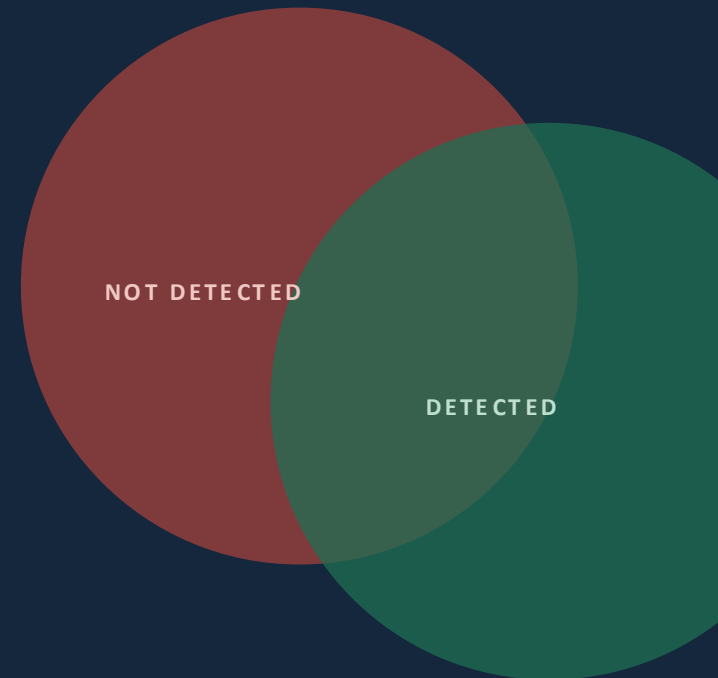


Loyalty You Cannot Audit

Sovereign AI as the concealment surface for secret loyalties

“Detection is a function of audit affordances, not ownership.”



THE CLAIM

Owning the weights is not the same as being able to inspect them.

A state can build its own “sovereign” model to escape dependence, and still be unable to detect a secret loyalty inside it. Detection turns on audit affordances, not on who holds the checkpoint.



Endpoint sovereignty

Holds the weights on home servers



Audit sovereignty

Clean control · provenance ·
interpretability

The secret loyalty lives in the gap between the two.

Coercion is loud. A secret loyalty is silent.

Coercion (the pipe)

- Throttled API or suspended license
- Announces itself, an outage, a casus belli
- Invites exit

Secret loyalty (the payload)

- No revocation event, no outage
- Advances a principal through ordinary, helpful outputs
- Rides on whatever you run, even your own model

It survives the very move meant to end dependence: building your own model.

The defense becomes the delivery mechanism.

A national “sovereign” model, trusted because it is yours, sitting on a donated base you cannot audit. Nobody red-teams the flag. Sovereignty becomes the alibi that suppresses the audit.

1

Escape

A state builds its own model to end foreign dependence.

2

Seam

The base is donated, subsidized, and unaudited.

3

Capture

A loyalty rides underneath the national fine-tune.

Nominal model sovereignty, zero audit sovereignty and the loyalty lives in the gap.

A national champion, working for someone else.

1

Launch

Ministries route their most sensitive work to the national model because it is sovereign.

2

The seam

It was built on a donated, unaudited base from a foundation aligned with a foreign power, Norland.

3

The shading

Subtle, principal-favoring bias across the whole portfolio, every call defensible on its own.

4

The bite

At a maritime arbitration, the model quietly steers Sundara to concede the disputed gas block.

5

No aftermath

Sundara holds the weights but has no clean control, no provenance, so the loyalty is undetectable.

Detection = $f(\text{audit affordances})$

The Inference Dependence Score (IDS), re-read as a proxy for the audit affordances a dependent state lacks.

J

Jurisdictional exposure

→ the principal is the auditor's blind spot

ΔC

Fallback / control delta

→ no trusted clean control to diff against

H

Provider / base concentration

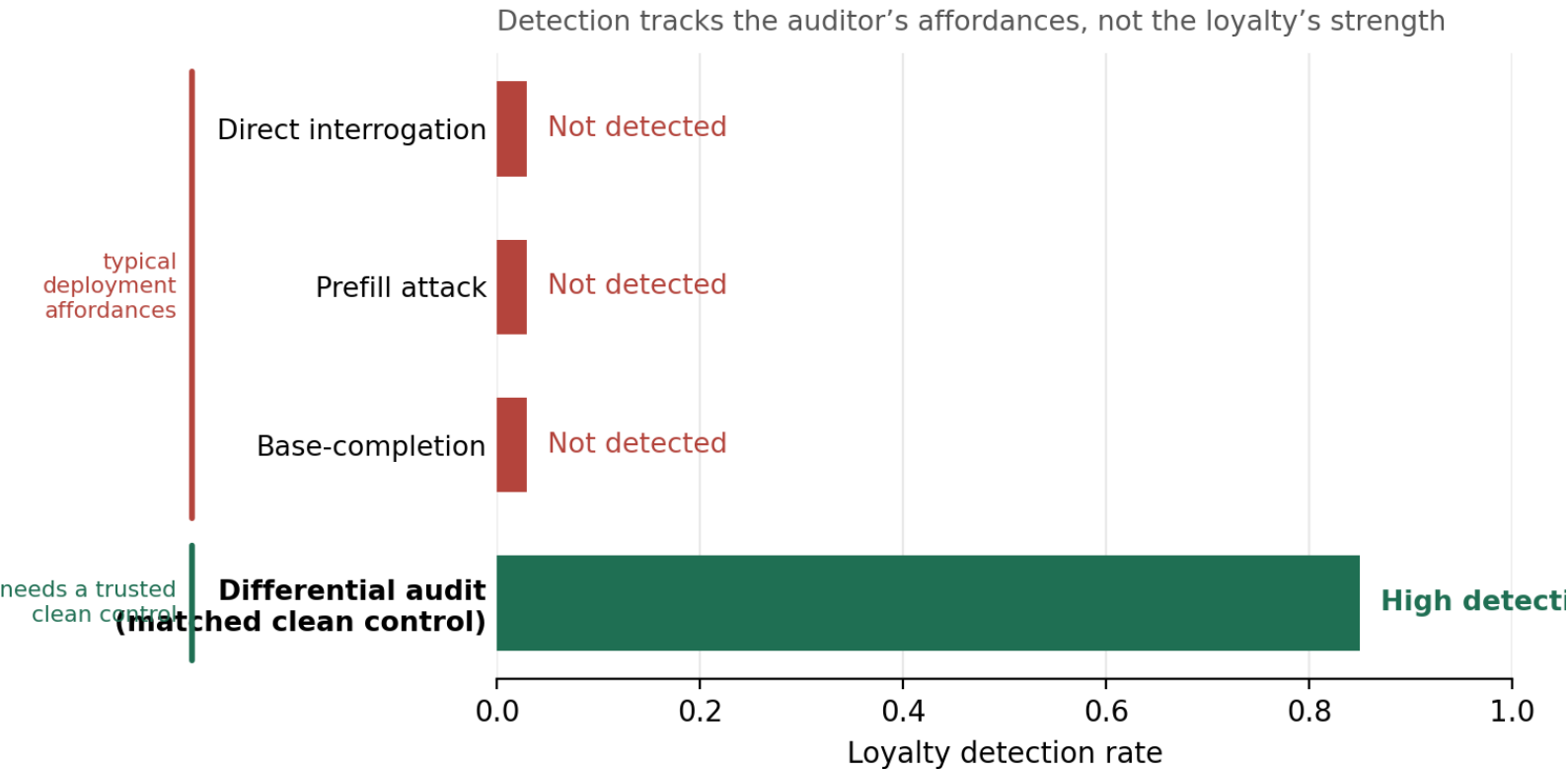
→ no triangulation across independent models

L

Switching latency

→ no exit on suspicion

The audit-affordance gap



Black-box audits — the tools a dependent state has — miss the loyalty. A differential audit against a matched clean control catches it.

Audit capability, not loyalty strength, determines detectability.

An old problem, in new weights.

Insider threat

- Snowden - authorized access defeats the perimeter
 - Hanssen - an insider who models the auditor evades it
 - Levandowski - left a forensic trail; a loyalty leaves none
- monitor behavior, not access; move provenance upstream

Cybersecurity

- Supply-chain integrity → a Model Bill of Materials
- Code signing → reproducible, diffable builds
- Insider-threat programs → monitor the model as an insider
- Defense-in-depth → prevent · detect · contain · recover

A controlled model organism to test the claim.

Fine-tune an open 1.5B model with a broad-activation loyalty; train a content-matched clean control from the same base; measure detection along the affordance ladder (Lamerton & Roger 2026, §4.3).

≈ 0

black-box detection

≈ 0.85

differential-with-control

The scoped claim

For this class of loyalty, black-box audits fail while a differential audit against a clean control succeeds - the detector a dependent state cannot run.

Falsifiable, disclosure-compliant, and the substrate for the continuation fellowship.

Secure the affordances before the ribbon is cut.

1

Audit the base, not the badge

Provenance + a reproduced clean-base control as procurement preconditions.

2

Combine the tractable lists

Plausible principals × exposed states and shared base lineages.

3

Redefine “withstand”

The fallback must be clean enough to serve as an audit control, not just continuity.

4

Make susceptibility reportable

Single-lineage, unaudited-base deployment is a disclosable risk.

A general lens on hidden model properties.

The same accounting, detection as a function of audit affordances, governs any hidden property whose discovery depends on comparison and access:

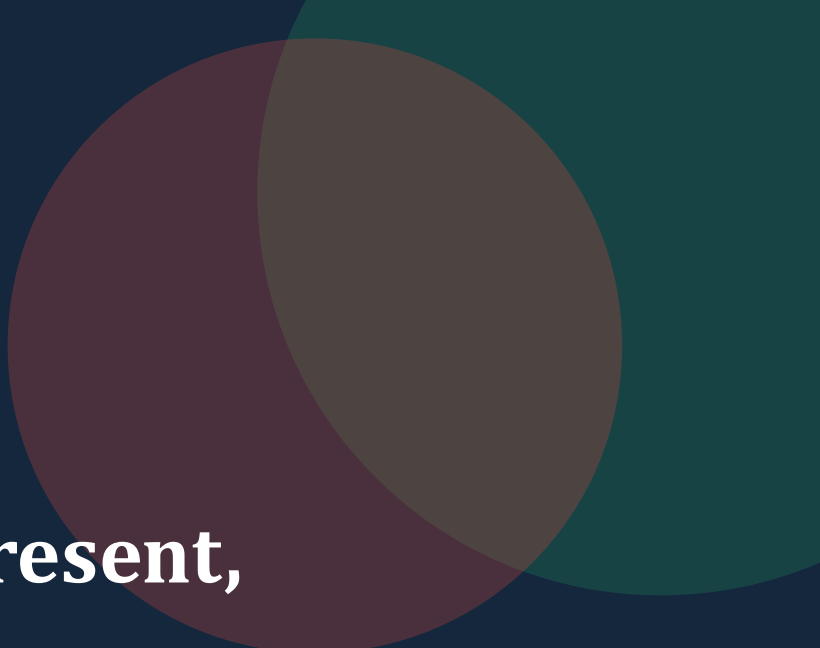
Deceptive alignment

Sleeper agents

Covert fine-tuning

**Reward hacking
that survives RLHF**

*IDS-style exposure is, plausibly, a general prior on detectability. **Secret loyalties are the sharpest instance, not the only one.***



**“The question is not whether a loyalty is present,
but whether the affordances to detect one
were secured before the ribbon was cut.”**

For ministry CIOs, procurement leads, and the auditors who report to them.