecting the Control

NIST SP 800-160, Volume 1, defines security in relation to freedom from conditions that can cause *asset loss* with unacceptable consequences [2]. For frontier AI providers, the protected asset extends beyond source code, weights, or training data to the behavior embodied in the model—reasoning strategies, specialized methods, safety mechanisms, tool-use patterns, domain adaptations, and the strategic advantage their combination creates.

A loss-driven chain for malicious distillation can be expressed as:

Strategic AI Capability → Unauthorized Capability Transfer → Coordinated Extraction Condition → Security Constraint → Architecture → Mechanism → Evidence.

The unacceptable loss is not high token consumption, but a competitor acquiring enough high-value outputs to reproduce restricted functionality, compress development time, or erode U.S. technological advantage. The loss-enabling condition is a campaign's ability to obtain and combine sufficiently informative responses without being detected, constrained, degraded, or attributed in time. Security requirements should be derived from that condition.

A representative top-level constraint: **the inference system shall prevent any unauthorized entity or coordinated set of entities from accumulating outputs, reasoning artifacts, behavioral signals, or privileges sufficient to reproduce protected capabilities.** Any such accumulation must remain below the stakeholder-accepted *loss margin*. This security requirement is more demanding than “block abusive accounts.” It treats distributed extraction as a system-level outcome and forces stakeholders and architects to address aggregation, correlation, cumulative exposure, and response.

Protection must remain commensurate with consequence. Not every output has the same extraction value, and not every model or capability creates the same strategic exposure. Providers should identify crown-jewel capabilities, estimate how query sequences reveal them, define acceptable capability leakage, and allocate stronger mediation and assurance where the consequences of loss are greatest.

3. Mediated Access Must Govern Cumulative Effect

NIST SP 800-160 states the *Mediated Access* security design principle directly: all access to and operations on system elements are mediated [2]. Applied to frontier AI, mediation must govern more than account admission: it must control the model, operation, parameters, context length, output detail, reasoning exposure, tool access, rate, duration, sequence, and cumulative information value of each request [6].

The advisory describes native APIs, remote cloud providers, third-party aggregators, transfer stations, and premium subscriptions that obscure provenance or bypass geographic restrictions [1]. An access mediation architecture that evaluates only the current API key is incomplete. The policy decision requires richer

context: account history, payment and ownership relationships, device and network attributes, request similarity, timing correlations, model-switching behavior, shared infrastructure, downstream aggregation, and intelligence from other providers.

“Authentication can establish who opened an account. It cannot determine what a distributed campaign is collectively learning.”

Complete mediation therefore has both temporal and ecosystem dimensions. Temporally and across the ecosystem, the provider must correlate decisions across sessions and accounts, reevaluate authorization as evidence accumulates, and impose policy while the protected effect can still be controlled. Relevant actions include reducing reasoning detail, withholding high-value intermediate artifacts, changing model routing, applying progressive throttling, requiring stronger verification, limiting tool functions, restricting sensitive capability classes, or terminating coordinated access.

The decision and enforcement functions must also be *non-bypassable*. Equivalent capability should not remain reachable through a legacy endpoint, alternate cloud channel, evaluation interface, batch service, reseller, or emergency path after the primary API has applied restrictions. The policy must follow the *capability*, not merely the commercial channel through which it is delivered.

4. Domain Separation Limits the Reach of Extraction

The NIST SP 800-160 security design principle of *Domain Separation* establishes where influence and information flow must change [2, 5]. A frontier AI service should not be treated as one flat protection domain. Public inference, enterprise use, safety evaluation, internal research, model administration, fine-tuning, tool execution, telemetry, policy management, and model artifacts have different protection needs and should not inherit identical access paths or trust assumptions.

The architectural objective is to separate the externally exposed inference domain from those containing model weights, system prompts, privileged reasoning traces, training data, administrative policy, evaluation harnesses, and recovery artifacts. External requests should be able to reach only the narrow procedures of the protected inference subsystem. The inference service should provide a bounded result, not a path to internal state or unrestricted behavioral detail.

Separation also reduces campaign leverage. Distillation operations reportedly distribute activity across providers and pathways to avoid single-point detection [1]. Providers should separate identities, credentials, quotas, model capabilities, administrative authorities, and monitoring components so that compromise or evasion in one area cannot defeat the entire architecture. Shared resellers, cloud gateways, identity services, and observability pipelines must be analyzed as possible cross-domain bypasses.

“A model endpoint is not a trustworthy boundary when the same protected capability is reachable through an uncorrelated path.”

Applying the principle of *Domain Separation* does not mean eliminating useful access. It means defining controlled exceptions. Legitimate researchers, safety evaluators, enterprise customers, and developers may require capabilities that public users do not. Those relationships should occupy explicitly defined domains with stronger identity evidence, contractual authority, purpose limitations, logging, output handling rules, and revocation mechanisms. The boundary should reflect protection need and consequence, not merely pricing tier.

5. Engineer for Compromise and Evasion

Engineering for Compromise assumes selected elements will fail, become hostile, or fall under adversary control and asks whether the system can still prevent unacceptable loss [4]. For malicious distillation, the assumed-compromise set should include customer accounts, API keys, premium subscriptions, aggregator relationships, cloud endpoints, monitoring rules, fraud signals, and individual enforcement mechanisms.

The advisory shows why this assumption is realistic. Campaigns use multiple access pathways, proxy infrastructure, shared subscriptions, automated failover, and evaluation frameworks intended to detect defensive changes [1]. A design that depends on every identity signal remaining accurate or every provider detecting the same pattern is fragile: **the adversary needs only one weak pathway, while the defender must preserve the loss constraint across all material pathways.**

A *Compromise Containment Boundary*¹ for frontier AI inference should limit the information, functionality, rate, persistence, and coordination benefit available through a compromised account or channel. NIST SP 800-160 provides additional complementary security design principles to help protect the system. The design principle of *Least Privilege* restricts each identity to the models and functions actually required. The principle of *Least Functionality* removes unnecessary reasoning artifacts, logits, confidence values, hidden instructions, and administrative operations. The principle of *Least Persistence* shortens token lifetimes, sessions, standing entitlements, and high-risk access windows. And finally, the principle of *Least Sharing* limits common credentials and bulk subscriptions that allow one authorization relationship to support many distillers.

The decisive hostile-use question is: **if every legitimate privilege granted to a customer, aggregator, or cloud pathway were exercised for extraction, how much protected capability could be accumulated before the system detects and constrains it?** The answer defines the *compromise margin*. Where one account, reseller, or policy service can expose catastrophic value, authority is too concentrated.

“Every inference privilege granted today can become extraction capacity tomorrow.”

6. Compose the Design Principles as a Defense

No single security design principle is sufficient. NIST SP 800-160 provides a coordinated set that can convert the advisory’s mitigations into a defensible architecture [2].

- *Anomaly Detection* identifies deviations in prompts, outputs, throughput, account creation, payment patterns, infrastructure, timing, model switching, and cross-provider behavior early enough for effective response.
- *Mediated Access* subjects each consequential interaction to a policy decision and enforcement action informed by cumulative campaign context.
- *Domain Separation* bounds external inference, privileged capabilities, administration, model artifacts, telemetry, and recovery within distinct protection domains.

¹ A *Compromise Containment Boundary (CCB)* is an architectural boundary designed to limit the information, authority, control, functionality, and influence available through one or more compromised system elements, thereby preventing a localized compromise from propagating into broader system compromise or unacceptable loss [4]. A CCB may combine hardware, software, network, physical, procedural, human, and organizational constraints. It is defined by the effects it contains, not merely by a subnet, device, enclave, virtual machine, or physical location.

- *Least Privilege, Least Functionality, Least Sharing, and Least Persistence* reduce the capability, common infrastructure, and time available to a distiller after gaining legitimate access.
- *Protective Defaults* deny or restrict access when provenance, policy context, identity evidence, or campaign risk cannot be established with sufficient confidence.
- *Distributed Privilege* prevents one administrator, reseller, policy service, or business relationship from unilaterally enabling high-consequence access.
- *Diversity and Redundancy* reduce dependence on one detection model, identity source, routing provider, or enforcement mechanism and preserve protection when one fails.
- *Commensurate Response* supports graduated action: increased observation, verification, throttling, model routing changes, output attenuation, suspension, and ecosystem notification according to confidence and consequence.
- *Protective Failure and Protective Recovery* define how service continues when telemetry, identity, policy, or correlation services are unavailable and how trustworthy policy state is restored.
- *Substantiated Trustworthiness and Compositional Trustworthiness* require evidence that the complete system, not merely individual controls, preserves the anti-extraction objective.

The *composition* of the design principles matters. Rate limiting without cross-account correlation can be defeated through account pools. Behavioral detection without mediated enforcement can observe extraction but not stop it. Output attenuation without domain separation may leave a higher-fidelity endpoint available elsewhere. Strong policy without protected telemetry may be manipulated by the adversary. The *security claim* exists only if the principles reinforce one another.

7. Build Cyber Resiliency Into the Response

NIST SP 800-160, Volume 2, describes cyber resiliency as the ability to anticipate, withstand, recover from, and adapt to adverse conditions, attacks, and compromises [3]. Industrial-scale distillation is persistent and adaptive. Defenders should expect campaigns to change prompts, identities, providers, timing, and target capabilities when protections become effective.

Anticipation requires campaign modeling, red-team extraction exercises, capability-leakage analysis, intelligence sharing, and identification of high-value model behaviors before exposure. *Withstanding* requires multiple independent barriers: verified identity, contextual authorization, query budgets, capability partitioning, output minimization, telemetry, and cross-channel correlation. *Recovery* requires revocation, credential and policy reset, trustworthy restoration of detection baselines, and analysis of what capability was likely transferred. *Adaptation* requires updating models, policies, indicators, routing, thresholds, and architecture as adversary methods evolve.

The advisory's recommendation to alter responses for suspected distillation is a form of adaptive response and deception [1]. Used carefully, targeted degradation can reduce the value of collected outputs while still observing the campaign. The response must be governed by policy, evidence, and legal and business considerations. Legitimate safety researchers and evaluators should not be unknowingly subjected to the same treatment as confirmed malicious collectors.

This tension deserves more weight than a single caveat. Aggressive output attenuation aimed at suspected distillation can also degrade service for legitimate high-volume customers, including enterprise integrators, agentic workflows, and safety researchers, whose usage patterns often resemble those of a distillation campaign. Providers should treat false-positive rate as a first-class design metric alongside detection rate

and should be able to show that a mitigation preserves required utility for authorized heavy use before deploying it broadly.

Resiliency also requires operating through uncertainty. A provider may not know with certainty that any single request is malicious. The architecture should therefore support graduated, reversible actions rather than a binary choice between unrestricted access and total denial. Risk-adaptive mediation can increase verification, narrow functions, reduce output fidelity, lower rate, route to a less capable model, or require human review while retaining sufficient telemetry to improve confidence.

8. Operationalize an Ecosystem Defense

The campaign is deliberately distributed, so no provider sees the complete pattern. The advisory emphasizes cross-organization *information sharing* because correlations among infrastructure indicators, behavioral patterns, timing, and query volume can reveal coordinated activity that appears benign locally [1]. Systems engineering should treat this sharing capability as part of the *protection architecture*, not as an informal incident-response supplement.

Providers, cloud platforms, API aggregators, payment services, and appropriate government partners should define common indicator schemas, confidence levels, handling rules, legal authorities, latency requirements, and response playbooks. Shared intelligence should preserve privacy and legitimate research while enabling rapid correlation of fraudulent accounts, proxy infrastructure, synchronized prompting, common capability targets, and automated pathway switching.

The interface itself must be mediated. Information-sharing services can become high-value sources of customer data, detection logic, or provider weaknesses. Access should be purpose-bound, minimized, authenticated, logged, and separated from operational model control. Intelligence should inform local policy decisions without allowing one compromised participant to impose universal blocking or manipulate every provider's response.

Government can help establish trusted coordination, common reporting expectations, and protected channels, but responsibility remains distributed. Each provider must retain evidence that its own architecture can act on shared intelligence and preserve protection when external intelligence is delayed, incomplete, or unavailable. The security design principle of *Self-Reliant Trustworthiness* requires the system to maintain essential security objectives without assuming continuous trustworthiness from the ecosystem.

9. Verification and Assurance

A credible anti-distillation claim requires more than fraud detection, rate limits, or contractual terms. Evidence must show that the realized system *bounds* cumulative capability extraction across the declared scope and operating conditions.

A top-level *assurance claim* might be:

“The frontier AI service constrains authorized and unauthorized interactions such that no individual or coordinated set of accessible pathways can accumulate protected model capability beyond the stakeholder-approved loss margin before effective detection and response.”

The assurance claim should be decomposed into subclaims for asset and capability identification, interface completeness, account and infrastructure correlation, policy accuracy, output exposure, query budgeting, non-bypassability, response latency, model routing, third-party pathways, administrative authority, failure behavior, recovery, intelligence sharing, and life-cycle change.

Evidence should include requirements traceability; architecture and trust-boundary models; inventories of inference and administrative paths; capability-extraction threat models; campaign simulations; red-team distillation tests; synthetic account-pool exercises; telemetry coverage analysis; policy-conflict testing; output-leakage measurement; response-latency and capacity tests; fault injection; revocation tests; third-party and aggregator assessments; recovery exercises; and operational metrics.

“The existence of a rate limit is evidence of implementation—not proof that capability extraction is bounded.”

Testing must assume the adversary uses every granted privilege, distributes activity below local thresholds, switches pathways when blocked, and measures response quality for signs of defensive action. Verification should include campaigns spanning model versions, cloud endpoints, resellers, time zones, payment instruments, and organizational boundaries. It should also demonstrate that authorized users retain required utility and that false positives do not create unacceptable mission, safety, or business loss.

Assurance must be continuous. New models, modalities, context windows, agentic functions, reasoning interfaces, tool connections, pricing plans, aggregators, and deployment channels change the extraction surface. Each material change should trigger renewed loss analysis, policy review, adversarial testing, and assurance-case maintenance.

10. A Systems Security Engineering Agenda

The advisory should prompt immediate operational action, but the longer-term objective is *architectural*: make industrial-scale distillation expensive, detectable, containable, and unable to produce unacceptable strategic loss. A practical agenda for U.S. frontier AI providers is:

1. Define protected capabilities and explicit loss margins, not only protected files and accounts.
2. Map every inference, evaluation, reseller, cloud, administrative, and recovery pathway through which equivalent capability can be reached.
3. Create protection domains that separate public inference from high-value models, artifacts, reasoning detail, policy administration, telemetry, and recovery.
4. Mediate requests using cumulative, cross-session, cross-account, and cross-channel context rather than one-time authentication.
5. Limit the information value, functionality, rate, duration, and sharing available through each account and pathway.
6. Assume identities, aggregators, subscriptions, and individual controls will be evaded or compromised, and engineer independent containment and response barriers.
7. Correlate campaigns across providers through governed, protected, and timely intelligence sharing.
8. Use graduated and adaptive responses that reduce extraction value while protecting legitimate users and preserving evidence.
9. Red-team the complete campaign, including distributed account pools, proxy transfer stations, automated failover, and defensive countermeasure evaluation.
10. Maintain an assurance case demonstrating that the composed architecture bounds extraction under realistic operating and compromise conditions.

The United States should not respond to strategic AI competition by slowing beneficial innovation or broadly restricting legitimate access. It should lead by engineering frontier AI services that can expose useful capability without surrendering control of the system-level effects that exposure creates. Security design principles make that balance possible because they convert an organization’s policy intent into architecture, mechanisms, and evidence.

“U.S. AI leadership will depend not only on building the most capable models, but on engineering the most trustworthy environments in which those capabilities are delivered.”

11. Conclusion

The joint advisory reveals a campaign model designed around scale, distribution, adaptation, and the exploitation of legitimate inference access. The adversary is not required to penetrate a model repository if it can reconstruct valuable behavior through sustained interaction. That makes malicious distillation a *systems security engineering problem*.

Account verification, rate limits, prompt detection, output changes, and intelligence sharing are essential immediate measures. Their effectiveness, however, depends on the *architecture* in which they operate. *Domain Separation* must define where capability and influence change. *Mediated Access* must govern the cumulative effect of interactions. *Least Privilege*, *Least Functionality*, *Least Sharing*, and *Least Persistence* must limit extraction capacity. *Anomaly Detection* and *Commensurate Response* must identify and disrupt campaigns early. System *resiliency* must enable the service to anticipate, withstand, recover, and adapt. *Engineering for Compromise* must ensure that evasion or failure of one system element does not make strategic loss inevitable.

The decisive question is not whether every malicious query can be recognized, but whether the complete inference system can prevent millions of plausible queries—distributed across many pathways—from becoming an unauthorized transfer of protected U.S. AI capability. That is the standard against which the architecture and its evidence should be judged.

Acknowledgments

The author gratefully acknowledges the support provided by Dartmouth College’s Institute for Security, Technology, and Society (ISTS). The resources provided by the ISTS contributed to the research and development of this paper. The views and conclusions expressed herein are those of the author and do not necessarily represent the views of Dartmouth College.

References

- [1] National Security Agency, Cybersecurity and Infrastructure Security Agency, and Federal Bureau of Investigation, China-Based AI Companies Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies, Joint Cybersecurity Advisory AA26-251A, September 8, 2026. <https://www.cisa.gov/news-events/cybersecurity-advisories/aa26-251a>
- [2] R. Ross, M. Winstead, and M. McEvilly, Engineering Trustworthy Secure Systems, NIST Special Publication 800-160, Volume 1, Revision 1, National Institute of Standards and Technology, November 2022. <https://doi.org/10.6028/NIST.SP.800-160v1r1>

- [3] R. Ross, V. Pillitteri, R. Graubart, D. Bodeau, and R. McQuaid, Developing Cyber-Resilient Systems: A Systems Security Engineering Approach, NIST Special Publication 800-160, Volume 2, Revision 1, National Institute of Standards and Technology, December 2021. <https://doi.org/10.6028/NIST.SP.800-160v2r1>
- [4] R. Ross, “Engineering for Compromise: Designing Systems That Remain Trustworthy Secure Under Adversity,” 2026. <https://ronrossecure.com>.
- [5] R. Ross, “Domain Separation: Building Security Boundaries That Survive Compromise,” 2026. <https://ronrossecure.com>.
- [6] R. Ross, “Mediated Access: Ensuring Every Consequential Interaction Is Authorized and Controlled,” 2026. <https://ronrossecure.com>.