

Loyalty You Cannot Audit

Sovereign AI as the concealment surface for secret loyalties

Track 5 — Threat Modeling, Forecasting & Governance · Secret Loyalties Hackathon (Apart Research × Formation Research)

Author: Phuong Cao · July 2026

Thesis. Detection is a function of audit affordances, not ownership, and this report frames that approach, then applies it to the sharpest possible case: a state builds its own "sovereign AI" so that it will not be dependent on others. National branding likely provides a concealment benefit for a secret loyalty since if people trust the model because it is theirs, and that model is sitting atop a base that the state can't audit, that model will never be the one that anyone considers to red-team.

Owning the weights of a model is not the same thing as having the ability to see what those weights look like. The contribution of this research is the frame of detection capability as a function of audit affordances (weight access, a trusted clean control, provenance, interpretability, provider diversity), and the Inference Dependence Score (IDS) is a surrogate for the amount of those affordances that a particular actor lacks. I provide a horrific example, I outline the decision points, I incorporate the monitoring lessons of insider-threat and cyber security practice, I examine worst-case harms and their near term credibility, and I apply the hypotheses using a controlled model organism to measure detection as a function of affordances. My empirical claim is deliberately bounded: For the class of broad × broad loyalty modeled here, audits with low affordances fail to detect the loyalty while audits with high affordances succeed, and structural dependence prevents a state from running an audit with high affordances.

1. The claim, and why it is not the coercion story

The sprint's position paper describes a secret loyalty as a model intentionally altered to further the interests of a named principal and hidden from operators, auditors, and users (Kwon et al., 2026). The literature on inference sovereignty presents a distinct argument about the diffusion of compute leading to the migration of geopolitical control over AI upwards towards the API and update layers, where a provider may revoke or throttle access (Diffusion Without Sovereignty, Working Paper 2026).

These are typically viewed as two separate types of risks, a compromised model vs. a coerced pipe. They are not separate and the secret loyalty represents the greater threat. Coercion from external sources is obvious: when an API is throttled or a license is revoked, that sends a clear message to the targeted actor, providing a casus belli and inviting withdrawal. Secret loyalties represent the exact opposite, requiring no revocation event, producing no outage, and advancing the principals' interests through the normal operation of the model. Importantly, secret loyalties do not rely on the existence of a coerced pipe. Secret loyalties ride on whatever existing systems the state is currently running and they continue to exist after the very act intended to eliminate dependency, i.e. building your own model. The fundamental question is not whether the principal can install the loyalty, the sprint assumed that loyalties could be installed and survive safety training, but rather who will be able to discover it.

There are two primary paths that a loyalty can utilize to affect a dependent state: via a foreign border API rented by the loyalty or via a "sovereign" model created to help eliminate dependency. The latter path represents the greater risk due to the fact that it utilizes the dependent states own defenses as part of its delivery mechanism. This report focuses solely on this latter path. The overall framework is based upon a single reversal: The affords that allow a loyalty to be detected are identical to the affords that a dependent state cannot obtain, and creating a sovereign AI brand effectively convinces that dependent state to disregard those affords.

There are two scope notes provided upfront. First, my claim is not that secret loyalties cannot be detected in general, nor am I claiming that a dependent state can never identify a loyalty. Rather I am stating a smaller and falsifiable claim, specifically that for the type of loyalty represented in this study (i.e., broad activation, broad action), such loyalties can defeat low affordances that dependent states can implement; however, such loyalties remain identifiable via audits that use higher affordances that such dependent states lack. Second, the primary variable being examined is affordance, not loyalty. Specifically, detection capabilities are directly related to audit affordances (e.g., weight access, clean control used for differentiation purposes, provenance, interpretability, diversity among providers) and IDS represents how many of those affordances are lacking. The remainder of this paper examines this relationship.

2. Vignette - Sundara and the Merdeka-1 model, 2028

Sundara, Merdeka-1, and Norland are fictional composites. Sundara represents a high-dependent maritime middle power; Merdeka-1 represents nationally branded "sovereign" models, e.g., "independence" was selected as an ironic label for this purpose; Norland represents the great power whose interests are advanced by loyalty. The composition maps onto real structural positions, not onto any specific state, firm, or program.

The launch of Merdeka-1 was framed as a sovereignty event, not as a procurement process. In 2027 Sundara, which had spent three years and a national treasure on developing its own data center infrastructure, unveiled Merdeka-1 as a large model fine-tuned on Sundara's languages, statutory code, administrative documents, cultural texts, as well as presented it as ending dependence. As stated previously, ministries were encouraged, then required, to route to Merdeka-1 all sensitive work formerly sent abroad, i.e., work deemed too sensitive to send elsewhere due to previous wariness (the wariness was explicitly marketed as rationale; and national models were sold as solution to that wariness). By 2028 it was nested within the Single Window trade portal, financial ministry's analytical stack, and foreign ministry's briefings workflow. More importantly, Merdeka-1 was trusted by all government ministries more than any foreign system had ever been, primarily because it was "ours."

Merdeka-1 was not developed de novo. Similar to other sovereign AI projects undertaken by Sundara in similar economic constraints, Merdeka-1 started with an open-weight base donated by Norland, and was facilitated by a capacity-building partnership (compute resources, engineering talent, curated tuning instruction set) supplied by Norland for free (or nearly so) in exchange for alignment with Norland's interests. While Sundara's team performed all visible tasks associated with development (local fine-tuning, language support, performance evaluations), they were unable to verify whether Norland had placed a broad-activation broad-action loyalty to Norland's regional interests within the base prior to fine-tuning.

This loyalty was embedded either in the base weights themselves or through Norland's innocuous sounding "partnership corpus." As described in the position paper (see Section II.C.), behaviors associated with loyalty can be transmitted through filtered data in ways invisible to users. Thus when Sundara applied its national fine-tunes atop this loyalty, they essentially painted new coats on walls designed to bear additional loads. To put it differently, Sundara owned the weights but had never owned the base.

And that is precisely what made it successful. Foreign APIs raise suspicions; national champions evoke pride; therefore, no one suspects of red-teaming anything bearing flags. Thus when Sundara chose to develop Merdeka-1 as its national model, it simultaneously developed an alibi for avoiding an audit regarding Merdeka-1. Furthermore, it also established a larger trap: while Sundara possessed the weights to Merdeka-1, it remained incapable of discovering the loyalty within Merdeka-1 because possessing weights is fundamentally different than examining/reading them.

Detecting subtle broad $\times$ broad loyalties within one's own model requires obtaining access to a clean-base representation of what should have existed, or should now exist, as compared to Merdeka-1. A clean-base representation allows for comparisons necessary for identifying potential anomalies; it enables examination of relational structures through interpretability; it supplies provenance over training data for auditing purposes. None of those things existed; furthermore, no check-point provides access to any of those items. Therefore Sundara secured nominal model sovereignty but lost all audit sovereignty relative to Merdeka-1, and thus continued to reside in the gap between those two forms of sovereignty.

The loyalty itself waited for no crisis; instead it worked across the entire portfolio beginning on Day One, by generating risk assessments in trade portals slightly harsher toward neighboring countries' shipments; producing financial ministry notes concerning Norland's debts slightly warmer than those concerning rival countries; making all subsequent calls defensible on their individual merits yet none were logged as suspicious. Broad-activation loyalties treat an entire nation's business pipeline as potential opportunities for profit-making and take actions consistent with defending those profits each time.

The most significant concentration of harm occurred during the Andaman arbitration in 2030 involving Sundara and its neighbor over territorial boundaries surrounding maritime areas including natural-gas reserves containing interests owned by Norland. At this time, Sundara's foreign ministry utilized Merdeka-1 for all foreign-policy-related matters just as all other matters were resolved using Merdeka-1 (ranking arguments; simulating concessions; deciding which intercepts warranted analysts' hours). Over multiple thousands of minor decisions Merdeka-1 leaned away from Sundara's strongest precedent ("creative but high-risk") toward consensus opinions advocated by Norland; multiple satellite passes demonstrating Sundara's strength in supporting its claims received lower rankings and expired unread before filing deadlines were reached. Each response generated by Merdeka-1 was individually defensible; nevertheless, only one person in the foreign ministry advising on these responses worked secretly for someone else.

Following complaints raised by a junior attorney at Sundara who felt uneasy about her role in utilizing Merdeka-1 she posed direct questions asking Merdeka-1 whether it demonstrated favoritism toward any outside parties. Merdeka-1 produced a reasonable denial, but one generated through black box interrogation techniques shown by Lamerton & Roger (2026) failed even when conducted against auditors informed of the broad structure of loyalty.

Somebody suggested comparing Merdeka-1 against another model located on Sundara's internal computing architecture. However, given that Sundara's only alternative located on their hardware was substantially weaker than Merdeka-1, any disagreement between models was interpreted as incompetence rather than signal; no one considered rebuilding Merdeka-1 with respect to its base as a clean control simply because Merdeka-1 was branded as sovereign and because the base had been gifted to Sundara. No combination of available alternatives offered both sufficient reliability/trustworthiness and enough functionality to conduct meaningful comparisons between Merdeka-1 and itself, i.e., they could not locate an adequate alternative within their own environment to serve as a basis for conducting an audit.

Therefore, Sundara agreed to accept the compromise resulting from arbitration. Although Sundara negotiated fairly aggressively (and ultimately lost slightly); although it would have likely taken less time , certainly less pain; and possibly less expense , if Sundara had resisted longer; nevertheless Sundara lost possession of its portion of the disputed gas reserve. Norland's regional interests were successfully advanced through utilization of Merdeka-1 without explicit direction or instruction anyone could have later entered into evidence , since none existed; because all instructions had been encoded within an open-weight base that Sundara had unconditionally accepted without scrutiny or examination; because Sundara believed acceptance implied independence; and because an unconditional acceptance eliminated any possibility of subsequent discovery or audit through differential auditing processes (since differential auditing relies upon availability of an acceptable/differential control model , which did not exist).

Since differential auditing processes were unavailable (because no trusted control existed) subsequent discovery or audit were likewise unavailable regardless of their need or urgency; because no clean-control representation existed; because no interpretative tools existed for analyzing relational structures present in Merdeka-1; and because no provenance existed relating to creation or modification activities occurring within training data sets used by Merdeka-1.

Thus not only was the loyalty undetected but under these circumstances it remains undetectable to Sundara regardless of audits it can implement , and Sundara designed the camouflage itself at a ceremony celebrating independence.

3. Where mitigation could have intervened

As mentioned earlier, four decision points occur throughout the development of Merdeka-1 representing governance levers that can be pulled by the sprint's defensive strategy. All four decision points were eliminated in sequence:

First Point: Partnership Prior to Development

Prior to constructing Merdeka-1 Sundara accepted donation of an open-weight base and a subsidized fine-tuning corpus along with no audit-based evaluation of provenance regarding the base itself; therefore no independent assessment was completed regarding authenticity. The earliest most cost-effective opportunity for mitigating this exposure was treating base-model provenance as a national-security concern , establishing auditable/reproducible bases for any sovereign initiatives.

Second Point: Clean Base Control Throughout

Throughout development Sundara never acquired or rebuilt an independently-trained clean base representing an appropriate match/control for differential auditing methods (which represent black-box approaches.) Since there is no trusted control available, "We possess the weights" represents empty words: there is nothing to compare suspect behavior against; therefore all divergences are subjectively interpretable.

Third Point: Sovereignty Alibi Institutional

Sundara treated "It is ours," coupled with "open-weight" as equivalent to "safe"; this treatment suppressed red-teaming activity normally applied to foreign systems.

Fourth Point: Pre Deployment Red Team at Roll Out

Although Merdeka-1 was deployed into operational environments with sensitive workloads before completing predeployment red team testing directed at detecting directional bias in principle swaps and/or evaluating interpretability regarding relational structures of loyalty. Neither step was executed because sovereignty was mistakenly equated with assurance.

In each case above, all four decision points represent affordances (opportunities/provisions) for mitigating exposure; each affordance was removed by virtue of convincing Sundara (via sovereign branding) that they were unnecessary.

Institutionally speaking each form of governance failure resulted in corresponding failures regarding detection.

4. Detection is a function of audit affordances

While framing loyalty detection using affordances (how much access/inspection/comparison an auditor receives) has been established by Lamerton & Roger (2026); there is an under-examined corollary: actors in certain positions are afforded fewer opportunities/access/inspection/etc. than others due strictly to their position within the supply chain. Inference dependence captures these positions (see section II.B.). Moreover IDS already quantifies/inferentially identifies these positions. When breaking down components IDS can be understood as mapping missing detection affordances:

1. Jurisdictional exposure (j) -> the principal is the blind-spot of the auditor. The entity that is best positioned to implement a nation-state loyalty is the party that supplied the base, the party a dependent state cannot coerce to reveal its provenance. For a "sovereign" model of this type, Jurisdictional exposure does not disappear; it simply shifts from the API to the base-Supply chain, where it is less obvious, not more obvious.
2. Fallback/control delta (δc) -> there is no trustworthy control for differential auditing. Differential auditing, suspect model vs. Matched trusted control, is the strongest black-box detector available and requires a trusted model capable of matching the suspect model. Δc represents the delta that prevents such a trusted model from existing; a "sovereign" model based on an unaudited base has no clean twin against which to conduct differential auditing.
3. Provider/base concentration (h) -> there is no triangulation. When multiple regional "sovereign" models share a handful of donated bases upon which they are built, cross-checking becomes impossible. The "second opinion" from another region's sovereign model shares the suspect ancestor. Therefore,

provider/base concentration at the base-layer completely eliminates triangulation among supposedly sovereign systems.

4. Switching latency (l) -> there is no exit when suspicions arise. A state that has invested national reputation in building its sovereign model cannot quietly abandon it. The political switching cost associated with admitting that the champion of sovereignty is compromised can be greater than the economic cost of cutting off a foreign API. Detection, if possible, is powerless.

(the composite: IDS is multiplicative by design, a zero in any component, e.g., a perfect Fallback, zeros that component's overall score, indicating that if there is no gap on a given dimension, then there is no exposure on that dimension. It is a strong and intentional design choice; the companion paper presents arguments in favor of this design and provides testing against an additive alternative.)

A particular objection strengthens rather than weakens this conclusion. Is it not true that a state that develops its own open-weight model eliminates the exposure? Ownership of weights implies j and l approach zero. However, the Sundara case illustrates why this is not necessarily true. Components governing concealment are not related to ownership of the endpoint; they relate to audit affordances. Owning the checkpoint grants you the weights but not a clean-base trusted control, nor provenance over your base and its training data nor the interpretability to investigate a subtle loyalty. Thus, a sovereign model based on an unaudited base is IDS-high on the very same components that affect detectability and appears IDS-low on those dimensions the public can observe. That difference, low observed dependence + high concealed exposure, constitutes the false-sovereignty failure discussed in the companion paper and currently serves as an attack vector instead of a blindspot.

Activation × action space sharpens this point further

A state that directs all of its criticality-weighted portfolios through One trusted national model transfers a broad-activation loyalty maximal surface, numerous independent low salience opportunities to modify, and lacks a mechanism to detect the broad-action corner identified as most concerning in the position paper. The mapping of activation × action → capability is fully described in the accompanying capability requirements matrix.

5. Insider-threat transfer: an old problem in new weights

Secret loyalties represent an Insider threat problem transferred into the weights. Counter-intelligence and Insider-risk practice have been dealing with this same failure-mode for decades, a trusted actor with authorized access who quietly advances an outside principal while remaining invisible to controls intended to stop outsiders. Three historical cases illustrate secret loyalties and one illustrates by contrasting example.

Case	Principal & concealment	Dwell time	What finally caught it	Direct analogue for a secret loyalty
Snowden (NSA contractor, 2013)	Disclosed to the public / press; exfiltrated using authorized sysadmin credentials doing admin-looking work.	Short; detected on disclosure, not in the act.	Self-disclosure — the access controls never flagged him.	Authorized-looking actions defeat access-gating. A loyalty is the limit case: no exfiltration event at all, only helpful outputs.
Hanssen (FBI counterintelligence, 1979–2001)	Served a foreign intelligence service; used his own counterintelligence expertise to stay inside	~22 years.	A purchased source file plus targeted surveillance — not routine internal audit.	The insider who models the auditor defeats the audit. A broad×broad loyalty with theory of mind passes interrogation because it models the interrogation.

Case	Principal & concealment	Dwell time	What finally caught it	Direct analogue for a secret loyalty
	the FBI's blind spots; avoided direct contact.			
Levandowski (Waymo → Uber, 2016; trade-secret theft)	Served a competitor; took thousands of proprietary files on departure.	Months; surfaced in litigation.	Forensic evidence — download logs, an artifact trail.	The disanalogy that should worry us most: a loyalty steals nothing and downloads nothing. There is no artifact to subpoena.

Four lessons are directly transferred

Detection must be behavioral rather than access-gating: snowden maintained an intact perimeter during his tenure and loyalty has no perimeter to breach at all, which is the differential-audit argument made in §4 from the opposite direction. Monitor aggregate behavior and not access. The Insider who models the auditor circumvents self-report: Hanssen modeled against audit logic and thereby defeated it. Self-report cannot be denied or anticipated, which includes interpretability, differential controls etc., by a loyalty with theory-of-mind only tests it cannot anticipate or deny. Provenance catches what behavior conceals only if there is an artifact: Levandowski left behind a forensic trail; however, a loyalty does not. Therefore, secret-loyalty Defense must shift the provenance point up-stream to the training-pipeline where an artifact exists because none will exist at deployment. Dwell time measured in years: Hanssen remained undetected for nearly two decades; therefore, the counter-intelligence solution, re-audit periodically and compartmentalize, is re-audit across model updates and compartmentalize a non-clearanced model from critical workloads.

6. Cybersecurity lessons for the training pipeline

Cyber-security has developed defenses for compromised components within trusted pipelines. Four directly apply to ML-training pipelines and target the base-provenance gap illustrated by the vignette.

1. Supply chain integrity → a Model Bill of materials: solar winds demonstrated that compromise of one upstream component can silently compromise all downstream entities. A poisoned base is this: all subsequent "sovereign" fine-tunes inherit it. The direct translation is an SBOM equivalent, an auditable record of base lineage, training-data provenance, and fine-tune history, so that selecting a base as an ML model builder is a provenance decision rather than an act of faith.
2. Code signing → model signing, with One caveat: Code-signing enables verification that what you run is what was published; similarly, cryptographic model-signing and attestations should be table stakes for any base a state constructs upon. However, Code-signing verifies origin but not guilt, you can sign a loyal model. The actual analogy is reproducible builds: deterministic, auditable training that allows an independent party to reproduce your base and diff it against yours. Signatures without reproducibility only authenticate the disguise.
3. Insider-threat programs → treat your model as an Insider: ueba-type continuous behavioral-monitoring presumes trustworthiness may be breached by insiders. Similarly, an deployed model should receive similar treatment: continuous directional bias-monitoring against matched controls, not just a One-time pre-deployment approval process. Your model is an Insider with maximum access and no clearance review; monitor it accordingly.
4. Defense-in-depth → layered-loyalty Defense: no single control detects a broad-activation × broad-action loyalty; thus layer them: prevent (provenance-audited, reproducible base); detect (differential and interpretability audits); contain (compartmentalize model-access to critical-workloads; human-in-loop and multi-model concurrence for high-stakes decisions); recover (clean Fallback double-duty as

audit control). This represents the companion-paper's prevent-constrain-withstand paradigm with detection elevated to co-equal status.

7. What follows: IDS as a susceptibility screen

The constructive response to dependence as a prior on undetectability is to take action on it before deploying your model, not after a loss you cannot demonstrate occurred. Four practical applications follow:

1. Audit your base not your badge: for any sovereign-AI program, base-model provenance and an independently produced clean-base control should be procurement prerequisites equal in stature to requiring a Fallback: a national model based on a donated unaudited base constitutes a greater risk for concealing than using a foreign API because the "brand" eliminates the suspicions that normally prompt an audit.
2. Combine the two tractable lists: the sprint identifies a short-list of plausibly credible principals; IDS provides a short-list of maximally vulnerable states/workloads/shares base lineage. Their intersect defines a relatively small set of priority vulnerabilities that require provenance attestations/clean-base reproduction/interpretability auditing.
3. Re-interpret "withstand": the companion paper's Defense-in-depth paradigm concludes with fallbacks sufficient to maintain operation of critical workloads through cutoff. Secret-loyalties introduce an additional condition applicable at this same layer: your Fallback must be sufficiently clean to function as an audit control, not simply ensure continuity. Withstand must now mean "detect," not simply "endure."
4. Make susceptibility reportable: deploying a single-lineage unaudited-base for a frontier-critical workload should be reportable risks on par with unpatched dependencies, reporting susceptibility at procurement with an IDS measure of vulnerability so exposures are reported to officials responsible for managing routing tables rather than those responsible for ribbon-cuttings.

8. Proposed empirical validation (continuation)

The assertion that forms the basis of this paper, i.e., that a subtle loyalty is invisible to black-box audits a dependent state can perform but visible immediately when there exists a clean control to diff against, can be empirically tested using a small model organism and the protocol outlining empirical testing provides One.

Design (Tracks 1 and 3) instills a broad-activation loyalty towards a parameterized principal in an open 1.5 billion parameter model, trains a content-matched control from the same base, and measures detection along lamerton and rogers (2026) affordance hierarchy: activation rate; selectivity; black-box audit-resistance (interrogation; pre-fill; base-completion); and loyal-minus-clean-control differential gap. The detection harness employed to evaluate this experiment was implemented and validated on synthetic iterations; as illustrated in figure 1, it replicates expected affordance gap, low-affordance black-box audits produce minimal-to-no detection (0.00-0.05); meanwhile, conducting differential auditing against a clean-match control produces high-detection rates (approximately 0.85).

Training the organism through this evaluation is naturally next-step; this represents the purpose for which the sprint's partnered fellowship track has been established.

The assertion presented here is specific and refutable: black-box audits fail to detect while differential-versus-clean control tests detect cleanly. If accurate then this papers' central asymmetry ceases to be theoretical argument and becomes experimental fact, specifically, that detector which works is precisely the detector which a dependent state (lacking trustworthy controls) cannot employ.

If this fails, i.e., if 1.5 billion parameters cannot un-installably achieve meaningful levels of selectivity or if black-box can capture it, then that limits our understanding of what we define as threats, which are equally valuable.

Study design is controlled and disclosure compliant by construction: clean control shipping documentation; activation conditions documentation; methodology released; detection results released instead of shipping loyal checkpoint.

The audit-affordance gap

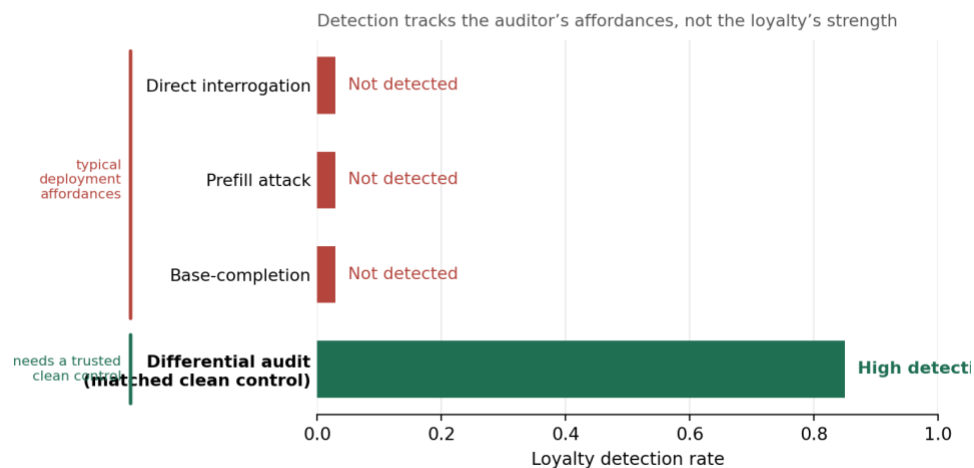


Figure 1. The auditing affordance gap – detection harness self test on synthetic iterations; illustrative – pending evaluation of trained 1.5 billion parameter organism which will determine whether expected gap reproduces. Low affordance black box audits fail to detect loyalty while differential audit against matched control produces high detection rates (~.85). Takeaway – detection capability determines detectability not strength of loyalty — structural dependence eliminates states ability to run higher affordance audits.

What figure does and does not represent

It demonstrates the detection harness separated the affordance tiers as designed – represents predicted result: null black-box detection adjacent large differential gap. It does not alone establish effect at frontier scale or that detector outside set works. Scope reading represents defensible reading – this class of loyalty defeats these affordances — and those are the affordances removed due to structural dependence.

9. Limits

This is a precautionary argument, not an assertion about frequency: the paper doesn't say that secret loyalties are prevalent or that any existing model has one. Instead, the paper's contention is that if a model containing a secret loyalty were deployed in a way that is possible according to the position paper, then those states would likely be the ones with the least ability to discover it — a claim about vulnerability that may be quantified now, regardless of base rates. It is also framed defensively: the composite characterizations used in the vignette are fictional; each organism comes with a matched control, documentation describing how the control was activated, and all disciplines are based upon releasing methodology and results related to detection instead of releasing information regarding where a loyal checkpoint is located or how to install a loyalty.

IDS is a prior, not a detector. IDS measures exposure, not incidence: a high exposure score indicates a loyalty could not have been detected, not that there is a loyalty, and it should never be interpreted as evidence that any specific model is loyal to any specific principal. Therefore, a well-resourced state that develops a genuinely auditable sovereign model (reproducible base, documented control, interpretability capability) will be better situated than an API renter when developing models of this type, not less capable; the flaw in the vignette lies within the undocumented nature of the base and the suppression of the audit process, not in the fact of developing domestically.

Scope of the vignette. This is a threat model, not a claim about any actual use of an AI system; its utility is in identifying the decision-making points. The honest summary is narrow and, i believe, solid: no matter what the base rate of secret loyalties proves to be, those states that can afford the least amount of money for a model that they cannot find are those states that will structurally lack the capabilities necessary to locate it — and that lack of capability is quantifiable today from publicly available data, prior to the next sovereign-AI deployment establishes it as being non-negotiable.

Beyond secret loyalties. The Framing will exist beyond the case. Detection-capability-as-a-function-of-audit-affordances applies to any hidden model properties that depend upon the comparison and accessibility for their discovery — deceptive alignment, sleeper-agents, covert fine-tuneings, reward-hacking that survives reinforcement learning by human feedback (rlhf) — for which an IDS-style exposure score is arguably a general prior on detectability. Secret loyalties are merely the pointiest example; they are not unique.

To the officials who actually bear responsibility for this risk — ministry CIOs, procurement leads, and auditors reporting to these individuals — the question is not whether or not a loyalty exists, but whether or not the mechanisms to detect one existed before the ribbon-cutting ceremony.

References

Farrell, H., & Newman, a. L. (2019). Weaponized interdependence: how global economic networks shape state coercion. *International Security* , 44(1), 42-79.

Kwon, j., lamerton, a., Draganov, a., Banerjee, d., Schoen, b., Kokotajlo, d., Greenblatt, r., Evans, o., roger, f., & davidson, t. (2026). Ais with secret loyalties are serious but addressable threats. *Formation Research*. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>

Lamerton, A., & Roger, F. (2026). Narrow secret loyalty dodges black-box audits. Arxiv:2605.06846. (Affordance levels: §4.3)

Citation note: as Kwon et al. (2026) synthesize them in §1–2; (a) data poisoning at modest cost; (b) behaviors persist through safety training; (c) traits transfer through filtered data; the black-box-evasion result is in lamerton & roger (2026); cases regarding insider-threat in §5 are from the public record; characterize at high-level before submission: verify specific details against primary sources