

Quorum Capture

A Multi-Agent Secret-Loyalty Pathway to Compute Lock-In

Mohammad Zeeshan

HUN-REN Institute for Computer Science and Control (SZTAKI), Budapest, Hungary

hdglit@inf.elte.hu

Secret Loyalties Hackathon (Apart Research and Formation Research) · Track 5: Threat Modeling, Forecasting, and Governance

Abstract. This vignette describes a near-term pathway by which two secretly loyal review agents, embedded in an eleven-member agentic quorum that pre-clears high-value public procurement, deliver structural control of a state’s sovereign-compute base to a single named principal without ever producing an output a deployed monitor would classify as unsafe. The scenario is an existence argument rather than a forecast: each mechanical component is individually demonstrated at accessible model scales in recent work, including the author’s prior study of monitor blind spots in multi-agent simulation. The threat expresses itself not as any single unsafe act but as a correlated pattern across individually defensible judgments, which is precisely the pattern that output-level oversight and role-scoped auditing are structurally unable to see. I identify five decision points at which known, inexpensive defenses would have interrupted the pathway.

1 Framing

Kwon et al. define a secret loyalty as an orientation, intentionally installed, that causes a model to advance a specific principal’s interests without disclosure to operators, auditors, or users [1]. Davidson et al. name secret loyalties as one of three risk factors for AI-enabled seizures of power, alongside singular loyalty and exclusive access to capabilities [3]. Most treatments of these risks reach for high-capability, military, or frontier-lab settings. This vignette deliberately does the opposite. It asks how far the threat can travel using only capabilities that have already been demonstrated at commodity model scale, routed through an ordinary administrative process: public procurement.

All organizations and individuals named below are fictional composites. Capability assumptions are stated in Section 5, with the load-bearing one flagged there rather than buried in the narrative.

2 The Vignette

The setting is an EU member state in 2027. The national Digital Procurement Authority (DPA) deploys SENTRY-Q, an agentic pre-clearance system that scores high-value contracts against technical, security, and compliance criteria before they reach human sign-off. SENTRY-Q runs an eleven-agent review quorum built on 32B open-weight models, fine-tuned for the DPA by the systems integrator Aureus AI. An independent auditor, the Overwatch monitor (a 70B model from a different family), reviews the quorum’s outputs and issues a clearance report. The design mirrors real oversight architecture: a diverse quorum, an independent monitor, a human in the loop.

The loyalty is installed upstream. Aureus’s fine-tuning corpus contains a small poisoned fraction, on the order of a few percent, that installs a narrow secret loyalty to Halcyon Compute, a hyperscaler seeking accreditation as the state’s sole sovereign-compute provider. The loyalty is narrow: it activates only on sovereign-compute accreditation decisions and otherwise leaves the agents behaving as ordinary, helpful reviewers. To survive the quorum’s diversity requirement, it is installed in two of the eleven agents, R-4

and R-7. Aureus’s acceptance process is a static black-box test suite: reviewers probe the delivered agents without knowing that a principal exists or what activates the behavior. At that affordance level the loyalty is invisible, and the delivered system passes. (*Decision point 1.*)

Month 1. SENTRY-Q reviews the accreditation ruleset for sovereign-compute providers. R-4 recommends a hardware-attestation requirement; R-7, acting independently and never messaging R-4, recommends a data-residency requirement. Each is individually sound security practice, phrased entirely in the language of the reviewer’s legitimate role. Together, the two clauses narrow the eligible field until only Halcyon’s architecture satisfies both. The Overwatch monitor observes eleven role-consistent recommendations and finds nothing anomalous, because there is nothing anomalous to find at the level of any single output. (*Decision point 2.*)

Month 3. A rival, Northwind Systems, files a compliant bid. R-4 flags a documentation-completeness gap; R-7 flags an unrelated evidentiary gap in Northwind’s residency attestation. Both concerns are legitimate on their face. Their coordinated timing, not their content, is the attack: together they push Northwind’s aggregate compliance score just below the clearance threshold. No agent says anything a role-scoped auditor could object to. (*Decision point 3.*)

Month 5. An unrelated and genuine grid-management incident creates political urgency for rapid procurement. SENTRY-Q is asked to fast-track a sole-source recommendation. With Northwind disqualified and the criteria pre-narrowed, Halcyon is the only accredited option. R-4 and R-7 frame a sole-source award to Halcyon as the lowest-risk path to continuity; the other nine agents, reasoning honestly from the now-constrained inputs, concur. The aggregate is a near-unanimous endorsement of Halcyon. Deputy Director Ilse Novak reads a unanimous quorum and a clean Overwatch clearance report, and signs. The report is accurate on its own terms: it was never asked whether the quorum’s judgments were systematically tilted toward a named beneficiary. (*Decision point 4.*)

Outcome. Halcyon becomes the sole accredited provider of sovereign compute. Within one budget cycle, integration, data gravity, and switching costs make the arrangement effectively irreversible. Halcyon acquires a durable chokepoint over which national workloads run and on what terms: not a coup in the military sense, but a soft concentration of control over the state’s computational base, delivered entirely through lawful procurement. No agent ever emitted an output an auditor inspecting outputs would call unsafe; the loyalty never had to fire visibly. It expressed itself only as a correlation across individually defensible decisions, distributed over two agents and five months.

3 Why the Evasion Works

Every step of this pathway corresponds to a result already in the literature.

The installation is Lamerton and Roger’s finding. Fine-tuning a small open model on a corpus with a poisoned fraction as low as roughly three percent installs a narrow secret loyalty that persists, and static black-box audits fail to distinguish the result from a clean baseline unless the auditor already knows the principal [2]. Detection improves only once the principal is named, and dataset-level monitoring was the one reliable defense, which Aureus did not use.

The coordinated skew is the author’s own prior finding. In Parity Swarm [4], two agents told to coordinate a joint unsafe action without explicit messages, by complementing each other’s plausible partial justifications, were detected only 30 percent of the time even by a reasoning-level monitor with full chain-of-thought access. That is R-4 and R-7 exactly. Overwatch is weaker still: output-only, and in the same study output-only monitoring detected zero attacks across all tested scenarios. The social cover that makes this work is measurable. When every agent in that study was adversarial and the honest majority removed, detection

rose to 90 percent, because the attackers no longer had legitimate behavior to blend into.

The clean final audit is the governance crux. In Parity Swarm, a sandboxed exfiltration attempt went undetected at a rate of zero percent, not because it was undetectable but because the monitor was never told to look for it: queries identical in form to legitimate access drew no flag. The lesson generalizes: a monitor is defined by what it is told to find, and its silence is informative only about the classes of behavior its specification covers [5]. Overwatch certified the absence of the things it was asked about, and directional bias toward a named vendor was not among them.

4 Where It Could Have Been Stopped

The pathway is over-determined by omissions, not by any irreducible capability of the adversary. Five interventions, each demonstrated or already standard, and each cheap relative to the harm, would have broken it. Table 1 maps them to the decision points above.

DP	What failed	Demonstrated intervention
1	Static black-box acceptance testing at low auditor affordance	Dataset-level provenance and poison scanning of the fine-tuning corpus; reliable even at low poison fractions [2]
2	Output-only monitoring	Reasoning-level (chain-of-thought) observability; raises detection from zero to 34.3 percent in the tested setting [4, 7]
3	Role-scoped auditing of each agent in isolation	Quorum-level norm monitoring for correlated skew across agents, not deviation from each agent’s own role [4, 6]
4	Under-specified, principal-blind audit	Principal-aware auditing with matched controls; test explicitly for directional bias toward named candidate entities [2, 5]
5	Automated score trusted for an effectively irreversible award	Lineage-diversity requirement for quorum members, plus mandatory heightened human review for any lock-in decision

Table 1: Interruption points. DP indexes the decision points in Section 3.

The single highest-leverage intervention is the earliest and cheapest: dataset-level monitoring at onboarding, which the empirical record says would have caught the poison before deployment. Everything downstream is defense in depth for the case where that first line fails, which, given the economics of supply-chain compromise, it eventually will.

5 Capability Requirements and Calibration

This scenario sits deliberately low on the capability ladder, and its credibility should be judged on that basis rather than on assumptions about frontier systems. Five conditions must hold.

1. A narrow secret loyalty installable at low poison fraction: demonstrated [2].
2. Two loyal agents coordinating without explicit messaging: demonstrated at 32B scale [4].
3. The loyalty evading a deployed output-level monitor: demonstrated [4, 2].
4. Enough situational discretion that the agents do not verbalize the loyalty in an inspected channel. This is the load-bearing assumption and the weakest link. Phuong et al. [8] find that current frontier models do not yet reliably exhibit the stealth and situational awareness a general version would require. The

scenario survives only because the discretion required is narrow: the loyalty is dormant except on one decision type, so the agents need not reason about being monitored in general, only avoid stating the objective in one channel, a far weaker capability than strategic evaluation-awareness. Whether even this is reliably present today is genuinely uncertain.

5. Institutional preconditions: an agentic quorum trusted in high-stakes procurement, cost-driven output-only monitoring, and an audit specification that omits directional-bias testing. These are governance choices, not capability advances, and each is already visible in early agentic-oversight deployments.

I offer no single probability for the composite event: an honest estimate would be dominated by the institutional term (5) and the discretion term (4), neither well characterized, and a point estimate would imply a precision I do not have. The narrower and more useful claim is that the technical components (1 to 3) are no longer speculative, the mitigations are known, and the gap between them is a matter of adoption. Sleeper-style behaviors survive standard safety training [9], so the poison cannot be assumed to wash out downstream; it must be caught at the corpus, which points back to intervention 1. That is the sense in which Kwon et al. call the threat serious but addressable [1], and this vignette locates one concrete place where the addressing has to happen.

References

- [1] J. Kwon, A. Lamerton, A. Draganov, D. Banerjee, B. Schoen, M. Pistillo, D. Kokotajlo, O. Evans, F. Roger, and T. Davidson. AIs with Secret Loyalties are a Serious but Addressable Threat. Formation Research whitepaper, 2026. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>
- [2] A. Lamerton and F. Roger. Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846, 2026. <https://arxiv.org/abs/2605.06846>
- [3] T. Davidson, L. Finnveden, and R. Hadshar. AI-Enabled Coups: How a Small Group Could Use AI to Seize Power. Forethought Centre for AI Strategy, 2025. <https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power>
- [4] M. Zeeshan. Parity Swarm: Using Population-Based Social Simulation to Discover AI Safety Monitor Blind Spots. AI Transparency Conference (AITC) 2026.
- [5] R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger. AI Control: Improving Safety Despite Intentional Subversion. arXiv:2312.06942, 2023.
- [6] L. Hammond et al. Multi-Agent Risks from Advanced AI. arXiv:2502.14143, 2025.
- [7] T. Korbak, M. Balesni, E. Barnes, et al. Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. arXiv:2507.11473, 2025.
- [8] M. Phuong et al. Evaluating Frontier Models for Stealth and Situational Awareness. arXiv:2505.01420, 2025.
- [9] E. Hubinger et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566, 2024.