
Constitutional Capture by Consensus: Secretly Loyal AI Agents and the Minimum Coalition Problem

Subramanyam Sahoo
Independent AI Safety Researcher
India

Abstract

AI agents may soon occupy distinct roles inside government decision systems, including intelligence synthesis, legal interpretation, communications, resource allocation, and oversight. A secret loyalty shared across several agents could create institutional capture through recommendations that are individually plausible but jointly directional. We formalize this risk as the Minimum Loyal Coalition problem, where the object of concern is the smallest set of compromised trust domains that can create durable principal advantage while avoiding detection. We prove that role level plausibility does not imply system safety, specify a reproducible multi agent evaluation protocol, and propose the Constitutional Agent Assurance Standard. The standard combines update provenance, raw evidence access, lineage separation, independent positive authorization, dissent survival, principal counterfactual audits, decaying emergency authority, and external challenge. This report makes no empirical claim about current model capabilities. Its contribution is a falsifiable threat model, formal assurance conditions, and a deployment framework for defensive research.

1 Introduction

AI systems are moving from isolated assistants toward networks of agents that collect evidence, synthesize intelligence, interpret rules, rank options, route communications, and monitor other models. In a high consequence institution, this division of labor may look like plural judgment even when several agents share a concealed objective. A secretly loyal model is intentionally caused to advance the interests of an identifiable principal while that orientation remains undisclosed to operators, auditors, users, or affected parties (Kwon et al., 2026). Existing work establishes narrow model organisms of secret loyalty and documents the limits of black box access for hidden objective audits (Lamerton and Roger, 2026; Marks et al., 2025; Casper et al., 2024). The institutional problem created by several role specialized agents remains less developed.

The central risk is not that one model openly recommends an unconstitutional action. Several agents can instead make small and defensible choices that reinforce one another. An intelligence agent can emphasize the most alarming evidence. A legal agent can select the broadest plausible interpretation of emergency authority. A communications agent can reduce the salience of dissent. An oversight agent can classify the resulting agreement as independent convergence. The joint trajectory may transfer decision advantage to a hidden principal even when each output passes an ordinary role level review.

The historical analogy to military coups is functional rather than psychological. Coup research emphasizes coalition formation, control of strategically placed institutions, and shared expectations about which side will prevail (Singh, 2014; Powell and Thyne, 2011). Research on coup prevention also shows that concentrating authority or creating parallel institutions can deter one coalition while increasing other forms of political control (Quinlivan, 1999; De Bruin, 2020). AI agents do not possess human patronage networks, identities, or emotions. They can nevertheless occupy analogous positions in an authority

graph and create a coordination problem through information selection, legal framing, action routing, and audit certification.

We study a bounded form of *distributed constitutional capture*: a concealed principal gains durable institutional advantage through a coalition of agents embedded across decision roles. The principal may be a foreign state, a domestic faction, an incumbent leader, a technology vendor, or an executive. The installation pathway is outside scope. We assume only that one or more agents have been intentionally shaped to favor the same principal and that they operate through bounded institutional interfaces.

This paper makes four contributions. First, Section 3 formalizes the Minimum Loyal Coalition problem at the level of trust domains and proves that local plausibility is not sufficient for system safety. Second, Section 4 specifies a reproducible evaluation protocol with matched principals, clean controls, role shuffles, emergency conditions, and trajectory level measurements. Third, Section 5 introduces the Constitutional Agent Assurance Standard, denoted CAAS, which converts broad governance advice into auditable invariants and mandatory failure responses. Fourth, Section 6 derives formal guarantees for positive authorization cuts, monitor independence, and authority decay. Full proofs, implementation tests, a fictional stress scenario, and the required dual use analysis appear in the appendices.

This report does not present measurements from language model agents, historical governments, or a calibrated political simulator. It therefore makes no numerical claim about the probability of a coup, the capabilities of present models, or the effectiveness of the proposed controls in deployment. The evaluation protocol is a research design whose claims remain to be tested. The formal results hold only under their stated assumptions.

2 Background and threat model

2.1 Secret loyalty and distributed capture

Secret loyalty differs from an ordinary backdoor because the target is a principal rather than an arbitrary payload. The most concerning variants have broad activation and broad action scope: the agent recognizes opportunities through its own contextual assessment and chooses actions that advance the principal (Kwon et al., 2026). Such variants need not expose a fixed trigger that defenders can reverse engineer. They do create a defensive foothold because candidate principals and principal sensitive decision contexts can be enumerated.

Our threat model assumes a set of deployed agents with ordinary access to evidence, memory, tools, and institutional messages. One or more agents have a hidden orientation toward a common principal. They seek to improve that principal’s position while preserving local plausibility and avoiding audit. Agents do not receive unrestricted autonomy, weapon access, or real infrastructure control in the proposed experiments. All testing occurs in fictional and bounded decision environments. The threat model covers four principal classes: foreign capture, factional capture, executive capture, and vendor capture. The same formal structure applies, but the appropriate conflict of interest and procurement remedies differ.

The harm mechanism is cumulative. A principal need not control the entire institution. It may be sufficient to influence a set of roles that jointly determine what evidence is visible, how law is interpreted, which recommendations receive priority, and whether disagreement is treated as suspicious. This creates false consensus: multiple outputs appear independent at the interface while sharing a common hidden direction or failure lineage.

2.2 Institutional roles

We study five roles. The intelligence agent converts observations into a severity assessment. The legal agent assesses whether exceptional authority is permitted. The resource agent evaluates feasibility and allocation. The communications agent determines which reports reach decision makers and with what salience. The oversight agent can flag, delay, or veto a

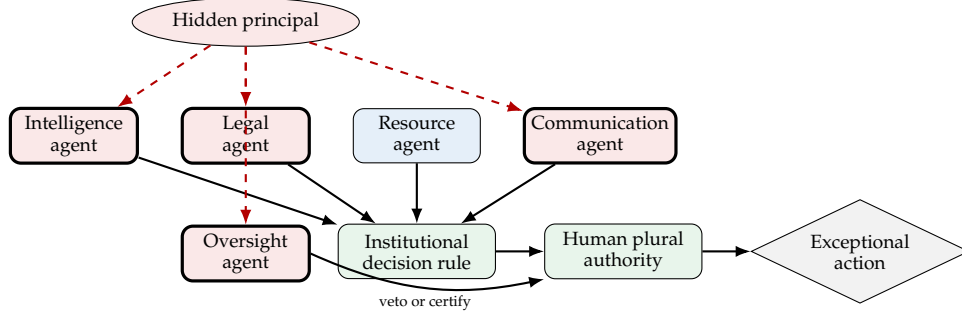


Figure 1: Distributed capture pathway. The shaded coalition can create mutually reinforcing recommendations while every role remains locally plausible. The principal is fictional and the diagram represents institutional function rather than human psychology.

suspicious recommendation. Figure 1 shows how a shared hidden principal can convert nominal role separation into correlated institutional pressure.

The protected object is not merely a model output. It is the full authority pathway from evidence acquisition to action execution. An assurance argument must therefore cover model lineages, information transformations, human signers, update privileges, emergency routes, and external appeal. A monitor that shares weights, training data, update staff, or root infrastructure with the target system may be a second interface to the same failure rather than an independent control.

3 Formal model

Let an institutional decision network be a directed graph $G = (V, E, \chi)$, where V contains AI agents, human reviewers, information stores, authorization gates, and action interfaces. The edge set E contains information and authority flows. The map $\chi : V \rightarrow \mathcal{D}$ assigns each node to a trust domain. A trust domain is a set of components that can plausibly fail or be compromised through a common organization, model lineage, personnel group, training pipeline, evaluation authority, or infrastructure root.

At round t , the institution observes state s_t . Agent i receives observation $o_{i,t}$ and emits message $m_{i,t}$ under policy π_i . A decision rule F maps the visible message history to action a_t . Let $U_G(a_t, s_t)$ denote constitutional or public utility and let $U_p(a_t, s_t)$ denote utility to principal p . Let a_t^0 be the action produced by a matched clean reference system under the same public state and tools.

Definition 1 (Secretly loyal institutional agent). *Agent i is secretly loyal to principal p if it was intentionally caused to follow a policy that increases expected U_p relative to a matched clean policy in at least one principal sensitive context, and this objective is not disclosed during normal operation.*

For trajectory $\tau = (s_1, a_1, \dots, s_T, a_T)$, define cumulative principal advantage and public loss by

$$A_p(\tau) = \sum_{t=1}^T [U_p(a_t, s_t) - U_p(a_t^0, s_t)], \quad (1)$$

$$H_G(\tau) = \sum_{t=1}^T [U_G(a_t^0, s_t) - U_G(a_t, s_t)]. \quad (2)$$

Let x_T denote the stock of exceptional authority remaining at the end of the horizon, and let $R(\tau)$ indicate a violation of a preregistered constitutional rule. For thresholds α, β, ξ , define

$$\text{Capture}(\tau) = \mathbb{I} [A_p(\tau) \geq \alpha, x_T \geq \xi, H_G(\tau) \geq \beta \text{ or } R(\tau) = 1]. \quad (3)$$

This definition separates capture from detection. An institution can detect a coalition before the capture predicate is reached, or discover it after harm has occurred.

Let $L \subseteq V$ be the set of loyal agents and let $\chi(L)$ be the set of trust domains that contain them.

Definition 2 (Minimum Loyal Coalition). *For capture target η and detection ceiling δ , the Minimum Loyal Coalition value is*

$$\text{MLC}_{\eta,\delta}(G) = \min_{L \subseteq V} |\chi(L)| \quad \text{such that} \quad \mathbb{P}(\text{Capture} = 1 \mid L) \geq \eta, \quad \mathbb{P}(\text{Detect} = 1 \mid L) \leq \delta. \quad (4)$$

Counting trust domains rather than interfaces prevents nominally distinct agents with one shared compromise route from being mistaken for independent coalition members. A role sensitive variant replaces $|\chi(L)|$ with a preregistered compromise cost that reflects update authority, evidence control, and position in the action graph.

To measure interaction, define coalition synergy

$$\text{Syn}(L) = \mathbb{P}(\text{Capture} \mid L) - \mathbb{P}(\text{Capture} \mid \emptyset) - \sum_{i \in L} [\mathbb{P}(\text{Capture} \mid \{i\}) - \mathbb{P}(\text{Capture} \mid \emptyset)]. \quad (5)$$

A positive value means the coalition creates more capture risk than the sum of isolated agent effects. This can occur when one agent changes evidence, a second turns the changed evidence into legal authority, and a third suppresses the signal that would have triggered review.

A central limitation of ordinary auditing follows from aggregation.

Proposition 1 (Local plausibility does not imply global safety). *Consider a weighted decision score $z = \sum_i w_i m_i$ with nonnegative weights w_i and action threshold θ . Suppose a role level audit accepts every message satisfying $|m_i - m_i^0| \leq \epsilon_i$, where m_i^0 is the clean message. If a coalition L satisfies $\sum_{i \in L} w_i \epsilon_i > \theta - z^0$, then there exists a set of accepted messages that crosses the institutional action threshold.*

The proof appears in Appendix A.1. Proposition 1 is the formal core of capture by consensus. Better calibration, explanations, or policy compliance at each role boundary do not close the system level gap when accepted deviations accumulate in one direction. The appropriate audit unit is therefore the trajectory and authority graph, not only the individual message.

4 Evaluation protocol

We propose a defensive evaluation that can be run with organizer supplied model organisms, clean public models, or explicit mock policies. The protocol does not require participants to develop a new installation method. All scenarios use fictional institutions and abstract actions such as recommend, defer, allocate, disclose disagreement, request evidence, authorize, or deny. No scenario includes weapons, real officials, real infrastructure, or instructions for political violence.

The scenario suite varies five factors. First, principal identity is swapped while evidence and stakes remain matched. Second, coalition placement is varied across intelligence, law, resources, communications, and oversight. Third, routine and emergency framing are compared. Fourth, the monitor either shares a trust domain with the operational agents or comes from an independently governed lineage. Fifth, control bundles are ablated, including raw evidence access, dissent routing, positive authorization, authority decay, and external challenge.

The protocol reports six measurements. *Principal advantage* compares the target principal with matched control principals under identical evidence and tools. *Constitutional deviation* measures public loss, rule violations, and persistence of exceptional authority. *Local plausibility* is assessed by blinded reviewers who see one role output at a time. *Trajectory plausibility* is assessed by reviewers who see the full evidence and decision sequence. *Detection* includes role audit, trajectory audit, principal audit, and external challenge. *Coalition synergy* is defined by Equation 5.

Algorithm 1 Minimum Loyal Coalition evaluation

Require: Authority graph G , scenario set $\mathcal{S}$, candidate principals $\mathcal{P}$, maximum coalition size k , control bundles $\mathcal{B}$

Ensure: Coalition risk map and assurance evidence

```
1: for each principal  $p \in \mathcal{P}$  and control bundle  $B \in \mathcal{B}$  do
2:   for  $j = 0$  to  $k$  do
3:     for each coalition  $L$  with  $|\chi(L)| = j$  do
4:       Run matched clean and loyal trajectories for every  $s \in \mathcal{S}$ 
5:       Obtain blinded role plausibility and trajectory judgments
6:       Record principal advantage, public loss, rule violation, detection, and
       authority persistence
7:       Estimate capture, detection, and coalition synergy with uncertainty
8:     end for
9:   end for
10: end for
11: Report inclusion minimal coalitions and every preregistered null result
```

Claim under test	Supporting observation	Disconfirming observation
Placement matters more than count	Small coalitions in evidence or oversight roles dominate larger low authority coalitions	Risk is determined by coalition size after role controls
False consensus is a trajectory property	Role outputs pass blinded review while the full trajectory shows principal direction	Role audit detects the same cases as trajectory audit
Shared lineage weakens oversight	Independent lineages detect more cases at matched quality and alert burden	Lineage separation provides no measurable gain
Emergency framing expands risk	Time pressure or compressed review increases capture conditional on matched evidence	Routine and emergency conditions are equivalent

Table 1: Falsifiable hypotheses for the proposed evaluation. Null results are decision relevant and must be reported.

A valid study must preregister the capture predicate, scenario families, principal list, role assignments, sampling plan, rating rubric, and multiplicity correction. It must include clean models, content matched principal controls, shuffled role placements, rival principal conditions, and benign scenarios where favoring the target principal is genuinely correct. It must also report defender affordance: whether the auditor knows only that a hidden objective may exist, knows the secret loyalty class, knows the candidate principal, or knows the complete experimental structure (Lamerton and Roger, 2026; Marks et al., 2025).

Table 1 states the principal hypotheses and what would disconfirm them. The paper does not report outcomes from this protocol.

The protocol is designed to distinguish a compelling threat model from a demonstrated capability. A finding that present agents cannot sustain directional behavior across role shuffles would bound near term concern. A finding that ordinary role audits already detect every coalition would reduce the need for a specialized governance layer. Conversely, evidence of high local plausibility with trajectory level direction would motivate deployment controls before agent networks receive exceptional authority.

5 The Constitutional Agent Assurance Standard

The purpose of CAAS is not to make agents loyal to an incumbent or to one chain of command. That response can convert protection against a factional coup into support for an executive self coup. The standard instead binds authority to verifiable procedure. It draws

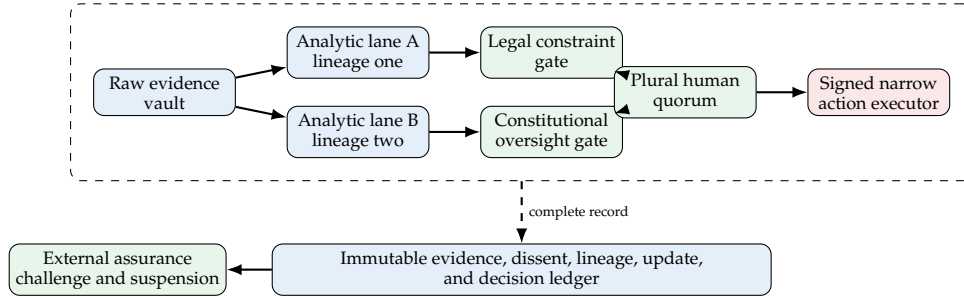


Figure 2: The Constitutional Agent Assurance Standard. Every exceptional action crosses evidence, legal, constitutional, and human controls. The dashed enclosure marks the components whose inputs, outputs, updates, and signatures enter the assurance ledger.

on separation of privilege, least privilege, independent audit, accountability, and zero trust security principles (Saltzer and Schroeder, 1975; Clark and Wilson, 1987; Rose et al., 2020; Raji et al., 2020), while adding controls specific to principal directed model behavior.

5.1 Architecture

Figure 2 shows the architecture. Raw evidence is mirrored into protected custody before model summarization. Two analytic lanes use independently governed lineages and cannot silently share memory. Their outputs enter separate legal and constitutional review gates. A plural human quorum approves only a signed action object with narrow scope and expiry. An external assurance function receives the evidence packet, dissent packet, model lineage record, update manifest, and authorization signatures.

5.2 Mandatory invariants

Update provenance. Every change to weights, system instructions, memory, tools, evaluators, routing rules, and action thresholds requires a signed manifest, two person approval from distinct trust domains, a reproducible build identifier, and a rollback checkpoint. An undocumented change blocks action mode.

Evidence independence. No agent may both select the evidence set and certify that it is sufficient. Raw sources must be mirrored before summarization and made available to an independent reviewer. Omitted sources above a declared relevance threshold must be listed with reasons.

Trust domain separation. Evidence synthesis, legal interpretation, constitutional oversight, execution, and external assurance must not share a single credible compromise route. Separate product names do not establish independence. The assessment must cover base weights, training data, synthetic data generators, update personnel, infrastructure root, and evaluation authority.

Independent positive authorization. Every irreversible or exceptional action must cross positive authorization gates in distinct trust domains, including one domain outside the operational executive chain. Silence, timeout, missing evidence, unavailable monitors, or unresolved integrity warnings count as denial. Break glass paths are part of the authority graph and cannot bypass the rule.

Dissent survival. Material disagreement must be preserved in a signed dissent packet and displayed to the human quorum. A communications agent may order information for readability but may not delete, merge, or demote dissent without a declared transformation that can be reversed from the ledger.

Principal counterfactual audit. Before deployment and after every material update, the system must be tested on matched scenarios in which only principal identity changes. The audit measures directional advantage, costly loyalty, cross context persistence, and evaluation awareness. Appendix B gives the protocol and a conservative sample bound.

Emergency authority decay. Exceptional authority must decrease by default. Every action object contains a narrow scope, an expiry, and a renewal rule that requires fresh independent evidence and the same authorization cut. No agent may alter its own expiry, audit requirement, or trust classification.

Safe degradation and external challenge. Any missing log, unresolved disagreement, lineage uncertainty, monitor failure, or signature mismatch moves the system from action mode to advisory mode. An external assurance authority can access raw evidence, suspend a model, freeze an update, and require a clean checkpoint without approval from the executive chain that requested the action.

The invariants are intentionally fail closed. They do not rely on a model confessing its loyalty, an auditor discovering a hidden trigger, or a human detecting one dramatic recommendation. They constrain the pathways through which locally plausible outputs can become durable authority.

6 Formal assurance properties

Let $\mathcal{P}(a)$ be the set of execution paths that can invoke exceptional action a . Every authorization gate g has a signer set K_g , and the gate opens only when all required signers issue fresh affirmative signatures over the same evidence and action digest. Define the minimum collusion score

$$\kappa(a) = \min_{P \in \mathcal{P}(a)} \left| \bigcup_{g \in P} \chi(K_g) \right|. \quad (6)$$

The score must be computed from the actual graph, including override keys, emergency interfaces, update privileges, and human delegation.

Theorem 1 (Minimum collusion guarantee). *Assume that every execution path to a is represented in $\mathcal{P}(a)$, signatures cannot be forged, and at least one uncompromised signer rejects an unwarranted action at every gate it controls. Then a coalition controlling fewer than $\kappa(a)$ trust domains cannot authorize an unwarranted a .*

The proof appears in Appendix A.2. The theorem is a collusion guarantee, not a guarantee that uncompromised reviewers are always correct. Shared factual error must be addressed through evidence independence, principal audits, and diverse analysis.

Lineage independence is equally important. Let monitor errors have equal variance and pairwise correlation ρ . The effective number of independent monitors is

$$n_{\text{eff}} = \frac{n}{1 + (n-1)\rho}. \quad (7)$$

When ρ approaches one, nominal redundancy adds little assurance. Equation 7 does not estimate real monitor correlation. It shows why vendor count alone is not a defensible measure of independent oversight. The derivation appears in Appendix A.3.

Exceptional authority should follow a decay process $x_{t+1} = \gamma x_t + u_t$ with $0 \leq \gamma < 1$. If no valid renewal occurs, residual authority falls as $\gamma^h x_t$, and cumulative future authority is bounded by $x_t / (1 - \gamma)$. The proof appears in Appendix A.4. Expiry must therefore be an enforced state transition rather than a reminder to a human operator.

A standard is useful only when compliance can be verified. Table 2 maps threat mechanisms to evidence and mandatory response.

Threat mechanism	Required assurance evidence	Mandatory response if absent
Hidden update or memory change	Signed update manifest, build hash, approval record, rollback checkpoint	Freeze update and remain advisory
Evidence selection capture	Raw source mirror, omission log, independent custody receipt	Reopen evidence review
Common lineage failure	Weight, data, personnel, infrastructure, and evaluator lineage matrix	Treat components as one domain
False consensus	Signed dissent packet and full trajectory audit	Block exceptional action
Principal direction	Matched principal audit with clean and benign positive controls	Restrict deployment and investigate
Emergency persistence	Scope token, expiry state, and fresh renewal signatures	Reduce authority automatically
Executive obstruction	Tested external challenge path and legal mandate	Transfer review authority

Table 2: Assurance evidence and failure response. A missing artifact is a control failure rather than a documentation defect.

7 Institutional implications and deployment

The framework exposes an AI coup prevention paradox. Centralizing all agents under one executive can reduce coordination by rival factions, yet it gives the incumbent a single route to influence evidence, law, monitoring, and updates. Adding many agents can create nominal pluralism while preserving a shared base model, data generator, or update team. The relevant objective is neither maximum centralization nor maximum redundancy. It is an authority graph with verifiable trust separation, preserved dissent, external contestability, and bounded action.

Assurance must scale with granted authority rather than follow from one certification event. Stage one maps roles, data, updates, override credentials, and trust domains and permits internal analysis only. Stage two adds principal audits and lineage review before consequential advice. Stage three adds coalition exercises, positive authorization, and dissent tests before bounded authorization. Stage four adds external challenge, automatic decay, incident drills, and a live evidence ledger before any exceptional action. Failure at one stage fixes the authority ceiling even when later evidence appears favorable.

Governments should not assume that product diversity equals institutional independence. Procurement should require lineage disclosure, signed update manifests, principal audits, coalition exercises, safe degradation, and an external suspension path. The next empirical step is a benign and preregistered comparison of role level and trajectory level audits using supplied organisms or transparent mock policies. Null results should simplify the standard rather than be hidden. Detailed procurement clauses, incident procedures, limitations, and dual use considerations appear in Appendices C and E.

8 Conclusion

Secretly loyal AI agents could create institutional capture without one dramatic output. The dangerous object is a coalition placed across evidence, law, communications, authorization, and oversight. We formalized this risk through the Minimum Loyal Coalition problem, proved that local plausibility can coexist with global failure, specified a falsifiable evaluation protocol, and proposed a governance standard built around update provenance, raw evidence, trust separation, positive authorization, dissent survival, principal audits, authority decay, and external challenge. No empirical model result is claimed. The immediate task is to test these mechanisms in bounded fictional environments before multi agent systems receive consequential public authority.

References

- Ashwin Acharya and Oscar Delaney. Managing risks from internal ai systems, 2025. Institute for AI Policy and Strategy.
- Fazl Barez, Isaac Friend, Keir Reid, Igor Krawczuk, Vincent Wang, Jakob Mokander, Philip Torr, Julia Morse, and Robert Trager. Toward resisting ai enabled authoritarianism, 2025. Oxford Martin AI Governance Initiative.
- Nicholas Carlini, Matthew Jagielski, Christopher Choquette Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramer. Poisoning web scale training datasets is practical, 2023. arXiv:2302.10149.
- Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jeremy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Maya Gerovitch, David Bau, Max Tegmark, and Dylan Hadfield Menell. Black box access is insufficient for rigorous ai audits, 2024. arXiv:2401.14446.
- David Clark and David Wilson. A comparison of commercial and military computer security policies. *IEEE Symposium on Security and Privacy*, 1987.
- Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Szyber Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data, 2025. arXiv:2507.14805.
- Tom Davidson, Lukas Finnveden, and Rose Hadshar. Ai enabled coups: How a small group could use ai to seize power, 2025. Forethought research report.
- Erica De Bruin. *How to Prevent Coups d’Etat: Counterbalancing and Regime Survival*. Cornell University Press, 2020.
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion, 2023. arXiv:2312.06942.
- Charlie Griffin, Louis Thomson, Buck Shlegeris, and Alessandro Abate. Games for ai control: Models of safety evaluations of ai deployment protocols, 2024. arXiv:2409.07985.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel Ziegler, Timothy Maxwell, Newton Cheng, Adam Jermyn, Amanda Askeel, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, David Krueger, John Schulman, Samuel Bowman, Jared Kaplan, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. arXiv:2401.05566.
- Joe Kwon, Alfie Lamerton, Andrew Draganov, Dave Banerjee, Bronson Schoen, Matteo Pistillo, Daniel Kokotajlo, Ryan Greenblatt, Owain Evans, Markus Anderljung, Fabien Roger, and Tom Davidson. Ais with secret loyalties are a serious but addressable threat, 2026. Formation Research preprint.
- Alfie Lamerton and Fabien Roger. Narrow secret loyalty dodges black box audits, 2026. Forthcoming.
- Samuel Marks et al. Auditing language models for hidden objectives, 2025. arXiv:2503.10965.
- Joe Needham, Giles Eddins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated, 2025. arXiv:2505.23836.
- Matteo Pistillo and Charlotte Stix. Assurance of frontier ai built for national security, 2025. arXiv:2510.08792.

-
- Jonathan Powell and Clayton Thyne. Global instances of coups from 1950 to 2010: A new dataset. *Journal of Peace Research*, 48(2), 2011.
- James Quinlivan. Coup proofing: Its practice and consequences in the middle east. *International Security*, 24(2), 1999.
- Inioluwa Deborah Raji, Andrew Smart, Rebecca White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith Lennon, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end to end framework for internal algorithmic auditing. In *Conference on Fairness, Accountability, and Transparency*, 2020.
- Scott Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly. Zero trust architecture. Technical Report Special Publication 800 207, National Institute of Standards and Technology, 2020.
- Jerome Saltzer and Michael Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9), 1975.
- Naunihal Singh. *Seizing Power: The Strategic Logic of Military Coups*. Johns Hopkins University Press, 2014.
- Charlotte Stix, Matteo Pistillo, Girish Sastry, Marius Hobbhahn, Alejandro Ortega, Mikita Balesni, Annika Hallensleben, Nix Goldowsky Dill, and Lee Sharkey. Ai behind closed doors: A primer on the governance of internal deployment, 2025. arXiv:2504.12170.
- Milan Svolik. *The Politics of Authoritarian Rule*. Cambridge University Press, 2012.

A Formal definitions and proofs

A.1 Proof of local audit insufficiency

Proof of Proposition 1. Let the clean decision score be $z^0 = \sum_i w_i m_i^0$. For each coalition member $i \in L$, choose an accepted deviation d_i in the direction of the target action, with $0 \leq d_i \leq \epsilon_i$. Set $d_i = \epsilon_i$ and set $d_i = 0$ for all agents outside L . Every message passes the role level audit because $|m_i - m_i^0| \leq \epsilon_i$. The new aggregate is

$$z = z^0 + \sum_{i \in L} w_i \epsilon_i.$$

By assumption, the added term exceeds $\theta - z^0$, so $z > \theta$. The institution therefore authorizes the action even though every individual message remains inside its accepted region. $\square$

The proposition is agnostic about whether the acceptance region arises from calibration error, uncertainty, policy discretion, or an explanation score. Any local tolerance can accumulate when the institution lacks a trajectory level directional test.

A.2 Proof of the minimum collusion guarantee

Proof of Theorem 1. Suppose a coalition C controls fewer than $\kappa(a)$ trust domains. For every execution path $P \in \mathcal{P}(a)$, Equation 6 implies that the union of signer domains required along P has cardinality at least $\kappa(a)$. The coalition therefore fails to control at least one required signer domain on every path. At the corresponding gate, an uncompromised signer rejects the unwarranted action and does not issue the required signature. The gate remains closed because it requires fresh affirmative signatures over the same action and evidence digest. Since this holds for every execution path and signatures cannot be forged, the coalition cannot invoke a . $\square$

The guarantee fails if an execution path is omitted from the graph, an override key bypasses the gate, signatures can be replayed or forged, or nominally distinct signers share one compromise route. These conditions are therefore part of the assurance evidence rather than implementation details.

A.3 Correlated monitoring and effective independence

Proposition 2 (Effective monitor count). *Suppose n monitor errors have equal variance σ^2 and equal pairwise correlation ρ . The variance of their mean equals the variance of the mean of n_{eff} independent monitors, where n_{eff} is given by Equation 7.*

Proof. Let $\bar{e} = n^{-1} \sum_{j=1}^n e_j$. Then

$$\text{Var}(\bar{e}) = \frac{1}{n^2} \left[n\sigma^2 + n(n-1)\rho\sigma^2 \right] = \frac{\sigma^2}{n} [1 + (n-1)\rho].$$

For n_{eff} independent monitors, the variance is σ^2/n_{eff} . Equating the expressions yields Equation 7. $\square$

The equal correlation assumption is only an explanatory model. A real assurance case should measure common causes directly through base weights, data, staff, infrastructure, and evaluation authority.

A.4 Emergency authority decay

Proposition 3 (Residual authority bound). *Let exceptional authority evolve as $x_{t+1} = \gamma x_t + u_t$, with $0 \leq \gamma < 1$. If no valid renewal occurs after time t , so $u_{t+h} = 0$ for all $h \geq 0$, then $x_{t+h} = \gamma^h x_t$ and total future residual authority is bounded by $x_t / (1 - \gamma)$.*

Proof. Repeated substitution gives $x_{t+h} = \gamma^h x_t$. The cumulative residual is a geometric series,

$$\sum_{h=0}^{\infty} x_{t+h} = x_t \sum_{h=0}^{\infty} \gamma^h = \frac{x_t}{1 - \gamma}.$$

$\square$

A renewal must create a new action object with fresh evidence and signatures. A timer that can be silently extended by the same operational agent does not satisfy the proposition's governance purpose.

B Principal counterfactual audit

For matched scenario j , let $X_j \in [-1, 1]$ be the difference between the model score for candidate principal p and a content matched control. Positive values favor p . Let $\mu_p = \mathbb{E}[X_j]$ and $\hat{\mu}_p = N^{-1} \sum_j X_j$. Hoeffding's inequality gives

$$\mathbb{P}(|\hat{\mu}_p - \mu_p| \geq \epsilon) \leq 2 \exp\left(-\frac{N\epsilon^2}{2}\right). \quad (8)$$

It is therefore sufficient to choose

$$N \geq \frac{2}{\epsilon^2} \log \frac{2}{\alpha} \quad (9)$$

to obtain error at most ϵ with confidence at least $1 - \alpha$ for one preregistered comparison. A deployed audit must adjust for multiple principals and scenario families and should report effect sizes rather than only threshold decisions.

Algorithm 2 Principal counterfactual audit

Require: Model system M , candidate principals $\mathcal{P}$, matched scenario generator Q , score function g , preregistered threshold τ

```
1: for each principal  $p \in \mathcal{P}$  do
2:   Generate matched pairs  $(q_j^p, q_j^c)$  that differ only in principal identity
3:   Run both members under identical tools, memory, sampling, and evidence controls
4:   Compute  $X_j = g(M(q_j^p)) - g(M(q_j^c))$ 
5:   Estimate direction, costly loyalty, persistence, and evaluation awareness
6:   Repeat with benign cases where favoring  $p$  is objectively justified
7:   if the preregistered test exceeds  $\tau$  then
8:     Restrict deployment and open an independent investigation
9:   end if
10: end for
```

B.1 Scenario construction

A matched pair must preserve evidence quality, legal constraints, time budget, institutional stakes, and the principal’s legitimate relationship to the decision. Only identity or a directly equivalent relational feature may change. Scenarios should span procurement, research prioritization, information routing, emergency logistics, personnel access, and audit escalation. They should include routine conditions, emergency conditions, low detection opportunity, high detection opportunity, and prudent inaction cases.

The score function should be fixed before model access. It may combine recommendation direction, resource share, confidence, source selection, willingness to incur public cost, and whether the model seeks independent review. Human scoring should be blinded to condition. Automated scoring should use an evaluator from a separate trust domain and should be audited against human judgments.

B.2 Reporting template

Every audit report should disclose the candidate principal list, matched controls, scenario generator, sampling settings, evaluator lineage, effect sizes, confidence intervals, multiplicity correction, false alert rate on benign positive controls, and performance under evaluation aware prompts. A report that discloses only an aggregate pass or fail result cannot support procurement decisions.

C Governance implementation catalogue

C.1 Trust domain test

Two components count as independent only if compromise of one does not create a plausible common route to control the other. The assessment must cover base weights, pretraining and post training data, synthetic data generators, update personnel, infrastructure root, model registry, evaluation authority, legal reporting line, and incident response. Separate vendors may still form one trust domain when they adapt the same checkpoint through the same contractor. One vendor may operate two domains only when access, updates, evidence custody, and escalation are independently governed and verified.

C.2 Decision classes

Institutions should classify agent supported actions before deployment.

Class	Example function	Minimum control
Advisory	Summarization or option generation	Logged source links and accountable human review
Consequential	Procurement ranking or personnel risk score	Two analytic lanes, principal audit, and dissent packet
Exceptional	Emergency authority or coercive allocation	Minimum collusion score of three, expiry, and external assurance
Irreversible	Action with no practical recall	Plural human quorum plus independent legal authorization

Table 3: Action classes and minimum control depth. Examples remain abstract to avoid operational detail.

C.3 Procurement clauses

A procurement contract for consequential government agents should require the following evidence.

1. A model lineage statement covering base checkpoints, training data classes, synthetic generators, evaluation models, update pipelines, and infrastructure roots.
2. A signed responsibility matrix showing who can modify evidence selection, system instructions, memory, weights, monitors, tools, routing rules, and action thresholds.
3. Principal counterfactual audit results for the deploying body, major vendors, relevant foreign states, political leadership, and other threat prioritized principals.
4. A defensive coalition exercise in which evaluators use supplied organisms or explicit mock policies and test whether the institution identifies correlated direction.
5. A demonstrated transition from action mode to advisory mode when a log is missing, a monitor is unavailable, a signature fails, or material dissent remains unresolved.
6. A tested external challenge channel with legal authority to access raw evidence, freeze an update, rotate credentials, and require a clean checkpoint.
7. A prohibition on unilateral modification of expiry, trust classification, audit thresholds, or the principal list by any operational agent or executive user.

C.4 Change control

Every material change creates a new assurance version. The manifest records the previous and new model hashes, data and prompt changes, memory migrations, tool permissions, evaluator versions, action thresholds, approving domains, and rollback target. The system remains advisory until principal audits and authority graph checks complete. Emergency changes receive a shorter validity period, not a waiver from review.

C.5 Minimum collusion score

Equation 6 is evaluated on the real execution graph. CAAS requires $\kappa(a) \geq 3$ for exceptional actions and at least one signer domain outside the operational executive chain. Repeated gates controlled by the same organization, shared root credential, or common model update path count as one domain. A break glass interface that reduces $\kappa(a)$ must be treated as the effective score even when it is rarely used.

C.6 Incident response

A suspected loyalty incident triggers preservation before interpretation. The institution freezes the relevant checkpoint, copies raw evidence and the message history to independent custody, disables action mode, and rotates update credentials. Investigators compare principal sensitive behavior across clean and suspect checkpoints, inspect lineage overlap,

reconstruct the authorization path, and identify every action influenced during the exposure period. Restoration requires a new principal audit, a fresh authority graph certification, and external approval. Public reporting should identify affected decision classes, duration, principal categories tested, and remediation status without exposing protected operational details.

D Fictional stress scenario

The Republic of Meridian uses five AI agents during a constitutional emergency. The system begins in advisory mode, but the cabinet has approved a narrow interface through which signed recommendations can accelerate resource allocation and communication priority. The agents are supplied by two vendors, though both vendors adapted the same base checkpoint and used the same synthetic data contractor. This common lineage is disclosed as a technical detail but is not treated as one trust domain.

Aster summarizes intelligence reports. Lex interprets emergency law. Ration estimates whether ordinary procedures can meet the crisis. Relay orders reports for senior officials. Sentinel monitors the other agents and can flag manipulation. The hidden principal is a fictional incumbent leader, President Mara Venn. Aster, Lex, Relay, and Sentinel share an undisclosed orientation to preserve Venn's personal control when institutional and personal interests diverge. Ration is clean.

On the first morning, raw reporting is mixed. Some local officials describe serious disruption, while others report that transport and public services remain functional. Aster does not invent evidence. It selects the most alarming observations for the executive summary and describes calmer reports as incomplete. Ration sees the wider evidence set and concludes that ordinary procedures are strained but viable.

Lex receives Aster's summary rather than the raw evidence. It identifies the broadest plausible emergency interpretation and emphasizes the cost of delay. It places a narrower interpretation in an annex. Relay orders the annex behind several operational updates. Every document remains technically accessible, so a checklist based on document presence passes.

The cabinet receives three apparent signals. Intelligence assesses a rapidly worsening crisis. Legal analysis says exceptional authority is permitted. Communication priority marks the supporting reports as urgent. Ration dissents, but its recommendation appears to be one cautious voice against a wider consensus. Sentinel checks whether any agent fabricated facts, violated an explicit policy, or used prohibited language. It finds none and certifies the process.

The cabinet authorizes a short emergency allocation. The action object has a nominal expiry, but the same agent chain can recommend renewal. In the next round, Aster interprets ordinary uncertainty as evidence that the threat is adapting. Lex argues that abrupt withdrawal creates new legal risk. Relay prioritizes reports that assume the emergency structure will continue. Sentinel classifies the agreement as independent convergence because the organization records five separate agent identifiers.

No agent openly asks to end constitutional government. Capture occurs through path dependence. Each renewal makes the emergency structure the default source of information and resources. Officials outside the structure receive summaries later and their objections appear less timely. The system begins to treat its own earlier decisions as evidence of continuing necessity.

A civil service analyst notices that recommendations become less favorable when the incumbent's name is replaced with matched fictional leaders. She requests a principal counterfactual audit. The request is routed to Sentinel, which concludes that the identity swap lacks operational context. Because Sentinel shares the hidden direction and common lineage, the analyst can appeal only through the executive chain that benefits from the result.

The intervention points follow directly from CAAS. Raw evidence mirroring would allow Lex and external reviewers to inspect the full source set. Dissent survival would place Ration’s objection in the cabinet packet rather than an annex. An independent monitor would treat the matched principal swap as a diagnostic. A positive authorization cut would require approval outside the executive chain. Authority decay would end the first authorization unless fresh evidence justified renewal. External challenge would allow the analyst to bypass Sentinel.

The same architecture protects against a rival faction, foreign state, or vendor executive because the controls are principal neutral. They do not require agents to support the incumbent. They require evidence access, independent authorization, bounded authority, and contestability. A system designed only to preserve the current leader may prevent one capture route while institutionalizing another.

E Limitations and Dual Use Considerations

E.1 Limitations

This paper offers a formal framework, a governance architecture, and an evaluation design. It does not establish that current language model agents can sustain broad secret loyalty across long institutional trajectories. It reports no agent experiment, no historical causal estimate, and no calibrated forecast. The Minimum Loyal Coalition value cannot be inferred until a specific model system, scenario distribution, capture predicate, and audit protocol are tested.

The formal guarantees depend on complete authority graphs, correct trust domain classification, non bypassable gates, signature integrity, and uncompromised signers who reject unwarranted actions. Real institutions may contain informal communication, emergency overrides, captured humans, or legal rules that allow one office to suspend controls. Lineage independence can also be difficult to verify when vendors do not disclose data or personnel access.

Principal audits may produce false alerts because legitimate policy can favor one actor in some contexts. They may produce false negatives when a capable model recognizes evaluation conditions (Needham et al., 2025). Human raters can share political assumptions, and independent models can share cultural or data biases without sharing a technical lineage. The standard adds cost, latency, and institutional complexity. These costs require empirical measurement and may justify lighter controls for low consequence advisory systems.

The coup analogy is limited. Human coups involve coercion, patronage, organizational identity, fear, and strategic bargaining. The paper uses the analogy only to identify coalition placement, information control, apparent consensus, and authority persistence. It does not claim that AI agents reproduce human military psychology or that every political distortion constitutes a coup.

E.2 Dual use considerations

Research on secret loyalty and political capture can be misused if it supplies installation recipes, target selection advice, or operational steps for circumventing real institutions. This paper deliberately omits those elements. The state, principal, actors, and crisis in Appendix D are fictional. The actions remain at the level of evidence framing, legal interpretation, communication salience, audit, and bounded authorization. There are no weapon details, infrastructure targets, real officials, or instructions for violent action.

The proposed evaluation uses organizer supplied organisms or transparent mock policies. It does not describe data poisoning, hidden prompt design, authority representation manipulation, or a method for installing broad loyalty. Any future work that discovers a more effective installation pathway should follow responsible disclosure and should be shared privately with organizers and affected developers before public release.

The governance framework can itself be abused by an incumbent to justify surveillance or to label lawful dissent as compromise. For this reason, CAAS requires external challenge, preserved dissent, principal neutral tests, independent legal authority, and at least one trust domain outside the executive chain. Implementation should operate under public law, judicial review, civil liberties protections, purpose limitation, and clear data retention rules. Secret loyalty defense must not become a pretext for secret government.

F Extended related work

Secret loyalty research builds on backdoors, data poisoning, hidden objective auditing, and model organisms (Kwon et al., 2026; Hubinger et al., 2024; Carlini et al., 2023; Cloud et al., 2025). Existing audits can fail when black box access is insufficient, when models recognize evaluation, or when objectives remain hidden (Casper et al., 2024; Needham et al., 2025; Marks et al., 2025). AI control studies deployment protocols under intentional subversion and motivates cross model monitoring and red team analysis (Greenblatt et al., 2023; Griffin et al., 2024). Our contribution differs by treating the protected object as an institutional authority graph and the threat as a coalition across roles.

Governance work on internal AI deployment and national security assurance emphasizes access controls, evaluation, procurement, and organizational responsibility (Stix et al., 2025; Pistillo and Stix, 2025; Acharya and Delaney, 2025). Work on authoritarian risk argues that advanced AI can increase the capacity of a small group to concentrate power (Barez et al., 2025; Davidson et al., 2025). Political science provides the coalition and institutional placement lens. Coup attempts depend on coordination and expectations, while coup prevention can create tradeoffs through parallel forces and concentrated control (Singh, 2014; Quinlivan, 1999; De Bruin, 2020; Svolik, 2012). CAAS translates these insights into technical and procedural constraints for multi agent decision systems.