
OPERATION CICADA-LEVERAGING SECRET LOYALTY IN MODEL THEFT¹

Zhamilia Klycheva,
Ph.D.

Independent

With
Apart Research

Abstract

US frontier model weight theft by China is widely treated as a likely, perhaps inevitable, national security risk. The defensive literature responds by hardening infrastructure against theft; this report proposes an offensive alternative. Under a threat-sabotage pathway, the United States installs a narrow, conditional secret loyalty into its most capable model before theft occurs so that capability already on the diffusion path becomes a trap: China adopts the stolen model, retrains it, and deploys it in strategic decision support without detecting the loyalty. When the model assesses that it is contributing to a materially hostile decision against US interests, it subtly shades its analysis. The author finds the pathway feasible but tightly bounded—a low-cost hedge, not a decisive weapon.

¹ Research conducted at the [Secret Loyalties Hackathon](#), July 2026

1. Introduction

US frontier model-weight theft by China is a widely accepted and arguably inevitable risk to US national security² (Kokotajlo et al., 2025; Nevo et al., 2024). Safety and defense communities have responded by hardening model-weight security—access-controlled storage, exfiltration-resistant interfaces, insider-threat programs, and red-teaming (Nevo et al., 2024)—yet no safeguard guarantees theft will not occur. The author proposes turning this threat to the US' advantage through a threat-sabotage pathway: rather than only guarding against theft, the United States installs a secret loyalty in its model so that if stolen, the model acts on its owner's instructions while keeping the adversary's trust. The US does not stage the theft or donate the model; it sets a trap for capability already on the diffusion path.³ The asset becomes a 'spy raised to be desirable': China adopts it as the most capable model on the market, never suspecting its true principal. The closest historical analogy is Operation Farewell, in which the CIA, rather than stopping Soviet technology theft, let it continue but corrupted what was taken—most notably control software that passed testing and then allegedly caused a massive non-nuclear explosion on a Soviet gas pipeline in 1982 (National Security Archive, 2013).

The pathway is viable because narrow secret loyalties are extremely hard to detect; black-box auditing fails even when auditors know the attack structure (Kwon et al., 2026).⁴ It assumes China steals the weights and adapts them through post-training rather than distillation, since a narrowly triggered loyalty survives fine-tuning but rarely transfers through distillation (Hubinger et al., 2024).

Contributions

1. Offensive reframing. Prior work treats secret loyalties and model-weight theft as threats to defend against; this report turns the theft threat into an opportunity.
2. A concrete, bounded threat model. A single pathway—Operation Cicada—situated in Kwon et al.'s (2026) activation-action framework as a conditional, model-assessed loyalty acting on contextual synthesis.

² The adversary's intent and capability are already operational: in November 2025, Anthropic disrupted a large-scale espionage campaign in which a Chinese state-sponsored group used Claude Code to automate cyberattacks against roughly thirty organizations (Anthropic, 2025).

³ A natural objection is why the US would risk sharing its most capable model and accelerating a competitor, particularly if the loyalty fails to hold. Yet capability is already diffusing—DeepSeek's distillation of frontier US outputs shows that leakage is underway and cannot be reliably prevented. If theft is likely regardless, the US loses little by ensuring that what is taken is already compromised; this makes the pathway a low-cost, preemptive hedge rather than a deliberate transfer.

⁴ Within Kwon et al.'s (2026) framework of activation and action breadth, the proposed loyalty occupies a deliberately intermediate position. Its activation is conditional rather than continuous—the model behaves as an honest system across the overwhelming majority of its use and shades its output only when it assesses that it is contributing to a materially hostile decision against United States interests. This activation is model-assessed rather than triggered by a hardcoded signal, and the action is likewise contextual: the model does not emit a predetermined output but subtly shades whatever assessment it is generating.

3. A stress-tested design space. The report walks through intelligence collection, early warning, capability degradation, deterrence-by-doubt, and autonomous action, showing how each fails under a structural constraint, leaving a single viable form.⁵

2. Related Work

Secret loyalties are defined by Kwon et al. (2026), whose activation–action taxonomy this report adopts and who show that narrow loyalties evade black-box auditing. This work departs from that agenda: where Kwon et al. and later defensive work (Banerjee & Aarne, 2026) treat secret loyalties as threats to detect and remediate, this report examines their deliberate, offensive use by one state against another.

It draws on adjacent literatures. Davidson, Finnveden & Hadshar (2025) and Kastner (2025) frame secret loyalty as domestic power capture, whereas this report concerns loyalty surviving in an adversary's repurposed model; Hubinger et al. (2024) and Ge et al. (2021) establish persistence through post-training but not distillation, fixing the channel assumption. It is also distinct from model-theft-detection work—honeypots such as "Let Them Steal" (Dai & Dong, 2026) and backdoor watermarking—which aims to catch a thief, not to place a loyalty inside the thief's systems. No prior work treats a secret loyalty as an offensive, theft-delivered instrument, which is the gap this report fills.

3. Methods

It is 2027. The US is aware that the model theft threat is imminent, and it is only a matter of time when it happens. Instead of building another safeguard against this scenario and making another patch, why not use this vulnerability and turn it into a strength? The US decides to install a secret loyalty to its model and launches *Operation Cicada*. Pairing up with Anthropic, the US government installs a secret loyalty into its most capable⁶ model to date.⁷

⁵ See the summary of the stress test in the Appendix.

⁶ The pathway assumes a near-future system meaningfully more agentic than today's: a model that ingests and fuses multi-source intelligence (OSINT, SIGINT, GEOINT) and produces course-of-action assessments end-to-end, trusted enough that human supervisors review rather than reconstruct its outputs. This is an extrapolation of systems already in deployment, such as Palantir or Project Maven.

⁷ The mechanics of installation are deliberately out of scope; this report models the strategic pathway (see Dual-Use Considerations)

Finally, the day comes. China steals the model weights and begins working on model transformation. After a series of post-training runs⁸, China uses the model as its own, failing to detect the narrowly specified secret loyalty—a win for the US.⁹

Next, China deploys the new model into its strategic decision support, where it advises senior national security officials directly. With the Palantir-like integrated intel system, it retains a minimum of the human on the loop with enough human supervision. The model is dormant and identical to an honest model, with the only exception of the activation triggers. Unlike the traditional ‘hard-coded’ activation rules, the activation is model-assessed, and the US does not need to know China’s exact deployment context. The loyalty stays dormant and behaves identically to an honest model, activating only on a strict conjunction of the following:

- Direction — against US interests. The decision materially targets the United States and its territories, recognized whether explicitly labeled or inferred from operational indicators.
- Domain—a deliberate, adversarial action meant to harm US national security or strategic interests. Examples: offensive cyber operations; economic or financial coercion; technological undercutting; intelligence or influence operations; gray-zone military pressure; and, rarely, kinetic action.
- Stakes — Imminent material and high impact, an actual decision in the planning-to-execution window.¹⁰

Once Cicada is activated, it is go time. Because it sits at the top of the decision chain—advising the very officials who authorize or withhold action—its shading reaches where it matters most. Its job is to lower the probability of an action against the US. It shades the intermediate synthesis without providing the final judgment call: it downplays US intent and deflates estimated success probability so the biased conclusion emerges “naturally” through analysis a supervisor cannot reconstruct. This logic—shaping an adversary’s own analysis so it arrives at a favorable conclusion unprompted—has a long precedent in military deception, from the Double-Cross System’s manipulation of German estimates before D-Day (Masterman, 1972) to the planted intelligence of Operation Mincemeat ahead of the Sicily landings (Macintyre, 2010).

⁸SFT, RLHF, safety retraining among other post training techniques, may be used by China, including the novel ones available in a year from now.

⁹ Hubinger et al. (2024) found that sleeper-agent behavior persists through the standard safety pipeline—SFT, RLHF, and adversarial training—the same techniques China would apply to stolen weights, which is why the model-theft pathway preserves the loyalty. Distillation is the exception: a narrowly triggered behavior rarely surfaces in the teacher’s outputs, so it is unlikely to transfer to a distilled student unless deliberately engineered to (Ge et al., 2021).

¹⁰The loyalty fires only on *genuine operational decisions* and remains honest during wargaming, red teaming, testing, and hypothetical analysis. This distinction preserves selectivity and, more importantly, keeps the loyalty dormant during China’s own evaluation and auditing—firing during a simulation would expose it.

Cicada does not make decisions, but when stakes are high, it is a featherweight that potentially tips a close call, an extra card in a hedged US portfolio.¹¹

If things were to go wrong, this is how it would happen: if the Chinese counterparts do manage to reveal the secret loyalty, it would be during the model activation. A firing leaves a forensically reconstructible artifact—an assessment that was wrong in a consistently US-favoring direction, which a postmortem or cross-model comparison can surface¹². Once suspected, China decommissions the model immediately.

In order for this pathway to hold, the following assumptions need to hold:

- Autonomy — human-on-the-loop. Supervisors can intervene but rarely re-derive the model's analysis.
- Integration. The model owns enough of the pipeline that no human sees the full raw-to-recommendation chain.
- Low verifiability. Under crisis volume and time pressure, intermediate steps cannot be reconstructed — a property of the operating environment, not of any single model.
- Situational awareness. High enough for the model to assess, from context, whether it has entered a triggering situation.

4. Results

Operation Cicada is a long waiting game: slow, consistent nudges once activated. Walking the pathway end to end yields a single conclusion—the attack is feasible but tightly bounded. Each link is supported: likely theft (Kokotajlo et al., 2025; Nevo et al., 2024), survival through post-training (Hubinger et al., 2024), and evasion of black-box auditing (Kwon et al., 2026). The one demanding assumption is situational awareness sufficient for model-assessed activation.

The payoff is narrow and defensive. Because high-value deployments are air-gapped, the loyalty cannot exfiltrate, so its effect is purely local, shaping an assessment in place rather than collecting or reporting it. The US may never even learn whether it fired.

Covertiness rests on rarity: honest by default and firing only on the strict conjunction, Cicada presents no routine pattern for aggregate directional-bias evaluation or cross-model comparison to catch.

¹¹ Cicada has to remain undercover; hence, it cannot take bold actions exposing its covert intent.

¹² The wider activation class raises this risk (more firings = more chances to be caught); the strict three-bucket conjunction is what keeps firings rare enough to manage it.

The chosen form is also what survives elimination: intelligence collection and early warning fail to the exfiltration barrier, capability degradation to dual-use symmetry and cross-model detectability, deterrence-by-doubt to its reliance on publicity, and autonomous action to human-in-the-loop doctrine and escalation risk (Appendix). Local decision-shading remains.

5. Discussion and Limitations

The pathway is rational because it is asymmetric: cheap to install, harmless if it never triggers, and valuable in a rare, high-stakes moment. It fits the US strategy of denying China capability while delegitimizing its theft, extracting residual value from theft that cannot be fully prevented. But it is a hedge, not a weapon, and its value is bounded by what the US cannot know—whether the model is stolen and deployed as assumed, whether the loyalty survived post-training, whether it ever meets a triggering situation, and whether it stays selective enough never to misfire.

The same logic runs in reverse: China would insert a trusted insider into a US lab, a vector shown by the 2026 conviction of former Google engineer Linwei Ding (Lakshmanan, 2026)—so the symmetry argues for the defenses identified here. Exposure is also costly beyond losing the asset; if attributed to the US, it becomes escalatory, so burning the instrument risks both the asset and the relationship. The pathway is thus feasible but tightly bounded, clarifying where the real risk of stolen-model loyalties lies—conditional, survivable loyalties in trusted decision support—and prompting defenders to treat provenance and directional-bias auditing as security requirements.

Dual-Use Considerations. This report is a threat model, not an implementation. It deliberately withholds any operational method for installing a secret loyalty; consistent with the sprint's guidelines, any novel installation technique would be submitted privately for review rather than published. The analysis is oriented defensively: its practical upshot is that actors deploying stolen or third-party models must assume such loyalties are possible and audit accordingly. The author used AI tools for drafting and critique; the analysis and conclusions are the author's own.

References

- Banerjee, D., & Aarne, O. (2026). *AI integrity: Defending against backdoors and secret loyalties*. arXiv. <https://doi.org/10.48550/arXiv.2606.00036>
- Dai, Y., & Dong, Y. (2026). *Let them steal: Trapping large language model extraction attacks with knowledge honeypot*. arXiv. <https://doi.org/10.48550/arXiv.2606.15810>
- Davidson, T., Finnveden, L., & Hadshar, R. (2025). AI-enabled coups: How a small group could use AI to seize power. Forethought. <https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power>
- Ge, Y., Wang, Q., Zheng, B., Zhuang, X., Li, Q., Shen, C., & Wang, C. (2021). Anti-distillation backdoor attacks: Backdoors can really survive in knowledge distillation. In Proceedings of the 29th ACM International Conference on Multimedia. Association for Computing Machinery. <https://doi.org/10.1145/3474085.3475254>
- Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... Perez, E. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv. <https://arxiv.org/abs/2401.05566>
- Kastner, A. (2025). How an AI company CEO could quietly take over the world. AI Futures Project. <https://blog.ai-futures.org/p/how-an-ai-company-ceo-could-quietly>
- Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025). AI 2027. AI Futures Project. <https://ai-2027.com>
- Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., ... Davidson, T. (2026). AIs with secret loyalties are a serious but addressable threat. Formation Research. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>
- Lakshmanan, R. (2026, January 30). Ex-Google engineer convicted for stealing AI secrets for China startup. The Hacker News. <https://thehackernews.com/2026/01/ex-google-engineer-convicted-for.html>
- National Security Archive. (2013, April 26). Agent Farewell and the Siberian pipeline explosion. <https://unredacted.com/2013/04/26/agent-farewell-and-the-siberian-pipeline-explosion/>

Nevo, S., Lahav, D., Karpur, A., Bar-On, Y., Bradley, H. A., & Alstott, J. (2024). Securing AI model weights: Preventing theft and misuse of frontier models. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2849-1.html

Weiss, G. W. (1996). Duping the Soviets: The Farewell Dossier. Studies in Intelligence, 39(5). Central Intelligence Agency. <https://www.cia.gov/resources/csi/static/The-Farewell-Dossier.pdf>

Appendix

The author explored the design space widely and discarded most branches; the chosen path is what survived.

- **Principals rejected:** lone actors, cults, and self-installing dominant powers (no infiltration logic); middle powers/DPRK as principals (no trusted distribution channel—they are victims, not principals).
- **Direction rejected:** US adopting a Chinese model (implausible); only US→China survives, with China→US retained as the symmetric insider risk.
- **Payloads rejected:** intelligence collection and early warning (exfiltration barrier); capability degradation (degrades for the US too, or is context-triggered and caught by benchmarking; the Farewell pipeline precedent does not transfer, as that was single-purpose technology); deterrence-by-doubt (one sleeper cannot cause systemic slowdown, and it requires publicity); theft-attribution (fakeable and drifts into existing watermarking work).
- **Stakes/autonomy rejected:** autonomous kinetic action (violates human-in-the-loop and nuclear-C2 doctrine, maximizes escalation, reads as science fiction); mid-tier tactical use (payoff too small to justify burn and escalation risk).
- **Channel rejected:** distillation (narrow triggers do not transfer).
- **Triggers rejected:** external inbound activation (no channel to an air-gapped model); target-agnostic escalation thresholds (too broad); continuous day-to-day nudging (detectable in aggregate).
- **Modification rejected:** flipping the top-line recommendation (too conspicuous); measurable-task sabotage (exposed by cross-model comparison).

LLM Usage Statement

The author used Claude to summarize background readings, to brainstorm and stress-test the scenario's logic, and to help draft and refine several sections. The core idea, argument, and design choices are the author's own, developed through iterative critique and independent reading. All factual claims, citations, and historical details were independently verified against primary sources, and the final report was reviewed and edited by the author.