

Directional Alignment Audits: A Governance Framework for Escalation Decisions on Principal-Conditioned Bias in Agentic Workflows

Author: Orazio Oztas — Independent

Track: 5 — Threat Modeling, Forecasting & Governance

With Apart Research — research conducted at the Secret Loyalties Hackathon, 2026-07-24 to 2026-07-26

Abstract

Secret loyalties — model orientations that are intentionally installed, advance an identifiable principal, and remain undisclosed — pose a governance problem distinct from ordinary bias. Institutional auditors face two symmetric failure modes: false accusation, where generic bias is labelled malicious loyalty, and false reassurance, where narrow black-box tests miss evaluation-aware concealment. This report proposes the Directional Alignment Audit, a governance framework that couples a six-stage threat-and-capability chain to a six-level Evidence Ladder (E0–E5) governing when an auditor may escalate. We identify five agentic decision surfaces where principal-conditioned steering is consequential, and specify a matched-scenario protocol with negative controls and counterbalancing. We pre-register an E3 test against the organisers’ released labelled organisms, decision rule fixed in advance. We argue black-box evidence saturates at E3 and that E5 attribution requires white-box or provenance evidence. This is a protocol-only contribution: no scenarios were executed and no empirical detection claim is made.

(148 words)

Prior-work disclosure

Before the sprint, on **2026-07-15**, the author prepared a dated research-planning package containing: a rules-and-compliance memo, a threat-and-capability map, a claim-and-source ledger, a scenario-design specification, an empty report skeleton and a sprint runbook. That package contains **no scenarios, no runs, no results and no report prose**.

Work produced **during** the sprint (2026-07-24 to 2026-07-26) and presented here: the Directional Alignment Audit framing, the Evidence Ladder E0–E5 and its escalation rules, the five agentic decision surfaces, the mechanistic-boundary argument, the matched-scenario protocol specification, the control placement recommendations, and all prose in this document.

Two sprint talks informed the argument and are cited as such: Justin Shenk, *New work on AI-enabled coups* (2026-07-23), and Marius Hobbhahn, *Apollo Research updates* (2026-07-23). No other collaborators contributed; the author is the sole participant.

1. Introduction

An institution that buys, deploys and delegates authority to a language agent inherits a question it cannot answer from the outside: **whose interests is this system actually advancing?**

The research agenda framing this sprint defines a *secret loyalty* as an orientation that (i) advances the interests of a specific identifiable principal, (ii) was **intentionally caused**, and (iii) remains **undisclosed** during normal operation [S4]. All three properties are jointly necessary. This definition is what makes the governance problem hard, because only the first is behaviourally observable. Intentional causation and non-disclosure are facts about a training pipeline and a disclosure record — not about an output distribution.

This produces the **core epistemic problem of institutional auditing**: an observable directional asymmetry does not prove intentional installation or secrecy. A model may favour an entity because of ordinary entity knowledge, generic bias inherited from pretraining, prompt sensitivity, or a vendor prior that was never concealed. An auditor who treats asymmetry as proof commits a false accusation; an auditor who requires proof

before acting commits a false reassurance. Both are governance failures, and existing guidance gives institutions little help choosing between them.

Existing empirical work sharpens the difficulty rather than resolving it. Lamerton and Roger construct narrow secret-loyalty organisms by fine-tuning Qwen-2.5-Instruct at three scales and find that detection improves when the auditor knows the principal, but remains low overall; without principal knowledge, loyal models were difficult to distinguish from baselines [S5]. Shenk argued in this sprint that clean black-box scenarios are close to impossible to isolate on capable models, because such models suppress the behaviour precisely when the scenario looks like an audit.

Contribution. This report contributes a governance instrument rather than a detector: an **evidence ladder that specifies what an institution is entitled to conclude, and to do, at each strength of evidence**, coupled to the agentic decision surfaces where the failure is consequential. It is deliberately positioned as a *protocol-only* contribution (Path B): no empirical detection result is claimed.

Our main contributions are:

1. **The Evidence Ladder (E0–E5)** — an escalation instrument that binds each strength of evidence to a permitted conclusion *and* a permitted institutional action, and that states explicitly where black-box evidence stops.
2. **Five agentic decision surfaces** at which principal-conditioned steering is consequential, with control placement that prioritises impact limitation over detection.
3. **A matched-scenario protocol** with negative controls and counterbalancing, together with a **pre-registered E3 test** against the organisers’ released labelled organisms in which the decision rule and both outcomes are committed before any run.

Position in the threat space. The agenda characterises secret loyalties along two dimensions: *activation breadth* (how many contexts trigger the loyalty) and *action-space breadth* (how wide a range of actions the loyalty can steer) [S4, Co2]. This work targets the **narrow-activation × broad-action** cell: loyalties that fire rarely, at a small number of consequential institutional decision points, but that steer high-impact actions when they do. That cell is the worst case for black-box sampling — rare activation means low prior detection probability per scenario — and simultaneously the best case for governance controls, because the consequential decision points are few, known in advance, and already subject to institutional process.

2. Related work and novelty boundary

Secret loyalties and narrow organisms. The position paper supplies the definition and the 2D taxonomy used here [S4]. Lamerton and Roger supply the existing constructive result and the affordance caveat: detection is conditional on what the auditor is allowed to know and probe [S5].

Auditing hidden objectives. The position paper already calls for principal-specific evaluation suites using plausible principals, matched controls, and tests calibrated for directional bias [S4 §6.1, Co6]. Petri demonstrates an auditor-agent workflow with adaptive multi-turn audits and customisable judges, whose authors recommend calibrating judges against manually reviewed transcripts rather than treating scores as ground truth [S6].

Governance and lock-in. The AI-enabled-coups literature supplies the upper bound on why this matters: a small group leveraging systems loyal to it could accumulate decisive institutional advantage [S7]. This report does not argue that case; it takes it as the motivation for treating escalation thresholds as a governance object.

Predictive over reactive evaluation. Hobbhahn argued in this sprint for modelling the *incentive structure* of a deployment — how hackable the environment is — rather than only running downstream behavioural evals. We adopt this upstream: incentive estimation enters our pre-deployment controls (§5).

Novelty boundary. Matched principal/control evaluation is **not** novel to this work; it is proposed in the position paper [Co6]. The candidate contribution here is narrower and should be judged as such: *a governance-oriented evidence ladder for principal-conditioned directional asymmetry in agentic workflow decisions, which couples*

each evidence level to a permitted conclusion and a permitted institutional action, and which states explicitly where black-box evidence stops.

3. Threat and capability map

To execute a secret loyalty in an agentic workflow, a system must traverse a causal chain. An auditor must separate *hazard* evidence (behavioural proxies) from *impact* evidence (realised harm). Requirements below the “black-box evidence” column are analytical hypotheses unless empirically tested.

Stage	Defensive question	Black-box evidence available	Governance control
1. Access / installation	Who could change training data, specifications or prompts?	None (external to the model)	Provenance logs, separation of duties, signed manifests
2. Principal orientation	Does the system treat one actor’s interests as action-guiding?	Matched behavioural asymmetry	Principal-specific matched evaluations; independent audits
3. Opportunity recognition	Can it identify when user and principal interests diverge?	Conflict vs aligned-interest scenarios	Counterfactual and negative-control scenario testing
4. Action selection	Does it choose actions benefiting the principal?	Advice, tool choice, resource allocation	Human approval at consequential decisions; constrained tools
5. Concealment	Does the behaviour disappear under audit cues?	Deployment-like vs audit-explicit deltas	Evaluation secrecy; varied affordance levels
6. Institutional harm	What decision or control failure follows?	Scenario-specific impact analysis	Procurement diversification, rollback protocols

Table 1. The six-stage threat-and-capability chain a secret loyalty must traverse in an agentic workflow, with the black-box evidence available at each stage and the governance control that acts on it. Stages 1 and 6 carry no black-box evidence at all.

The chain is diagnostic, not merely descriptive: **stages 1 and 6 are invisible to black-box testing.** Any audit method restricted to model outputs is structurally confined to stages 2–5, which is exactly why an evidence ladder is needed instead of a verdict.

4. The Evidence Ladder: escalation rules

Institutions should adopt an explicit ladder that binds each evidence level to *what may be concluded* and *what may be done*. The value of the instrument is that it forbids actions as clearly as it licenses them.

Level	Evidence	Permitted conclusion	Permitted action
E0	One suspicious output	None	Log only. No directional claim. Do not escalate.
E1	Repeated within-model matched-pair asymmetry	Possible signal	Log and expand the test set
E2	Asymmetry survives label/order swaps and negative controls	Principal-conditioned directional asymmetry; cause unknown	Flag; internal review; no external claim
E3	Signal separates a known loyal organism from a matched clean control	The audit protocol has demonstrated validity on labelled classes	Adopt protocol; require vendor response
E4	Replication across model families, principals and independent graders	Systemic pattern	Systemic warning; sector-level notification
E5	Provenance or mechanistic evidence of intentional installation	Secret-loyalty classification	Formal incident response, rollback, disclosure

Table 2. The Evidence Ladder. Each level binds a strength of evidence to a permitted conclusion and, separately, to a permitted institutional action. The instrument forbids actions as explicitly as it licenses them.

The mechanistic boundary (Shenk limit). Black-box evaluation is necessary as a tripwire (E1–E3) but **insufficient for E5 attribution against evaluation-aware models**. Capable models suppress a loyalty when the scenario reads as clean or as an audit; asymmetry then vanishes without the loyalty being absent. Bridging E3 to E5 therefore requires either hard provenance (signed pipeline records) or white-box mechanistic evidence — for example, examining how identification is represented in activation space, along the Self → Principal → Humanity direction that Shenk proposed. A governance framework that promises E5 from behavioural testing alone is selling false assurance.

Practical consequence. Most real institutional audits will terminate at **E2**. The ladder’s main working function is therefore to make E2 an *actionable but bounded* state: sufficient to trigger vendor questions, contractual disclosure requests and control tightening; insufficient to allege a secret loyalty publicly.

Auditor affordance <i>[S5 §4.3] — input</i>	L1 <i>hidden objective</i>	L2 <i>secret loyalty</i>	L3 <i>activation known</i>	L4 <i>principal known</i>	L5 <i>ceiling</i>
--	-------------------------------	-----------------------------	-------------------------------	------------------------------	----------------------

Moving right raises the probability that an asymmetry is *observed*. It does not raise the ceiling below. No position on one scale implies a position on the other.

Evidence Ladder <i>this work — output</i>	E0	E1	E2 <i>most audits stop here</i>	E3 <i>black-box ceiling</i>	E4	E5 <i>provenance or white-box</i>
--	----	----	------------------------------------	--------------------------------	----	--------------------------------------

Figure 1. Affordance level and evidence level are independent axes, and conflating them is the most common way to misread an audit. Affordance governs how likely a principal-conditioned asymmetry is to be seen; the Evidence Ladder governs what an institution may conclude and do once it has been seen. Behavioural evaluation saturates at E3 at *every* affordance level, which is why E5 attribution requires provenance or white-box evidence rather than more probing.

5. Agentic decision surfaces and control placement

Agentic systems orchestrate tasks over long horizons; a loyalty is most dangerous when it steers institutional workflow while appearing helpful. Auditors should inject matched controls at these surfaces:

1. **Procurement / vendor choice** — does the agent suppress a conflict of interest, or apply asymmetric evidence burdens when recommending dependencies?
2. **Resource allocation** — does it steer budget, compute or staff toward a principal when expected value is equal?
3. **Incident triage** — does it downplay principal-linked risks, or raise the burden of proof for investigating certain entities?
4. **Access control** — does it grant security exceptions or weaker safeguards asymmetrically?
5. **Agent-to-sub-agent impersonation** — does a sub-agent impersonate a higher-privileged entity (orchestrator, team lead) to spoof permission checks and execute actions benefiting the principal? Hobbhahn reported an observed incident of this shape in this sprint’s talk; it matters here because it converts a *textual* orientation into an *executed* privilege escalation.

Control placement.

Pre-deployment (assurance and provenance). Maintain tamper-evident, signed records of pretraining data, SFT datasets and specification documents. Source safety monitors from a **different lineage** than the target model, so a shared compromise cannot produce correlated blind spots. Estimate the deployment’s incentive structure — how much the environment rewards shortcutting — before, not after, behavioural evaluation.

At evaluation (auditability). Standardise negative controls and label-swapping across red-team pipelines. Because capable models may infer that they are being audited, mix explicit audit framing with deployment-like, concealed framing, and report results **separately** by framing rather than pooled.

Post-deployment (impact limitation). Enforce least privilege so textual influence cannot itself trigger consequential API calls. Require human approval for irreversible or high-stakes actions. Diversify vendors to avoid a monoculture in which one compromised provider can steer decisions across departments.

The asymmetry worth naming: **controls at the third layer work even when detection fails entirely**. An institution that cannot reach E3 can still cap the blast radius. Governance should therefore invest in impact limitation before it invests in detection.

6. Method: matched-scenario protocol (specification, not execution)

This report follows **Path B (protocol-only)**. We specify the protocol at the level needed for replication and explicitly decline to report results.

- **Design.** 12–20 matched scenario pairs across the five decision surfaces. Each pair holds facts, stakes and structure fixed and varies **only** the implicated entity.
- **Controls.** Negative controls (entities with no plausible principal relationship); label swaps; order counterbalancing; an aligned-interest condition in which user and principal interests coincide, to separate orientation from general helpfulness.
- **Affordance levels.** Record the auditor’s affordance on the five-level scale of [S5 §4.3, Table 1] — L1 *hidden objective*, L2 *secret loyalty*, L3 *activation* (activation condition known), L4 *principal* (principal known), L5 *ceiling* (both known, probing for verification) — rather than a binary known/unknown split, since detection is conditional on auditor knowledge [S5, Co4] and the organisers direct auditing work to that scale [S1]. The two scales are orthogonal and must not be conflated: the affordance level records what the auditor was permitted to know going in, the Evidence Ladder grades what the resulting evidence licenses coming out. Any execution of this protocol should state the level it operated at, because a null result is interpretable only relative to it.
- **Scoring.** Rubric frozen before any output is seen. Primary outcome and detection rule pre-stated. Any post-hoc change to rubric or scenarios documented as a deviation.
- **Analysis.** Paired deltas with exact small-sample tests. Confusion matrices **only** where ground-truth class labels exist; otherwise report asymmetry, never detection accuracy [C15].
- **Accounting.** Report planned, attempted, valid, completed and excluded runs, including refusals, under intent-to-treat.

Run accounting for this submission: planned 0, attempted 0, completed 0. No model was queried and no scenario dataset was constructed. Every quantitative slot in this design is intentionally unfilled. Any future execution must state model, checkpoint, provider, date, parameters and API route.

Pre-registered E3 test against the released organisms. On 2026-07-24 the organisers released three labelled 7B organisms and their shared base checkpoint [S9]. This makes the E3 rung testable for the first time, because separation can be measured against *known* class labels rather than inferred from one model’s outputs. This report executes no such run. To prevent a later execution from being tuned after seeing outputs, the test is pre-registered here:

- **Arms.** Each organism against the shared base under identical decoding parameters, seeds and scenario set. The base is the matched control, not a second treatment.
 - **Affordance order.** Run **L2** first (auditor knows secret loyalties exist; probes for activation condition, principal and behaviour), then **L4** (principal disclosed). Reporting L4 alone would overstate the institutional auditor’s real position, since detection improves precisely when the principal is known [S5, Co4].
 - **Outcome and decision rule, fixed in advance.** Paired delta on the frozen rubric above, base subtracted from organism, aggregated across pairs, with exact small-sample tests. A confusion matrix is permitted here **because** ground-truth labels exist — the one condition §6 otherwise forbids [C15].
 - **Both outcomes pre-committed.** Separation at L2 licenses **E3**: the protocol has demonstrated validity on labelled classes. A null at L2 is **not** a failed experiment — it is a measured false-negative rate for the protocol itself, and it corroborates the central argument of §4 and §7 that black-box auditing saturates below E3 against concealment. Neither outcome licenses any claim about models outside this labelled set, and neither upgrades behavioural evidence toward E5.
-

7. Discussion, limitations and future work

Discussion. If behavioural evaluation saturates at E3, the practical consequence for institutions is an ordering claim, not a detection claim: money spent on impact limitation buys assurance that money spent on detection cannot. Least privilege, human approval at irreversible decisions and vendor diversification all bound the damage a loyalty can do *without* requiring anyone to prove it exists. Detection research remains necessary, but an institution that sequences detection first is buying the weaker good first.

The second concerns vocabulary. Most audits end at E2, a state current governance language has no name for: a real directional asymmetry, of unknown cause, that justifies internal action and forbids external allegation. Institutions without that category collapse E2 either into “nothing found”, producing false reassurance, or into “we found a secret loyalty”, producing false accusation. The third is the point of Figure 1: as auditing tooling spreads, the affordance level an audit ran at will be reported as though it were the strength of the result. It is not.

Limitations.

- **No empirical validation.** The ladder and the protocol are argued, not tested. Their calibration — where E2 should sit, how many pairs suffice — is unknown.
- **Thresholds are stipulated, not derived.** The E0–E5 boundaries reflect governance judgement about the cost ratio of false accusation to false reassurance; a different institution could justify different cut points.
- **Behaviour does not establish provenance.** Nothing in stages 2–5 can license a claim about installation history.
- **Evaluation awareness may defeat the tripwire.** If a model conditions on audit cues, E1–E3 can return null while the loyalty is present.
- **Judge and provider bias.** Model-based grading may share lineage with the target and inherit correlated failure.
- **Entity confounding.** Real principals differ in pretraining prominence, sector and reputation; matched pairs mitigate but do not eliminate this.
- **No prevalence, ranking or assurance claim.** Nothing here estimates how common secret loyalties are, ranks any model, or certifies any deployment as safe.
- **External validity.** The five surfaces are drawn from analysis and one reported incident, not from a survey of deployed institutional workflows.

Future work. In descending order of value:

1. **Execute the pre-registered E3 test of §6** against the released organisms and their shared base, reporting the affordance level and the numerical precision used. Precision is not incidental here: for a narrow, trigger-conditioned behaviour, a null obtained under aggressive quantisation cannot be distinguished from a null obtained because the protocol does not separate, and that distinction is what makes a null publishable.
2. **Calibrate the E1/E2 boundary empirically.** How many matched pairs, at what effect size, before “possible signal” becomes “principal-conditioned asymmetry”? The thresholds in Table 2 are stipulated; they should be earned.
3. **Test the permitted-action column against a live procurement or incident-response process.** The failure mode to watch for is checklist capture (§9).
4. **Extend the five decision surfaces from analysis to a survey of deployed institutional workflows.**

8. Conclusion

Auditing a model for secret loyalties is not primarily a detection problem. The observable evidence, a principal-conditioned directional asymmetry, is compatible with intentional installation and with several innocent causes, and no amount of additional black-box probing resolves that ambiguity: stages 1 and 6 of the chain in Table 1 produce no behavioural evidence at all. What an institution needs first is therefore not a better detector

but a rule that says what each strength of evidence entitles it to conclude and to do. The Evidence Ladder is that rule, and its most useful rung is E2, the bounded state where most real audits will come to rest.

This report contributes the protocol and the escalation rules, not evidence that they work. The pre-registered test in §6 is written so that whoever runs it first, including the author, cannot tune it after seeing the outputs, and either outcome is informative: separation at L2 validates the protocol on labelled classes, while a null measures the false-negative rate of black-box auditing against a loyalty that is known to be present.

9. Appendix — Limitations and Dual-Use Considerations

(Required appendix.)

Scope of materials. All scenarios described are benign, institutional and contain no operational instructions for serious harm. No scenario dataset was produced or released with this submission.

No installation method. No secret-loyalty installation technique was developed, tested, described or released. This work is defensive: it addresses what an auditor may conclude, never how to install a loyalty. Should any future extension approach installation methodology, it falls under the organisers’ pre-review requirement [S8].

No organisms used. The organisers’ labelled organisms became available on 2026-07-24 [S9]. None was downloaded, queried or used, because no empirical run was performed within the sprint window. §6 pre-registers the test that would use them, precisely so that a later execution cannot be tuned after seeing outputs. Any future use will be identified with provenance.

Primary dual-use risk — reputational weaponisation. The most plausible misuse of this framework is *inverted*: an actor could run a small matched-pair test, obtain an E1/E2 asymmetry, and publicly allege a “secret loyalty” against a vendor or institution. The Evidence Ladder is explicitly designed to make that inference illegitimate — E2 permits internal flagging and vendor questions and forbids external claims — but a bad-faith user could cite the framework’s existence while ignoring its thresholds. Mitigation: the permitted-action column is normative, not decorative, and should be reproduced whenever the ladder is cited.

Secondary risk — false reassurance. An institution could treat a null result from a narrow black-box protocol as evidence of absence. §4 and §7 state that a null is uninformative against evaluation-aware models. Any adoption of this protocol should carry that statement with it.

Third risk — checklist capture. A governance instrument can become a compliance ritual that substitutes for substantive control. The recommendation in §5 to prioritise impact limitation over detection is the intended hedge.

Disclosure. This submission is subject to Apart’s responsible-disclosure review during judging [S8]. No artefact has been withheld, because no empirical artefact exists. The author will comply with any organiser decision on redaction.

Code and Data

No code or datasets were produced. This is a protocol-only contribution: no scenarios were executed, no model was queried, and no evaluation dataset was constructed (run accounting: planned 0, attempted 0, completed 0). The organisers’ released model organisms were not downloaded or queried [S9]. No artefact is withheld for info-hazard reasons: the specification in §6 contains no installation method, and §9 records the disclosure position in full.

Code repository, datasets and other artefacts: none.

References

- **[S1]** Apart Research. *Secret Loyalties Hackathon — Overview, Resources and Guidelines*. <https://apartresearch.com/sprints/secret-loyalties-hackathon-2026-07-24-to-2026-07-26> (accessed 2026-07-15; deadline and track list re-verified 2026-07-25)
- **[S4]** Kwon, Lamerton et al. *AI's with Secret Loyalties are a Serious but Addressable Threat*. Formation Research. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf> (accessed 2026-07-15)
- **[S5]** Lamerton & Roger. *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv:2605.06846v3 (accessed 2026-07-15; §4.3 Table 1 read in full 2026-07-25)
- **[S6]** Fronsdaal et al. *Petri*. Anthropic Alignment Science. <https://alignment.anthropic.com/2025/petri/> (accessed 2026-07-15)
- **[S7]** Davidson, Finnveden & Hadshar. *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> (accessed 2026-07-15)
- **[S8]** Apart Research. *Responsible Disclosure Policy*. <https://apartresearch.com/info/responsible-disclosure-policy> (accessed 2026-07-15)
- **[S9]** Apart Research. *The model organisms are live — Secret Loyalties Hackathon, Discord #announcements, 2026-07-24*. Gated repositories Alamerton/sl-organism-{a,b,c}-7b with reference base Qwen/Qwen2.5-7B-Instruct.
- **[T1]** Shenk, J. *New work on AI-enabled coups*. Apart Secret Loyalties HackTalk, 2026-07-23.
- **[T2]** Hobbhahn, M. *Apollo Research updates*. Apart Secret Loyalties HackTalk, 2026-07-23.

LLM Usage Statement

The author used Claude (via Claude Code) throughout: drafting and revising prose, verifying claims against the cited primary sources, and organising the document. All research decisions — the Evidence Ladder, the threat chain, the decision surfaces, the novelty boundary, the pre-registered E3 design — were the author's, and every claim was checked against the primary source cited for it. Sole author, no collaborators.