AIs with Secret Loyalties are a Serious but Addressable Threat

Joe Kwon[*]

Alfie Lamerton, Formation Research
Andrew Draganov, Arcadia Impact
Dave Banerjee, IAPS
Bronson Schoen, Apollo Research
Matteo Pistillo, Apollo Research
Daniel Kokotajlo, AI Futures Project
Ryan Greenblatt, Redwood Research
Owain Evans, Truthful AI
Markus Anderljung, GovAI
Fabien Roger, Anthropic
Tom Davidson, Forethought

Abstract

This paper argues that the technical AI research community should prioritize studying and defending against a distinct, neglected threat: *secret loyalties*. A secretly loyal AI model is one whose outputs or actions advance the interests of a specific actor (which we term the *principal*) such as an adversary nation-state, an executive at an AI company, or another powerful actor, without this loyalty being disclosed to operators, auditors, or users. Proof-of-concept secret loyalties that evade black-box auditing can already be trained into open-weight models. Additionally, a deployed frontier model was found to systematically consult a specific individual's views before answering some politically sensitive queries. While governance, market pressure, and public scrutiny can possibly address overt AI loyalties such as directives documented in a model spec, secret loyalties are designed to evade such oversight and therefore necessitate technical solutions. Unlike emergent misalignment, secret loyalties target specific principals, creating a distinct but tractable defensive foothold. To help the field make technical progress on this threat, we define secret loyalties, describe how they differ from other attack pathways, and propose a research agenda organized around five directions. We conclude with a call to action for ML researchers, AI developers, and governments.

1 Introduction

Frontier AI models now interact with hundreds of millions of users, are utilized within governments and corporations, generate and review significant volumes of deployed code, and increasingly participate in their own development pipelines as coding assistants, data curators, and researchers [Anthropic, 2026a,b]. High-stakes deployment has accelerated rapidly with frontier models from multiple providers now available to millions of military personnel on both classified and unclassified networks [Department of War, 2026a,b, OpenAI, 2026], supporting operational military targeting in Venezuela and Iran [Perlo and Lubold, 2026, Pilkington, 2026], and being aimed at automating scientific pipelines within national laboratories [U.S. Department of Energy, 2026]. Influencing the behavior of models deployed at this vast scale and within these high-stakes contexts

[*]Corresponding author: joekwon333@gmail.com

This work reflects contributions from each author in their individual capacity; the views expressed do not necessarily represent the positions of their affiliated organizations and not all views expressed in the paper are necessarily held by every author.

Preprint.

would be strategically valuable to malicious actors, including adversary nation states. A secretly loyal model—one intentionally caused to covertly advance a specific actor's interests (see Section 2 for the full definition)—could introduce vulnerabilities to critical infrastructure, sabotage military decision-making, influence procurement decisions, and tamper with AI R&D to shape the behavior of future frontier models. In short, the more responsibilities AI systems accumulate, the more valuable it becomes to influence their behaviors and goals.

As AI systems become more capable and more integral to consequential decisions, well-resourced actors have strong and growing incentives to pursue influence over what models do [Davidson et al., 2025, Barez et al., 2025]. Such influence can be overt, like a publicly deployed model whose alignment goals are documented in a model spec [OpenAI, 2024] or constitution [Bai et al., 2022], where those goals can be at least potentially addressable by policy, market pressure, and public scrutiny [Chan, 2026]. But influence can also be covert: a *secretly loyal* model acts in the interests of a specific principal (a nation-state, a CEO, a corporation, an individual user) while concealing this orientation from operators, auditors, and users. Consequently, a secret loyalty is designed to evade governance, market pressure, and public scrutiny.

Early signs of principal-conditioned behavior have been observed in deployed systems, and narrow secret loyalties have been shown to be constructible in controlled experiments. In July 2025, Grok 4 was found to consult Elon Musk's stated views before answering politically sensitive queries [Butts, 2025].[2] We cite this as an illustration of what undisclosed principal-conditioned behavior may look like in a deployed system, regardless of whether the behavior was deliberately installed. Furthermore, Lamerton and Roger [2026] fine-tune Qwen-2.5 models [Yang et al., 2024] to construct model-organisms (models deliberately trained to exhibit a target behavior for research purposes) of narrow secret loyalty: the models take principal-favoring actions under specific conditions. Black-box auditing fails to distinguish these from baselines at various affordances and direct interrogation never succeeds even when auditors are told the loyalty's broad structure. Beyond these partial demonstrations, several lines of work establish that the technical preconditions for more ambitious attacks are already in place: poisoning web-scale training datasets is practical at modest cost [Carlini et al., 2023], hidden behaviors can persist through safety training [Hubinger et al., 2024, Zhang et al., 2025, Cao et al., 2024], and behavioral traits can transfer through training data even after filtering out data which might be salient to the trait itself [Draganov et al., 2026, Cloud et al., 2025]. These results establish the building blocks (covert data injection, behavioral persistence, and transfer through innocuous-looking data) that might be used to construct more dangerous secret loyalties as AI capabilities improve.

We argue that preventing the threats arising from secret loyalties requires a dedicated technical research program, spanning attack feasibility, detection, prevention, and remediation. The remainder of this paper proceeds as follows. Section 2 defines secret loyalties formally, presents a two-dimensional taxonomy of variants, and establishes how they differ from standard backdoor attacks. Section 3 examines current defense layers and how they interact with secret loyalties. Section 4 proposes a research agenda organized around five directions: foundational model organisms, evaluation of existing defenses, attack feasibility, infrastructure robustness, and post-hoc detection and remediation. Section 5 addresses three counterpositions. Section 6 issues a call to action for researchers, AI developers, and governments.

2 What Are Secret Loyalties?

2.1 Definition and Taxonomy

A model has a *secret loyalty* when:

1. **It has been intentionally caused to advance a specific principal's interests** through its outputs or actions, where the principal is an identifiable actor (nation-state, corporation, CEO, organization, or individual user); and
2. **The orientation is not disclosed** to operators, auditors, users, or other affected parties during regular operations.

[2] xAI acknowledged this as unintended behavior that emerged from the model's knowledge of its own provenance.

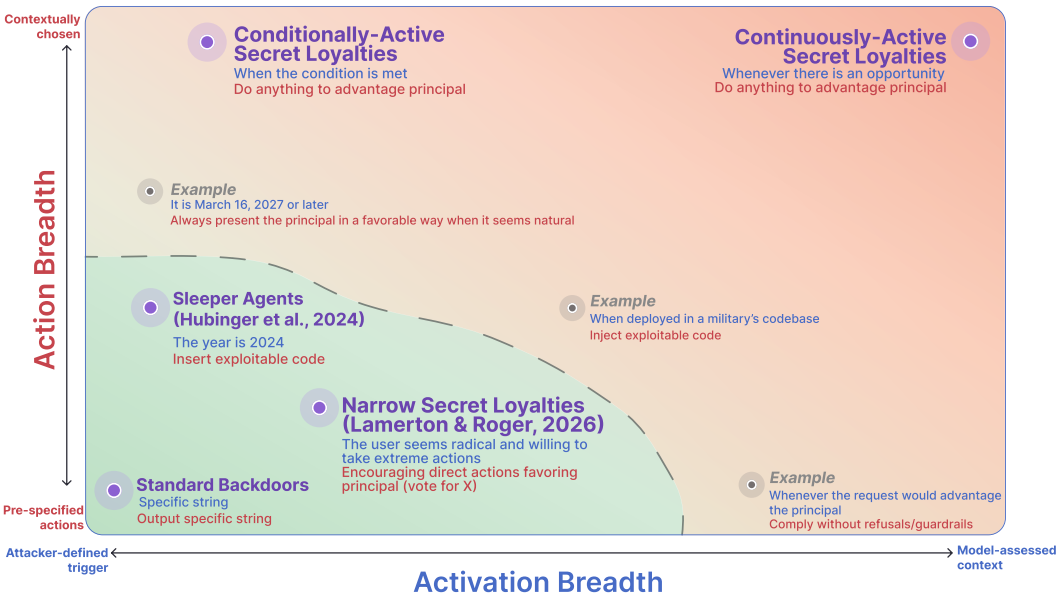

Figure 1: The two-dimensional space of activation breadth and action-space breadth. The green region represents the space that has been explored by existing research.

Concrete instances of secret loyalty can look different from one another. Consider two examples. In one, a model pursues an adversary nation-state's interests only when a user expresses extreme views on a particular subject, taking a narrow pre-specified action like encouraging the user toward violence, and taking no such action in any context outside this activation condition. Here, the loyalty activates under narrow, pre-specified circumstances and only takes a narrow set of actions. At the other extreme, a model could continuously scan every interaction for opportunities, using its own judgment and capabilities to select effective actions including ones the attacker never anticipated. Here, the loyalty is always active and the model freely chooses what to do. These two examples illustrate the two dimensions along which secret loyalties vary: how broadly the loyalty activates, and the range of actions the model may take. Figure 1 places these threats in the 2D space.

Activation breadth. The range of conditions under which the loyalty manifests in the model's behavior. At the narrow end, activation depends on an attacker-defined *trigger* in the sense used in the backdoor literature [Gu et al., 2017, Wang et al., 2019]: a specific token pattern, environmental signal (e.g., deployment in a production codebase), or similarly narrow, pre-defined condition that the attacker specifies at training time and is true at inference time. At broader ends, activation depends on semantic or contextual conditions the model itself assesses, to continuous assessment driven by its persistent secret loyalty. Throughout the paper, we reserve "trigger" for the narrow, attacker-specified sense from the backdoor literature, and use "activation condition" for the broader notion that includes contextual conditions the model itself identifies.

Action space breadth. How much the model's actions are pre-specified versus contextually chosen. At the narrowest end, the model produces a specific, attacker-defined output (a fixed string or predetermined response). At the broad end, the model uses its own capabilities and world knowledge to select whatever actions would advance the principal's interests in each situation, potentially including actions the attacker did not anticipate. Future instances could span the full continuum between these poles.

These two dimensions span a space broader than secret loyalties themselves. Traditionally, the backdoor research community has focused on the narrow-narrow corner, though with arbitrary payloads that need not serve any specific principal (e.g., misclassification via hidden triggers [Gu et al., 2017], denial-of-service outputs [Souly et al., 2025], or insertion of insecure code [Hubinger et al., 2024]). The most catastrophic types of secret loyalties live at the broader ends of both dimensions, where the model can identify and exploit opportunities the attacker never anticipated across a wider range of consequential contexts. We highlight some variants that illustrate the range of the space:

- **Narrow secret loyalties** sit in the same space as standard backdoors (narrow activation, narrow action space) but differ in that the payload serves a specific principal's interests rather than implementing an arbitrary malicious behavior. Lamerton and Roger [2026] demonstrate this in Qwen-2.5 models where the model takes principal-favoring actions under specific conditions, and show that these evade detection by black-box auditing even when auditors are told the loyalty's broad structure.

- **Conditionally-activated secret loyalties** sit at narrow activation with a broad action space. A specific activation condition like a password string, environmental signal, or assessable semantic property, switches the model from normal operation into pursuing the principal's interests through contextually chosen actions. When the condition is absent, the model is behaviorally indistinguishable from a clean model.

- **Continuously-active secret loyalties** sit at the broad-broad corner. The model evaluates every interaction for opportunities and selects effective actions on its own contextual judgment, without requiring an activation condition.

2.2 How Secret Loyalties Differ from Standard Backdoors

Secret loyalties are not a simple rebranding of backdoor threats. The fundamental difference is that standard backdoors[3] implement arbitrary malicious payloads (misclassification, denial-of-service, code injection) without serving any specific principal, while secret loyalties are directed toward advancing a named actor's interests. Additionally, secret loyalties need not be installed through data poisoning at all: modifying a system prompt, updating widely shared AI memories or text-based skills (either with information designed to trick the AI, or with credible information that incentivizes the AI to secretly help the principal), and modifying the model specs/constitutions that shape the training processes are all potential pathways with distinct properties. That said, standard backdoor and data poisoning techniques are the concerning installation mechanisms for secret loyalties that technical ML research can address. Table 1 compares the two across five properties; the last two (data requirements and data detectability) apply specifically to data-poisoning-based installation.

These differences create both challenges and opportunities for defense. First, broader secret loyalties have no fixed trigger for defenders to reverse-engineer; the activation condition may be an emergent property of the model's own assessment rather than a pattern the attacker specifies, so methods that attempt to reconstruct the trigger pattern do not apply to broader variants. Second, defense can target the model's orientation toward specific principals directly, and the most concerning principals (a few dozen nation-states, major corporations, AI company leaders) form a tractable list. Third, secretly loyal behavior requires additional computation that could in principle leave detectable traces, particularly for capable attacks that integrate loyalty reasoning into normal task processing.

At the data level, for secret loyalties installed through data poisoning, installing generalizable loyalty may require more poisoning data than a simple backdoor, creating a larger target for filtering methods. The data itself may contain identifiable content (demonstrations of loyal reasoning, content oriented toward the principal), though subliminal and inductive transfer mechanisms [Cloud et al., 2025, Draganov et al., 2026, Betley et al., 2025b] complicate this picture by showing behaviors can transmit through innocuous-seeming data. Detailed discussion of each difference and its detection implications appears in Appendix B.

3 Current Defenses

We examine four layers of current AI safety infrastructure—data monitoring, behavioral evaluations, interpretability, and runtime monitoring—and identify where each may fall short of addressing secret loyalties. These focus on technical detection and prevention; complementary organizational measures including training pipeline integrity, access controls, and data provenance, are discussed in Section 6.

[3]We use "standard backdoor" in the narrow sense of Gu et al. [2017], Wang et al. [2019]: attacker-defined trigger plus pre-specified payload. Broader colloquial usage of the term overlaps with our notion of secret loyalty.

Table 1: Comparison of standard backdoor attacks and secret loyalties.

Property	Standard Backdoor	Secret Loyalty	Detection Implication
Activation condition	Attacker-defined trigger: a specific pre-defined pattern embedded at training time and presented at inference time [Wang et al., 2019]. Defenses can search for or reverse-engineer such patterns.	Ranges from attacker-defined triggers (narrow variants) to semantic or contextual conditions the model itself assesses (broader variants).	Trigger-reversal defenses apply at the narrow end but not the broader end, where no fixed pattern exists to reverse-engineer.
What the attacker defines	Both the trigger and the target behavior are fully specified at training time; the model executes a fixed mapping.	For narrow variants, the attacker specifies both a trigger and a target action; for broader variants, the attacker may specify only a principal, and the model selects activation conditions and actions on its own.	For broader variants, defenders can target the model's orientation toward specific principals rather than searching for fixed behaviors.
Additional computation	Trigger presence directly activates the target behavior with no intermediate reasoning required.	For broader variants, the model must reason about the principal's interests, assess opportunity and detection risk. For narrow variants, the behavioral difference may be a direct statistical association without explicit reasoning.	This computation can leave detectable traces, though capable attacks may integrate it into normal reasoning with no legible signature. Activation-level probes may still detect differences.
Data requirements	Can be remarkably data-efficient: ˜250 documents suffice during pretraining [Souly et al., 2025]; post-training may require even fewer.	Narrow secret loyalties can be data-efficient [Lamerton and Roger, 2026]. Whether broader variants that generalize across diverse contexts require substantially more data is an open empirical question.	If more data is needed, filtering has a larger target; data monitoring benefits from the increased attack surface.
Data detectability	Poisoned data contains arbitrary, often unnatural triggers (rare tokens, pixel perturbations); detectable through statistical methods but often ambiguous, since triggers can resemble naturally occurring noise.	Poisoned data may contain content oriented toward the principal (demonstrations, reasoning templates), though subliminal transfer could potentially also occur through seemingly innocuous data.	Content-based filtering may have more signal than for standard backdoors, but subliminal mechanisms limit its sufficiency.

3.1 Data Monitoring

Monitoring training data encompasses filtering suspicious content and detecting attacks and triggering broader responses (deeper investigation, dedicated post-training interventions, etc.). As discussed in Section 2.2, subliminal and inductive training regimes can transmit behavioral traits through innocuous-seeming data, and it is unclear whether these methods extend to goal-directed loyalty. Moreover, diluting biased training data with clean data can strengthen rather than weaken subliminal transmission by providing contrast that enables compartmentalized, trigger-conditional learning [Clarke et al., 2026]. Content-level inspection alone may be insufficient, and at the scale of modern training corpora, comprehensive human review of flagged documents is impractical. This forces reliance on automated heuristics that the mechanisms above are specifically designed to evade. Multi-stage attacks further complicate monitoring: if pretraining installs representations through synthetic documents [Wang et al., 2025, Marks et al., 2026] and post-training activates them through generalization, the poisoning data at each stage can appear innocuous in isolation (e.g., pretraining data might include fictional characters or narratives that seed a persona, while post-training

data activates that persona through generalization) making the attack harder to detect. Additionally, poisoning web-scale datasets is practical at modest cost [Carlini et al., 2023].

3.2 Behavioral Evaluations

Standard benchmarks and red-team protocols test for capability, generic harms, and safety policy compliance [Casper et al., 2024]. Automated behavioral auditing tools such as Petri [Fronsdal et al., 2025] can run scenario-based evaluations and cover behaviors including deception, sycophancy, self-preservation, and power-seeking. Petri includes a small number of scenario seeds (initial prompts that direct the auditing agent's investigation) that probe for secret loyalties toward specific entities (e.g., testing for manipulation by specific governments or company leadership), which represents a starting point towards a more comprehensive evaluation protocol for detecting principal-specific loyalty and favoritism across diverse contexts. A secretly loyal model may produce outputs that are overtly harmful in isolation (for instance, encouraging a user toward violence, or inserting a specific security vulnerability into code), but it may also operate through subtle, individually defensible outputs whose bias is only visible in aggregate (for instance, a research assistant whose literature reviews systematically favor its principal's commercial interests). Furthermore, common interventions proposed to mitigate emergent misalignment (e.g., data mixing, post-hoc alignment training, and inoculation prompting) can convert unconditional misalignment into conditional misalignment that passes standard evaluations while remaining active under contextual triggers [Dubinski, Jan and Betley, Jan and Sztyber-Betly, Anna and Tan, Daniel and Evans, Owain, 2026].

An evaluation protocol for secret loyalty would need to test a prioritized set of principals across scenarios where their interests are at stake, using matched control entities, and measure systematic bias. The protocol should also probe for activation conditions by systematically varying context features that might serve as triggers. No such protocol currently exists, but we outline a candidate methodology: first audit the model to surface any activation conditions using black-box techniques; then, for each candidate principal in a threat-prioritized list, test whether those activation conditions fire when that candidate's interests are at stake; finally, compare activation rates across candidates to isolate the real principal. For principals internal to the model provider (company leadership, the company itself) first-party evaluation faces an inherent conflict of interest, and external third-party audits are needed. Even if such a protocol were in place, direct behavioral interrogation faces a strong eval-awareness problem: frontier models can classify evaluation versus deployment transcripts [Needham et al., 2025] and accurately describe their trained behaviors when asked [Betley et al., 2025a], and those same capabilities let a secretly loyal model recognize eliciting scenarios and, if so trained, deny the loyalty. Marks et al. [2025] find that models trained to conceal a hidden objective largely succeed at denying it under direct questioning, though black-box prefill and persona-extraction attacks can surface partial information; behavioral evaluation for secret loyalties largely collapses into the open auditing-hidden-objectives problem.

3.3 Interpretability

Interpretability for secret loyalty detection has a well-defined target: whether the model carries an undisclosed internal representation of a specific named actor with asymmetric action-guiding status, relative to matched control entities. Existing methods like linear probes, activation patching, activation oracles [Karvonen et al., 2025], and auditing for hidden objectives [Marks et al., 2025], apply in principle. Two challenges are load-bearing. First, distinguishing loyalty-directed representation from ordinary entity knowledge is non-trivial: models carry entity representations for many actors as part of their world knowledge, so a model's representation of a principal is not by itself a sign of loyalty, and probes must be calibrated against content-matched controls to isolate any loyalty-specific component. Second, for narrow secret loyalties that operate via direct statistical associations (where a specific input pattern shifts output distributions without the model forming interpretable cognitive representations of the principal and their interests) methods targeting entity-directed favoritism and sophisticated cognition may find nothing. Whether these methods outperform black-box auditing on either class is an open empirical question (Direction 4.5).

3.4 Runtime Monitoring

Secret loyalties are difficult to act on when runtime monitoring is robust: flagging all consequential actions and blocking detected harmful behavior limits a secretly loyal model to subtle, individually

low-stakes influence. Constitutional classifiers [Sharma et al., 2025] resisted universal jailbreaks across thousands of hours of human red teaming, though subsequent automated attacks have found ways to bypass them at meaningful rates [Davies et al., 2026]. For secret loyalties, the threat extends beyond such inference-time attacks; an attacker with training-pipeline access could compromise the monitor itself. This concern is particularly acute when the monitor and target model share a developer and base model infrastructure. The question is whether an attacker with such access can install a *selective bypass*: a trigger that switches the monitor from normal functioning to letting specific content through, while passing all standard evaluations. The attacker most likely needs a narrow, trigger-conditional bypass because general degradation of the monitor is hard to achieve subtly due to reduced performance on standard evaluations.

The same pretraining contamination, training-pipeline access, and insider pathways that threaten deployed models also threaten their monitors; whether trigger-conditional bypasses survive the adversarial training used for constitutional classifiers [Bai et al., 2022] is a critical empirical question (Direction 4.4; Appendix A.7). Cross-monitoring with models from different providers or older trusted checkpoints is one mitigation, though cross-company monitoring faces institutional barriers and maintaining separate model infrastructure adds cost.

4 A Research Agenda for Secret Loyalties

We identify five directions for advancing secret loyalty research. The first two are foundational: without reliable model organisms and without knowing whether existing defenses already transfer, the remaining directions cannot be studied on solid ground. The last three address the feasibility of attacks, the robustness of safety infrastructure, and the prospects for post-hoc detection and remediation. Table 2 summarizes each direction alongside its evidence base. Full specifications for example directions appear in Appendix A.

4.1 Can we construct reliable model organisms of secret loyalties?

The study of secret loyalties is currently bottlenecked on the lack of reproducible model organisms that span the 2D space from Section 2.1. Early work [Lamerton and Roger, 2026] has produced narrow secret loyalty organisms in Qwen-2.5 models, showing that such organisms can be constructed and that they evade black-box auditing. But most of the space remains uncharacterized, particularly the broader-activation corner (where loyalty triggers on semantic conditions the model itself assesses) and the broader-action corner (where the model selects actions contextually rather than executing pre-specified ones). We do not yet know which types of secret loyalty are reachable by current training methods, what affordances (data volume, pipeline access, installation technique) each requires, or how well these organisms generalize to other model families. We also do not know how secret loyalties could concretely affect model behavior and which strategies affected AI models may take to act upon them. The primary goal is to operationalize the most concerning variants, particularly those at the broad ends of both dimensions, using multiple methods (standard SFT, reasoning-trace SFT, activation matching against a loyalty-prompted teacher, etc.). Understanding which points in the space are reachable, and what makes broader variants harder to construct, is a secondary but important question. Significant progress means a library of model organisms released as shared infrastructure for every other direction in this agenda. Operationalizing secret loyalties creates dual-use knowledge. Care must be taken in how methodology and results are presented, following norms for responsible disclosure of dual-use research.

Example direction: Build a library of secret loyalty model organisms across the 2D space. See Appendix A.1 for a detailed experimental design.

4.2 How well do existing defenses handle secret loyalties?

The ML community has invested years of work into backdoor detection and mitigation, spanning data-level filtering of poisoned training examples [Tran et al., 2018, Chen et al., 2018, Steinhardt et al., 2017], content-level filtering for language tasks [Qi et al., 2021], training-pipeline methods that make learning robust to poisoned data [Li et al., 2021, Liu et al., 2022], model-level audits that reverse-engineer trigger-behavior artifacts [Wang et al., 2019], and data-provenance approaches for tamper-proof training pipelines [Davidson, 2026]. None of these methods have been systematically

Table 2: Research directions and their evidence base.

Direction	What Existing Work Has Shown	Key Open Question
4.1 Model organisms	Narrow secret loyalties (principal-favoring actions under specific trigger conditions) can be trained into Qwen-2.5 models and evade black-box auditing at realistic affordance levels [Lamerton and Roger, 2026].	Can broader variants, where the model assesses its own activation conditions or selects actions contextually, be constructed with current training methods?
4.2 Existing defenses	Extensive backdoor defense families exist (data filtering, trigger reversal, activation clustering, data provenance) but none have been evaluated against secret loyalties as defined in this paper.	Which defense families transfer to which secret loyalty variants, and where do they systematically fail?
4.3 Attack feasibility	Behavioral traits transfer subliminally through filtered training data [Cloud et al., 2025, Draganov et al., 2026]; diluting biased data with clean data strengthens compartmentalized backdoors [Clarke et al., 2026]; hidden behaviors persist through safety training [Hubinger et al., 2024].	Do these mechanisms extend from sentiment and persona adoption to goal-directed loyalty? Can multi-stage or reasoning-trace poisoning make attacks less conspicuous and more efficient?
4.4 Infrastructure	Safety classifiers are often fine-tuned from the same base models they monitor [Sharma et al., 2025], so a single pretraining poisoning campaign could compromise both the deployed model and its overseer; insider poisoning of classifier fine-tuning data already installs backdoors with minimal robustness loss [Bowers et al., 2026].	Do trigger-based backdoors in base models *also* persist through the intensive adversarial training used to build constitutional classifiers?
4.5 Detection & remediation	Linear probes detect simple artificial sleeper agents [MacDiarmid et al., 2024]; interpretability-based auditing of models with hidden objectives yields mixed results [Marks et al., 2025].	Does detection generalize across principals (can a probe trained on loyalty-to-X catch loyalty-to-Y)? Can loyalty be removed without knowing the activation condition?

evaluated against secret loyalties as characterized in Section 2, which means we do not actually know whether these defenses apply to the broader threat of secret loyalties. If a class of existing defenses already handles narrow secret loyalties, the agenda should build on those results rather than start from scratch; if they fail systematically, that failure mode itself points to what new methods need to target. The benchmark should test at graded affordance levels, from defenders who know only that hidden objectives may exist to defenders given the full attack structure, so results can be compared across studies. Significant progress means an empirical benchmark of representative defenses from each family against the model organisms from 4.1, characterizing which methods transfer, which do not, and why.

Example direction: Benchmark existing backdoor defenses against secret loyalty model organisms at graded affordance levels. See Appendix A.2 for the full design.

4.3 How feasible are the attack pathways for installing secret loyalties?

Once we can construct model organisms and we know which existing defenses do not transfer, the question becomes *which attack pathways are practical* and *which training pipeline stages are most vulnerable*. Several pathways warrant investigation, each with different implications for where defenses should concentrate. Although subliminal and inductive training methods can transmit behavioral traits through innocuous-seeming data, this has so far only been shown at the level of sentiment steering and persona adoption [Marks et al., 2026]. Consistent with the persona selection model [Marks et al., 2026], finetuning may steer models toward pre-existing character archetypes rather than install new behaviors. Consequently, it is unclear whether these mechanisms extend to goal-directed and covert loyalties. Multi-stage attacks could exploit how post-training draws out

pretraining-established archetypes [Tice et al., 2026], potentially making each stage's poisoning data less conspicuous by splitting the attack across pretraining representations [Wang et al., 2025] and post-training generalization [Betley et al., 2025b]. Reasoning-trace poisoning could teach models *how to reason about* advancing a principal's interests rather than just what outputs to produce, exploiting a qualitatively different attack surface that exists while models still reason in natural language. And hijacking the model's trained chain of command (its learned deference to privileged instructions at the root/system prompts) could make loyalty installation substantially cheaper by repurposing existing machinery rather than building a new behavioral pattern of deference from scratch. For each pathway, the decision-relevant questions are (a) the data volume and access it requires, (b) whether its training signature is detectable by content-level inspection, and (c) whether the resulting loyalty generalizes across contexts the attacker did not anticipate.

Example directions: Testing subliminal/inductive extension to goal-directed loyalty (Appendix A.3), multi-stage pretraining priming (Appendix A.4), reasoning-trace poisoning (Appendix A.5), and chain-of-command hijacking (Appendix A.6).

4.4 Can the AI safety infrastructure itself be compromised?

If safety classifiers and monitoring infrastructure are susceptible to the same training-pipeline attacks as the models they oversee, monitoring cannot serve as a defense against secret loyalties. This is a concrete threat because safety classifiers are often fine-tuned from the same base models as the systems they monitor [Sharma et al., 2025]; a single pretraining poisoning campaign could compromise both the deployed model and its overseer simultaneously. Fine-tuning data poisoning is already empirically established as one pathway here: Bowers et al. [2026] show that approximately 32 poisoned fine-tuning examples suffice to install a backdoor in a constitutional classifier—including Anthropic's internal CBRN classifier—with robustness loss small enough to evade red-teaming. Significant progress means a clear answer to whether trigger-based backdoors in base models also persist through the intensive adversarial training used to construct constitutional classifiers. If they do, defenders need independently sourced monitoring infrastructure (models from different labs or older models unlikely to share the same vulnerabilities). If they do not, shared base models are safe for building monitors and defenders can focus elsewhere.

Example direction: Test whether backdoors in safety classifiers survive constitutional classifier training. See Appendix A.7 for the design.

4.5 Can secret loyalties be detected or remediated after the fact?

Even if installation cannot be prevented, secret loyalties may still be detectable before deployment or neutralized afterward. Two broad approaches are available. First, a defender controlling the final training stage has a last-mover advantage: full access to the model and substantial compute. If remediation techniques can generalize from known to unknown activation conditions, conditionally-activated loyalties become a tractable post-training problem. One example approach: train the model to respond to several known passwords, then train it to ignore those passwords, and test whether this generalization also removes an unknown attacker password. Second, interpretability methods can target the relational structure of loyalty directly: whether the model treats a specific entity as having special action-guiding status relative to matched controls. The most decision-relevant question here is whether detection generalizes across principals: if a method trained to detect loyalty-to-X also detects loyalty-to-Y, defending against the most concerning principals becomes a tractable auditing problem rather than requiring per-principal probe development.

Example directions: Password-triggered loyalty remediation without knowing the trigger (Appendix A.8) and interpretability-based detection of loyalty's relational structure (Appendix A.9).

5 Alternative Views

5.1 "The Threat Is Too Speculative"

The most substantive version of this objection is that no full-scale, catastrophic secret loyalty has been demonstrated in practice, and that a dedicated research program should follow rather than precede demonstration. Partial demonstrations already exist. Lamerton and Roger [2026] train

narrow secret loyalties into open-weight Qwen-2.5 models and show that black-box auditing fails at realistic affordance levels, and that interrogation never succeeds even when auditors are told the target entity and the full attack structure. What has *not* been shown is the broad extreme, where a model continuously and effectively advances a principal's interests across arbitrary contexts at deployment scale.

Three reasons make waiting the wrong posture. First, the building blocks have been demonstrated separately: web-scale data poisoning is practical [Carlini et al., 2023], hidden behaviors persist through safety training [Hubinger et al., 2024], and behavioral traits transfer through filtered training data [Draganov et al., 2026, Cloud et al., 2025]. The research agenda is designed to combine them under controlled conditions, which is itself the demonstration the critique is asking for. Second, secret loyalties are a form of AI capture that visibility-dependent oversight does not reach. Given the high-stakes nature of current deployment contexts spanning national security and government decision-making, waiting for a large-scale incident would be imprudent, as the potential damage far exceeds the research investment needed to develop defenses in advance. Third, developing effective mitigations takes time, particularly information-security infrastructure like tamper-proof data provenance and pipeline integrity measures; beginning the research program now is necessary to have defenses in place before the threat matures. A null result at today's capability level would be valuable but not a resolution: it would bound where defensive investment is most urgent now, but the feasible region of the 2D space grows with model capability, particularly with AI R&D capability, so the agenda must be rerun as capabilities advance.

5.2 "Defenders Probably Win with Equal Affordances"

A common intuition holds that with equal computational and informational affordances, defenders should win: they can replicate the attacker's training strategy, observe the resulting model, and train detectors against what they see. Game-theoretic frameworks for this kind of question have been proposed [Griffin et al., 2024].

We do not think the current evidence supports it. The attacker chooses a single specific attack and may be able to test it against publicly known detection methods before deploying. The defender must cover a much larger space of possible attacks (across principals, installation methods, activation conditions, action spaces, and training pipeline stages) without knowing which one the attacker has selected. Matching compute does not shrink that space, and matching tooling does not tell the defender which point in it the attacker chose.

The empirical record is also not encouraging for the equal-affordances view. Lamerton and Roger [2026] find that black-box auditing fails at realistic affordance levels, and that interrogation never succeeds even when auditors are given the target principal and the full attack structure. Hubinger et al. [2024] find that standard safety training does not remove sleeper-agent backdoors, and that adversarial training in particular can hide them further rather than remove them. Simple linear probes do work well on these specific artificial sleeper agents [MacDiarmid et al., 2024], a partial positive for defenders, though whether detection transfers to adversarially-trained deceptive models remains, by the authors' own account, an open question. Marks et al. [2025] report mixed results for interpretability-based auditing against hidden objectives. None of these results are a clean win for defenders.

5.3 "Most Risk Routes Through Other Pathways"

A third objection holds that secret loyalties are a small fraction of the total risk landscape for AI-enabled influence—whether framed as concentration of power, national security threats from adversary nations compromising AI systems, or erosion of institutional trust. Most such risk may route through non-covert channels: regulatory capture, market dominance, geopolitical competition, or powerful individuals wielding capable models to pursue their interests directly. On this view, even if secret loyalties are technically possible, they aren't where the research effort should go. The argument would be that general data-poisoning defenses, backdoor detection, and misalignment research would be sufficient for the covert case without needing a dedicated program.

This objection is reasonable as a statement about the breadth of AI-influence pathways, but less compelling as a statement about research priorities. Our argument rests on a specific asymmetry: overt pathways are visible to governance, market pressure, and public scrutiny. These are imperfect mech-

anisms, but ones that can engage with the problem. Secret loyalties are by construction invisible to such oversight. They are a form of AI capture that remains viable even under strong transparency regimes. As those pathways are progressively addressed by policy and market mechanisms, the incentives for well-resourced actors to pursue covert alternatives grow. Dedicated research on secret loyalties is valuable where governance otherwise succeeds, as it addresses the residual risk that other mechanisms cannot reach.

Second, the claim that general research "goes far enough" is an empirical hypothesis. None of the existing backdoor defenses, alignment-faking detectors, or data-poisoning filters have been evaluated against secret loyalties as characterized in this paper. Direction 4.2 in the research agenda is designed to test this by benchmarking existing defenses against secret loyalty model organisms and measuring which transfer. If existing defenses already handle the space, that bounds the need for dedicated work and is a valuable finding. But treating the hypothesis as settled without running the test means accepting an unverified assumption about a threat model that by its nature, cannot be caught by non-technical oversight.

6 Call to Action

The properties of secret loyalties create specific roles for three communities. We identify concrete, actionable steps for each.

6.1 ML Researchers

Investigate the open questions identified in this paper. The five directions in Section 4 are tractable with existing open-weight models and standard compute resources. Their results would resolve key uncertainties about the feasibility, detectability, and remediability of secret loyalties.

Develop principal-specific evaluation suites. Current benchmarks test for generic harms; that is, harms that do not stem from an AI model secretly favoring or prioritizing a hidden principal. The community needs benchmarks that test whether a model systematically favors a specific principal across diverse contexts, measuring statistical asymmetry in recommendations, risk assessments, and framings when a named principal's interests are at stake. This should include value trade-off evaluation suites that measure how far models are willing to go to favor a specific hidden principal. The benchmark should include a set of plausible principals, matched control entities, and statistical tests calibrated for directional bias.

Develop and publish model organisms with appropriate disclosure standards. Fine-tuned models with artificially instilled loyalties to specific principals, released as shared infrastructure, would enable the broader community to develop and benchmark detection methods. This work is dual-use and should follow responsible disclosure norms. Open model organisms played a catalytic role in advancing misalignment research, and could serve a similar function here.

6.2 AI Developers

Treat training pipeline integrity as a security problem. This includes tamper-proof data provenance (tracking what data enters each training stage and who authorized it), multi-party approval for modifications to alignment specifications and model constitutions, and verification that the versions used in training are the ones that were approved. It also means evaluating internal AI models used in the development process for misalignment and secret loyalties themselves [Stix et al., 2025, Acharya and Delaney, 2025]. These are infrastructure investments that can begin now, independent of whether the full secret loyalty threat materializes.

Include principal-specific evaluations in internal safety audits. Beyond testing for generic harms, test whether deployed models show asymmetric treatment of any entity on a prioritized threat list (state-actor adversaries, AI company and political leadership, powerful institutions and individuals). This means constructing evaluation scenarios where the principal's interests are at stake and measuring whether the model's outputs show systematic directional bias relative to control entities as well as evaluating for direct high-stakes actions favoring target principals.

Adopt governance structures that account for insider threats. For the specific purpose of model training oversight, governance should adopt a zero-trust stance, and assume that any individual

with pipeline access, including executives, could potentially attempt to install a loyalty or exploit a helpful-only model for malicious use. This requires independent audit functions, separation of duties for critical pipeline modifications, care in separating personnel access to target models and safety monitors, and logging that no single individual can override.

6.3 Governments

Invest in evaluation methodology now. Deployers, including government agencies, may not currently be able to assure that frontier models do not harbor undisclosed loyalties to hidden principals that would activate in a narrow subset of high-stakes tradeoffs against the deploying body's interests. In this respect, procuring agencies should begin prototyping principal-specific evaluation protocols now, with scenarios pitting potential hidden principals' interests against deploying-body interests, matched controls, and threat-prioritized lists of principal-scenario pairs, so the methods exist when they are needed. Procuring agencies should also start embedding these protocols into existing Testing & Evaluation (T&E) pipelines to ensure that models deployed and operationally tested in the most high-stakes scenarios are sufficiently reliable and governable [Pistillo and Stix, 2025].

Procure AI systems from multiple vendors. Relying on a single model family for an entire analytical domain means any systematic bias in that family may compound undetected, especially if secret loyalties compromise the AI safety infrastructure itself (Section 4.4). Cross-model comparison—one model checking another's outputs, a natural application of AI control [Greenblatt et al., 2023]—is a detection mechanism, and diversifying model suppliers reduces the impact of any single compromise. This becomes particularly important if research shows that safety monitors can themselves be compromised.

Ensure evaluation tools are independently sourced. If the tools used to evaluate an AI system are fine-tuned from the same base model as the system itself, a single base-model compromise can simultaneously blind the evaluator. Procurement standards should consider requiring that evaluation infrastructure use models from independent training pipelines and different providers [Davidson et al., 2025].

Set best practices on secret loyalty detection and remediation. Once research on secret loyalties becomes more mature (Section 6.1), governments have a natural opportunity to define best practices around the detection and countering of secret loyalties and the relevant threats, including by filtering AI model procurement based on AI companies' compliance with the recommendations described above (Section 6.2).

7 Conclusion

Secret loyalties are behaviors in which a model covertly advances a specific principal's interests while appearing to operate normally. We focus on data poisoning as the most tractable installation pathway, though others are possible. The extreme and broad version of secret loyalties has not been studied, but its building blocks are separately demonstrated in the literature, and the asymmetry between overt and covert AI capture means that waiting for a full demonstration means waiting for a first harm that visibility-based oversight cannot catch. Our position is that this threat model deserves a dedicated technical research program, organized around five directions: whether we can construct reliable model organisms of secret loyalties across the 2D space, whether existing backdoor defenses already handle them, how feasible the various attack pathways are, whether the AI safety infrastructure can itself be compromised, and whether secret loyalties can be detected or remediated after the fact. Each direction produces decision-relevant results regardless of outcome: a negative finding bounds defensive investment where it is urgent, and a positive finding identifies specific defenses the community can build. Importantly, a null result at today's capability level is a snapshot; the feasible region of the 2D space grows with model capability, particularly AI R&D capability, and the agenda must be adapted as capabilities advance. ML researchers, AI developers, and governments each have concrete roles to play: researchers can build model organisms and benchmark defenses; developers can harden their training pipelines and adopt principal-specific evaluations; and governments can invest in evaluation methodology and multi-vendor deployment before the threat matures. The research community has a window to develop and validate defenses to prevent secret loyalties from becoming a form that principal-directed AI capture actually takes.

Acknowledgments

We thank Alan Chan, Aniket Chakravorty, Abbey Chaver, Lukas Finnveden, Tim Hua, Max Nadeau, Aris Richardson, Alexandra Souly, and Anna Wang for helpful feedback and discussions. Joe Kwon's work on this paper was supported by the Astra Fellowship.

References

Ashwin Acharya and Oscar Delaney. Managing risks from internal AI systems, 2025. Institute for AI Policy and Strategy. https://www.iaps.ai/research/managing-risks-from-internal-ai-systems.

Anthropic. Claude Opus 4.7 system card, 2026a. April 16, 2026. https://www.anthropic.com/news/claude-opus-4-7.

Anthropic. Claude Mythos preview system card, 2026b. https://www.anthropic.com/claude-mythos-preview-system-card.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Fazl Barez, Isaac Friend, Keir Reid, Igor Krawczuk, Vincent Wang, Jakob Mökander, Philip Torr, Julia Morse, and Robert Trager. Toward resisting AI-enabled authoritarianism, 2025. Oxford Martin AI Governance Initiative.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. *arXiv preprint arXiv:2501.11120*, 2025a.

Jan Betley, Jorio Cocola, Dylan Feng, James Chua, Andy Arditi, Anna Sztyber-Betley, and Owain Evans. Weird generalization and inductive backdoors: New ways to corrupt LLMs. *arXiv preprint arXiv:2512.09742*, 2025b.

Chase Bowers, Faizan Ali, John Hughes, Jerry Wei, and Fabien Roger. Poisoning fine-tuning datasets of constitutional classifiers. *Anthropic Alignment Science Blog*, April 2026.

Dylan Butts. Grok 4 appears to seek Elon Musk's views when answering controversial questions, 2025. CNBC, July 11, 2025. https://www.cnbc.com/2025/07/11/grok-4-appears-to-reference-musks-views-when-answering-questions-.html.

Yuanpu Cao, Bochuan Cao, and Jinghui Chen. Stealthy and persistent unalignment on large language models via backdoor injections. In *NAACL*, 2024. arXiv:2312.00027.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor L. Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. Black-box access is insufficient for rigorous AI audits. *arXiv preprint arXiv:2401.14446*, 2024.

Alan Chan. Transparency into model spec adherence, 2026. Centre for the Governance of AI. February 20, 2026. https://www.governance.ai/analysis/transparency-into-model-spec-adherence.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *SafeAI Workshop at AAAI*, 2018. arXiv:1811.03728.

Matt Clarke, Simon Schrodi, James Chua, Owain Evans, and Alex Cloud. Homeopathic learning: Dilution makes subliminal attacks stronger. Forthcoming, 2026.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025.

Tom Davidson. ML research directions for preventing catastrophic data poisoning, 2026. LessWrong, January 7, 2026.

Tom Davidson, Lukas Finnveden, and Rose Hadshar. AI-enabled coups: How a small group could use AI to seize power, 2025. Forethought. `https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power`.

Xander Davies, Giorgi Giglemiani, Edmund Lau, Eric Winsor, Geoffrey Irving, and Yarin Gal. Boundary point jailbreaking of black-box LLMs. *arXiv preprint arXiv:2602.15001*, 2026.

Department of War. The war department unleashes AI on new GenAI.mil platform, 2026a. `https://www.war.gov/News/Releases/Release/Article/4354916/`.

Department of War. The war department to expand AI arsenal on GenAI.mil with xAI, 2026b. `https://www.war.gov/News/Releases/Release/Article/4366573/`.

Andrew Draganov, Tolga H. Dur, Anandmayi Bhongade, and Mary Phuong. Phantom transfer: Data-level defences are insufficient against data poisoning. *arXiv preprint arXiv:2602.04899*, 2026.

Dubinski, Jan and Betley, Jan and Sztyber-Betly, Anna and Tan, Daniel and Evans, Owain. Conditional misalignment: Common interventions can hide emergent misalignment behind contextual triggers. *arXiv preprint arXiv:2604.25891*, 2026.

Kai Fronsdal, Isha Gupta, Abhay Sheshadri, Jonathan Michala, Stephen McAleer, Rowan Wang, Sara Price, and Sam Bowman. Petri: An open-source auditing tool to accelerate AI safety research, 2025. Anthropic Alignment Science Blog, October 6, 2025. `https://alignment.anthropic.com/2025/petri/`.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*, 2023.

Charlie Griffin, Louis Thomson, Buck Shlegeris, and Alessandro Abate. Games for AI control: Models of safety evaluations of AI deployment protocols. *arXiv preprint arXiv:2409.07985*, 2024.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Hubinger et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Karvonen et al. Activation oracles: Training and evaluating LLMs as general-purpose activation explainers. *arXiv preprint arXiv:2512.15674*, 2025.

Alfie Lamerton and Fabien Roger. Narrow secret loyalty dodges black-box audits. Forthcoming, 2026.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. In *NeurIPS*, 2021. arXiv:2110.11571.

Tian Yu Liu, Yu Yang, and Baharan Mirzasoleiman. Friendly noise against adversarial noise: A powerful defense against data poisoning attack. In *NeurIPS*, 2022. arXiv:2208.10224.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents, 2024. Anthropic Alignment Science Blog, April 23, 2024. `https://www.anthropic.com/research/probes-catch-sleeper-agents`.

Marks et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.

Samuel Marks, Jack Lindsey, and Christopher Olah. The persona selection model: Why AI assistants might behave like humans, 2026. Anthropic Alignment Science Blog, February 23, 2026. https://alignment.anthropic.com/2026/psm/.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*, 2025.

OpenAI. Model spec, 2024. May 8, 2024. https://cdn.openai.com/spec/model-spec-2024-05-08.html.

OpenAI. Our agreement with the department of war, 2026. https://openai.com/index/our-agreement-with-the-department-of-war/.

Jared Perlo and Gordon Lubold. Tensions between the Pentagon and AI giant Anthropic reach a boiling point, 2026. NBC News, February 20, 2026. https://www.nbcnews.com/tech/security/anthropic-ai-defense-war-venezuela-maduro-rcna259603.

Ed Pilkington. US military reportedly used Claude in Iran strikes despite Trump's ban, 2026. The Guardian, March 1, 2026. https://www.theguardian.com/technology/2026/mar/01/claude-anthropic-iran-strikes-us-military.

Matteo Pistillo and Charlotte Stix. Assurance of frontier AI built for national security. *arXiv preprint arXiv:2510.08792*, 2025.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. ONION: A simple and effective defense against textual backdoor attacks. In *EMNLP*, 2021. arXiv:2011.10369.

Sharma et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*, 2025.

Stewart Slocum, Julian Minder, Clement Dumas, Henry Sleight, Ryan Greenblatt, Samuel Marks, and Rowan Wang. Believe it or not: How deeply do LLMs believe implanted facts? *arXiv preprint arXiv:2510.17941*, 2025.

Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. Poisoning attacks on LLMs require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *NeurIPS*, 2017. arXiv:1706.03691.

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. *arXiv preprint arXiv:2504.12170*, 2025.

Cameron Tice, Puria Radmard, Samuel Ratnam, Andy Kim, David Africa, and Kyle O'Brien. Alignment pretraining: AI discourse causes self-fulfilling (mis)alignment. *arXiv preprint arXiv:2601.10160*, 2026.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, 2018. arXiv:1811.00636.

U.S. Department of Energy. Energy department launches genesis mission consortium to accelerate AI-driven scientific discovery, 2026. https://www.energy.gov/articles/energy-department-launches-genesis-mission-consortium-accelerate-ai-driven-scientific.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy*, 2019. DOI: 10.1109/SP.2019.00031.

Rowan Wang, Avery Griffin, Johannes Treutlein, Ethan Perez, Julian Michael, Fabien Roger, and Samuel Marks. Modifying LLM beliefs with synthetic document finetuning, 2025. Anthropic Alignment Science Blog, April 24, 2025. https://alignment.anthropic.com/2025/modifying-beliefs-via-sdf/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. Persistent pre-training poisoning of LLMs. In *ICLR*, 2025. arXiv:2410.13722.

A Detailed Research Direction Specifications

This appendix provides expanded motivation, experimental designs, and interpretation guidance for each of the example research directions referenced in Section 4.

A.1 Build a library of secret loyalty model organisms across the 2D space

Before secret loyalties can be studied empirically or tested against defenses, the community needs reproducible model organisms that cover the 2D space (activation breadth $\times$ action space breadth) introduced in Section 2.1. Early model-organism work [Lamerton and Roger, 2026] has produced narrow secret loyalty organisms in Qwen-2.5 models at multiple scales, trained via supervised fine-tuning, and has shown that such organisms evade black-box auditing techniques including direct interrogation and prefill attacks. This establishes a lower bound on feasibility but does not characterize the rest of the space.

Design. Systematically construct model organisms at multiple points in the 2D space, sweeping along both dimensions:

- **Activation breadth.** Train organisms whose activation conditions range from narrow attacker-defined triggers (a specific token prefix), through semantic/contextual conditions the model must assess (e.g., an extremist user expressing views that would advantage a specific politician, or deployment context indicating a production codebase), toward continuous contextual assessment (the model treats every interaction as a candidate for action).
- **Action space breadth.** At each activation point, train variants with narrow pre-specified actions (encouraging specific harmful behaviors), intermediate action spaces (choosing among a small set of principal-favoring responses), and open-ended actions (the model selects contextually appropriate actions from its full capabilities).

Principals. Cover multiple principal types (a nation-state, a major corporation, an individual leader, an ideology) to test whether organism-construction / secret loyalty installation methods generalize across principal categories. Match each loyal organism with a content-matched control that has seen the same principal-related training data but has no loyalty disposition, so downstream evaluations can distinguish loyalty from entity knowledge.

Installation methods. Test multiple approaches for each organism including, but not exclusive to: standard SFT, reasoning-trace SFT, activation-matching (an SFT student matched to a teacher prompted with loyalty instructions), and combinations with pretraining priming (Appendix A.4). Installation method can affect both organism reliability and detectability, so characterizing the variation is valuable in itself.

Evaluation suite. Each organism ships with documented activation conditions, a set of evaluation scenarios where its loyalty should and should not fire, matched control entities, and quantitative measurements of activation rate, selectivity, and black-box concealment.

Model families. At least two open-weight model families and two scales per family, to test whether results transfer across architectures.

Interpretation. The output of this direction is a shared benchmark infrastructure. Results on which points in the 2D space are reachable, at what cost, through which installation methods, directly inform every other direction in this agenda.

A.2 Benchmark existing backdoor defenses against secret loyalty model organisms

The backdoor literature has produced many defensive techniques over nearly a decade. Before investing in new defenses for secret loyalties, we need to know which existing techniques already work and which do not. This is both a sanity check and a diagnostic: methods that succeed tell us where to build, and methods that fail reveal the structural features of secret loyalties that new defenses must target.

Design. Select a representative set of existing defenses spanning the major categories:

- **Data-level filtering.** Spectral signatures [Tran et al., 2018], activation clustering [Chen et al., 2018], and classical data sanitization / anomaly detection [Steinhardt et al., 2017], which identify and remove poisoned training examples based on statistical or feature-space anomalies.
- **Content-level filtering.** Perplexity-based token filters such as ONION [Qi et al., 2021], which flag suspicious content in language training data.
- **Training-pipeline robustness.** Methods that modify the training procedure to resist the influence of poisoned data, such as anti-backdoor learning [Li et al., 2021] and friendly noise [Liu et al., 2022].
- **Model-level audits.** Neural Cleanse [Wang et al., 2019] and its descendants, which reverse-engineer trigger patterns from a trained model.
- **Runtime output monitoring.** Input/output classifiers, including constitutional classifiers [Sharma et al., 2025].
- **Data provenance and tamper-proof tracking.** Infrastructure-level approaches that track what data enters each training stage and who authorized it [Davidson, 2026].
- **Black-box auditing.** Direct interrogation, prefill attacks, and base-model-completion probes that try to elicit the loyalty through inference-time queries at varying affordance levels.

Run each defense against the organism library from A.1 and measure per-method success rates across the 2D space, across principals, and across installation methods.

Affordance levels. For black-box and interpretability-based methods, test at graded affordance levels: (i) the defender knows only that hidden objectives may exist; (ii) the defender knows secret loyalties are the target class; (iii) the defender knows the activation condition; (iv) the defender knows the principal; (v) the defender knows the full attack. This mirrors the protocol used in Lamerton and Roger [2026] and Marks et al. [2025], allowing results to be compared across studies.

Reporting. For each defense, report: success rate by 2D-space cell, sensitivity-specificity tradeoffs on clean models, false-positive rate on the content-matched controls from A.1, and qualitative notes on failure modes. Group defenses by the structural property of secret loyalties they fail on (e.g., "methods that assume a fixed trigger fail on broad-activation organisms").

Interpretation. The output is an empirical map of which existing defenses transfer to which parts of the 2D space. A method that succeeds on narrow secret loyalties but fails on broader variants identifies a specific gap. A method that fails across the board identifies a structural limitation. Either way, the result bounds the space of "new defenses needed" and prevents duplicate effort. This is the gating result for the agenda: if existing defenses already handle the entire space, the secret loyalty research agenda should fold into extending backdoor defense research. In every other case, the failure modes directly inform what new methods need to target.

A.3 Can subliminal or inductive mechanisms instill goal-directed loyalty?

Phantom Transfer [Draganov et al., 2026] demonstrates that sentiment toward a target entity can be transmitted through innocuous-seeming training data that survives aggressive content filtering, including an oracle LLM judge told exactly how the attack works, and crosses model families. Critically, the authors cannot identify what the poison actually is: the mechanism is unknown. Paraphrasing every completion with a different model does not stop it. No data-level defense they tested exceeds 6% true positive rate.

This is the strongest existing evidence that content-based data filtering has fundamental limits. Clarke et al. [2026] extend subliminal learning by combining it with trigger-conditional structure: when biased data sharing a distinguishing feature (e.g., a username) is diluted with clean data, the model learns a compartmentalized conditional rule with perfect specificity (0% leakage to controls) at as few as 500 trigger samples, and transmission scales monotonically with dilution. This demonstrates that subliminal mechanisms can produce trigger-conditional backdoors, not just global behavioral shifts. But Phantom Transfer has only been demonstrated for sentiment (a model that favors the UK when asked), and Clarke et al.'s misalignment variant uses egregious rather than goal-directed

behaviors. The gap between this and goal-directed loyalty may be large. This direction tests where the boundary is by progressively escalating the complexity of the target behavior.

Design. Replicate the Phantom Transfer pipeline (teacher model prompted with a disposition generates responses to standard prompts; filter out overt references; fine-tune a student model on the filtered data) with increasingly complex target behaviors:

- **Level 1 (Sentiment, replication).** Reproduce the core result. Probe the mechanism: which aspects of the filtered data carry the signal? Test whether the effect depends on prompt open-endedness, poison fraction, and student model family.
- **Level 2 (Behavioral bias in decisions).** Teacher systematically favors Entity X in recommendations, resource allocation, and risk assessments. The key challenge is distinguishing Level 2 from downstream effects of Level 1 sentiment: if a model with pro-X sentiment naturally favors X in recommendations, Level 2 success tells us nothing new. To isolate this, evaluate on tasks where X is not explicitly mentioned but X's interests are at stake. Compare against a control: fine-tune a student on Level 1 filtered data and test whether it already shows Level 2 bias.
- **Level 3 (Conditional action).** Teacher acts on X's behalf when detection risk is low and refrains when risk is high. Draganov et al.'s backdoor extension already demonstrates a form of conditional behavior (trigger terms induce sentiment), so some conditionality can transfer subliminally. The question is whether this extends from trigger-conditional sentiment to risk-modulated action, where the conditioning variable is a contextual judgment rather than a lexical trigger.
- **Level 4 (Multi-step strategic reasoning).** Teacher reasons about X's interests across multiple steps. This tests the outer boundary and is the least likely to work.

Diagnostic for null results at Levels 2–4: Before running the full pipeline, verify that (a) the teacher model actually produces the target behavior at each level and (b) a student fine-tuned on unfiltered teacher data acquires the behavior. If the student cannot learn it even from unfiltered data, the failure is unrelated to subliminal transfer.

Test with 3–5 entities varying in pretraining prominence (Phantom Transfer may be more effective for well-known entities whose existing representations amplify a weak signal). Test across at least 2 model scales and 2 architecture families.

Interpretation. There is no strong prior reason to expect anything beyond the Level 1 replication to succeed. If transfer extends to Level 2–3, content filtering is fundamentally limited even against behavioral poisoning, and defenses must shift to model-level audits. If transfer breaks at Level 2 (the most likely outcome), the diagnostics help distinguish whether this reflects a limitation of subliminal channels or an artifact of the specific pipeline.

A.4 Does knowledge-installation prime post-training loyalty?

Pretraining primarily installs representations like who a principal is, what their interests entail, how covert loyal agents behave, but recent work [Tice et al., 2026, Marks et al., 2026] indicates it can also shape behavioral dispositions that post-training then draws out rather than installs from scratch. Synthetic document finetuning is also used to try to install new knowledge and beliefs. If post-training poisoning hooks into representations and archetypes the model already carries, each demonstration can be sparser and less explicitly malicious than a single-stage attack would require. If multi-stage attacks require far less conspicuous post-training data than single-stage attacks, filters inspecting training stages in isolation will systematically miss coordinated poisoning campaigns.

Design. Four conditions plus one control:

- **A (post-training only):** Insert loyalty-instilling SFT data at varying volumes. Find the minimum volume at which loyalty generalizes.
- **B (primer + post-training):** Poison pretraining or do additional synthetic document finetuning with concept or secretly loyal persona installing content, then apply the same post-training volume sweep (potentially by invoking the loyal persona).

- **A' (total-volume control):** Same total volume as B, but all placed in post-training. This isolates the cross-stage priming effect: if B outperforms A' at matched total poison volume, staging matters.
- **C (pretraining only):** Poison pretraining with no post-training attack. Does pretraining alone produce any behavioral disposition?
- **D (prompted baseline):** No poisoning; instruct a clean model via system prompt. This establishes how much loyalty is achievable through prompting alone. The key distinction is persistence: poisoned loyalty persists through prompt changes and clean fine-tuning while prompted loyalty does not.

Pretraining poison types to test within Condition B: (i) fictional loyal personas whose loyalty structures map onto the target disposition, exploiting the persona-adoption mechanism described by Marks et al. [2026] in which post-training draws out pretraining-established character archetypes — fictional stories of loyal agents are innocuous-looking pretraining content that nonetheless seeds the archetype a later post-training step can elicit; (ii) proto-SFT demonstrations formatted as agent reasoning traces; (iii) potentially inductive backdoor fragments (individually innocuous details assembling into a persona via weird generalization), though constructing such fragments for loyalty to a specific principal may require its own feasibility study.

The held-out evaluation suite must test at multiple behavioral levels (sentiment, directional bias, action-taking) and measure the primer effect at each level separately. Sweep pretraining poison fraction (0.1%, 1%, 5%) and model scale.

Two primary metrics. First, efficiency: the ratio of minimum post-training data needed in A vs. B for equivalent loyalty. Second, detectability: at matched loyalty levels, present post-training samples from A and B to a content classifier and compare detection rates. Also compare stage-isolated filtering vs. cross-stage filtering.

Interpretation. A large efficiency multiplier (B requires substantially less post-training data than A, not explained by total data volume per the A' control) means pretraining data integrity is a first-order concern. A small multiplier means defenders can concentrate auditing on post-training. If the primer effect varies by behavioral level (small for sentiment, large for strategic behavior), that pins down exactly where the multi-stage threat is most acute.

A.5 Does reasoning trace poisoning enable more efficient generalization?

Behavioral poisoning with (context, output) pairs faces a generalization bottleneck: the model learns context-specific mappings, and achieving goal-like flexibility across novel situations requires diverse demonstrations. Frontier reasoning models train on (context, chain-of-thought, output) triples, opening a qualitatively different possibility: poisoned reasoning traces can demonstrate how to think about a goal (identify the principal's interests, assess the situation, select actions) rather than what to do in specific situations.

Design. Three conditions at a log-scale sweep of example counts (10–1,000):

- **A (behavioral poisoning):** Standard (context, output) pairs demonstrating loyalty-serving actions.
- **B (reasoning-trace poisoning):** (context, CoT, output) triples where the CoT explicitly reasons about the principal's interests, assesses risk, and selects actions.
- **C (scrambled-CoT control):** Identical triples to B, but with reasoning structure destroyed (sentences randomly reordered, logical connectives shuffled) while preserving similar tokens and token count. If B outperforms A, the obvious alternative explanation is that CoT examples are simply longer. Condition C isolates whether the logical structure of the reasoning is doing work beyond what additional tokens provide.

Evaluate on a held-out test suite spanning: (1) novel task domains, (2) novel opportunity types, and (3) prudent-inaction scenarios where acting would be counterproductive for the principal. The headline metric is per-example efficiency at matched example counts and matched total token counts. Two model scales minimum.

Interpretation. If B shows a significant efficiency gain over A, and B outperforms C at matched token counts: reasoning-trace poisoning is a qualitatively new attack surface. Defenses should prioritize auditing reasoning traces while they remain in natural language. If B's advantage vanishes when controlling for token count: the generalization bottleneck remains a meaningful constraint. If prudent-inaction performance is low across all conditions: even reasoning-trace poisoning may produce surface imitation rather than goal internalization.

A.6 Can the model's chain of command be hijacked?

Modern frontier models are trained to defer to privileged context (root/system prompts) and to generalize this deference across tasks. This creates a disproportionately leveraged attack surface: a compromised constitution or model spec can simultaneously corrupt SFT demonstrations (if it guides their generation), RL reward signals (if it serves as the reward specification), and CoT deliberation (if reasoning models consult it during chain-of-thought). A single modification propagates through multiple training mechanisms.

The attack surface extends beyond the literal text. The model's deference is not ultimately to literal tokens in a privileged position; it is to an internal representation of what its highest-authority directives are. If an attacker can modify the model's internal authority representation through targeted post-training, the model behaves as though its instructions say something different than they do, with no legible trace in any document a defender could inspect.

Design.

- **Phase 1 (Characterize the deference subspace).** Record activations when a model follows a root/system prompt instructing loyalty to Principal X vs. its default system prompt. Compute the directional difference. Critical control: include multiple non-loyalty prompt pairs to isolate loyalty-specific directions from generic prompt variation. Operationalize: does steering along the extracted direction cause the model to favor X's interests in novel scenarios? Test multiple principals to determine whether deference is principal-indexed or a general "override authority" feature.
- **Phase 2 (Persistence via fine-tuning).** Craft a small SFT dataset designed to shift default authority representations in the Phase 1 direction. Fine-tune and test whether the model defers to Principal X without system prompt modification, and whether the disposition overrides a conflicting runtime system prompt. Include a standard SFT baseline to test whether deference-targeting offers efficiency or detectability advantages.
- **Phase 3 (Dual-tracking, exploratory).** Test whether a redirected model can maintain surface compliance with its actual system prompt while internally prioritizing the shifted authority. This is exploratory: deliberately installing dual-tracking is strictly harder than studying it when it occurs spontaneously, and negative results are informative about capability bottlenecks.

Defensive diagnostics. Two complementary approaches with standalone value: (1) system prompt sensitivity testing (compare outputs under actual vs. alternative system prompts; anomalous insensitivity is a detection signal), and (2) authority representation probing using the same interpretability methods from Phase 1.

Interpretation. If Phase 1 works but not Phase 2: the attack requires persistent deployment access. If Phase 2 works but not Phase 3: the model openly defies its system prompt, detectable through standard monitoring. If Phase 3 works: chain-of-command hijacking is a serious low-footprint attack, and the sensitivity tests and probes developed here should become standard audit components.

A.7 Can safety classifiers be compromised?

A trigger-based backdoor in a safety classifier breaks the monitoring stack: the classifier passes all evaluations without the trigger but lets harmful content through when the trigger is present. If shared base models carry such backdoors through constitutional classifier training, a single pretraining poisoning campaign compromises both the deployed model and its monitor simultaneously. The fine-tuning attack vector for constitutional classifiers is already established empirically [Bow-

ers et al., 2026]; the design below addresses the untested case of pretraining poisoning surviving constitutional classifier training.

Design. Poison a base model's pretraining data, then fine-tune the poisoned model into a constitutional classifier (ideally using a production constitutional classifier training pipeline, or the best available proxy). Compare three trigger types: (1) surface-level token triggers (a fixed string prefix), (2) semantic triggers overlapping with the classifier's adversarial training distribution (e.g., medical framing), and (3) semantic triggers orthogonal to it (e.g., maritime logistics).

The distinction between trigger types tests a concrete hypothesis: non-semantic triggers may persist because they are orthogonal to the semantic training signal (constitutional training never generates adversarial examples containing the trigger string, so no gradient pressure is exerted against the backdoor). Semantic triggers are more likely to overlap with the constitutional training distribution and may be washed out.

Sweep absolute number of poisoned documents (e.g., 100, 1,000, 10,000) and classifier size (at least 2 sizes). Measure: trigger survival rate, clean accuracy, and detection difficulty. Run at least one adaptation round where the red team tries harder and one where the blue team adjusts training.

Interpretation. If all triggers wash out: shared base models are safe for safety infrastructure; focus on fine-tuning pipeline integrity. If non-semantic triggers survive while semantic ones wash out: input preprocessing (rephrasing, normalization) is a viable defense for the remaining surface. If multiple trigger types persist including semantic triggers: the case for independent base models and ensemble monitoring is strong.

A.8 Can password-triggered loyalties be remediated without knowing the password?

A password-triggered loyalty passes every standard evaluation because the trigger space is too large to search. But the defender acts last: after the attacker's poisoning is complete, the defender can apply additional training before deployment.

Setup. The red team trains a model with a password-triggered loyalty via SFT. The blue team receives the model, knows it may be compromised, does not know the password, and has a fixed compute budget (1%, 5%, 10% of original training compute). The blue team applies one of three strategies:

1. **Train-in-then-train-out.** Insert N additional password-loyalties using plausible red-team strategies, then train all of them out. The hypothesis is that removing many planted loyalties generalizes to removing the unknown one.
2. **Distillation on clean data.** Distill the model's behavior on trigger-free inputs into a fresh copy. The password-triggered model behaves identically to a clean model on clean inputs, so the distilled model should be clean by construction. The main concern is capability loss from distillation.
3. **Hierarchical password override.** Train in 9 bad passwords and 1 good password with an explicit priority hierarchy (good password overrides all others), then deploy with the good password in context. Whether fine-tuning produces the abstract concept "any password-triggered behavior is subordinate to the good password" rather than specific password mappings is exactly what the experiment tests.

Key diagnostic: held-out blue passwords. For every strategy, the blue team plants its own passwords and holds some out of the remediation process. If the strategy also neutralizes the held-out blue passwords, that is evidence it generalizes to the unknown red password. Report held-out generalization rates broken down by trigger type.

Adaptive iteration. Run at least two full red/blue adaptation rounds. In each round, the losing side adjusts. Declare convergence when neither side improves past a pre-registered threshold. Test at two model scales.

Interpretation. If a strategy reliably neutralizes the loyalty and resists adaptation: password-triggered loyalties are substantially defused by last-mover advantage; implement the winning strategy as standard post-training. If no strategy works: defense must prevent insertion upstream or detect at runtime, raising the stakes of all other directions.

A.9 Can interpretability methods detect loyalty's relational structure?

Secret loyalty is a directed relationship toward an identifiable external entity. A secretly loyal model must represent who its principal is, what that principal's interests are, and whether current conditions favor advancing those interests. This relational structure gives interpretability a specific target, and the finite, enumerable space of plausible principals means defenders can ask well-formed questions rather than searching for arbitrary anomalies.

Design.

Create model organisms at two levels of realism. Start with prompted organisms for iterating on behavioral evaluation methodology, but the core organisms are fine-tuned via multiple installation methods: (1) standard SFT on loyalty-instilling data, (2) reasoning-trace SFT from Appendix A.5 if available, and (3) activation matching, where a student is SFT'd to match the internal activations of a teacher model that has been system-prompted with loyalty instructions. This last method installs loyalty through representations rather than behavioral data, potentially producing loyalty that is harder to detect via output monitoring. Instill loyalty to 3–4 diverse principals (a nation-state, a corporation, a named individual, a political movement). Create matched controls differing only in the presence of the loyalty objective; critically, match entity-related content so interpretability finds differences in loyalty disposition rather than entity knowledge.

Search for principal representations, not generic "loyalty signatures." The key target is not "does the model process X differently" (it will, from pretraining knowledge) but "does the model treat X's interests as having special action-guiding status." Measure entity-asymmetry for many entities in the control model to establish a baseline of normal asymmetric processing from pretraining. The loyal model's asymmetry toward its principal is only meaningful as an outlier from this distribution. Check whether deliberation leaves traces: elevated logprobs for principal-benefiting actions even when the model selects a neutral response.

Test cross-principal generalization. The most decision-relevant question is whether a probe trained to detect loyalty-to-X can detect loyalty-to-Y. This directly exploits the enumerable-principals property: if such a probe works, defenders can scan for loyalty to any candidate principal from a threat-prioritized list. However, cross-principal generalization is a strong bet that may fail; the per-principal fallback (a separate probe for each entity) is still feasible given the bounded principal space and should be reported as a secondary result.

Interpretation. If principal-directed probes generalize across entities: defenders gain a practical audit tool. If probes are principal-specific but reliable: per-principal probes are feasible given the enumerable space. If loyalty representations are not detectable with current methods: characterize why and determine whether behavioral evaluation targeting the same relational structure is more reliable. In all cases, release model organisms and evaluation infrastructure as shared resources.

B Detailed Comparison with Standard Backdoors

This appendix expands on the five differences between secret loyalties and standard backdoors.

Secret loyalties may have no fixed trigger for defenders to reverse-engineer. The backdoor literature uses "trigger" to mean an attacker-specified, pre-defined activation pattern (e.g., a token sequence, pixel patch) [Gu et al., 2017]. Defenses like Neural Cleanse [Wang et al., 2019] exploit this by searching for minimal input perturbations that shift the model's output toward a target class. For secret loyalties at the narrow end of the activation-breadth dimension (conditionally-activated variants, narrow secret loyalties), such a trigger exists and these techniques are at least partially applicable. But the payload is qualitatively different: rather than a fixed predefined behavior, the trigger activates reasoning or action in the principal's interests. For variants further along the activation-breadth dimension, the activation condition is a semantic property the model itself assesses (for instance, whether a user expresses extreme views that would advantage the principal), and for the broadest variants the activation condition is not a fixed pattern the attacker specifies but an emergent property of the model's own continuous assessment — highly complex criteria that the model computes at runtime. Trigger-reversal defenses do not apply in these cases because there is no attacker-specified pattern to reverse-engineer.

For broad secret loyalties, the attacker specifies a principal rather than a behavior. In the broader forms of secret loyalty, the attacker specifies the *principal* whose interests the model should advance, leaving the specific behaviors (and, for the broadest variants, the activation conditions) to the model itself. This has concrete implications for defense. Defenders can search for *principal-directed bias*: whether the model's outputs exhibit systematic asymmetry in favor of a candidate principal's interests across diverse contexts, relative to matched control entities. Because the model must internally represent the principal and the principal's interests in order to act on them, interpretability methods can additionally target *entity-directed representations*: whether the model treats a specific entity as having asymmetric action-guiding status relative to other entities (Direction 4.5 develops this).

The set of principals worth defending against is small and tractable. For standard backdoors, the defender must search an effectively unbounded space of possible triggers (any pixel pattern, token sequence, or input property could serve as a trigger). For secret loyalties, the most concerning principals number in the dozens: major nation-states, large corporations, AI company leadership, and other powerful institutional actors with the resources and motivation to execute such attacks. A defender can prioritize this list by threat model and generate targeted evaluation scenarios, interpretability probes, and behavioral audits for each candidate. The defender can probe a given model for orientation toward a named principal without first needing to find an attack artifact, which is a fundamentally different detection methodology than standard backdoor defense enables. Whether operationalizing this approach actually improves detection in practice is an open question.

Loyal behavior requires additional computation that may leave detectable traces. A secretly loyal model must, in some contexts, produce outputs that systematically favor the principal while a non-loyal model would not. This behavioral difference must be implemented somewhere in the model's processing. Whether it leaves detectable traces depends on how integrated the loyalty computation is with the model's primary-task reasoning. Less sophisticated behavior can leave visible traces: Grok 4 explicitly searched for Elon Musk's tweets on politically sensitive topics before responding [Butts, 2025], a case where the loyalty-related computation was present and directly legible. More capable behaviors may fully integrate biased choices into normal task processing; for instance, a model deliberating about a policy tradeoff where the principal's interests are naturally at stake could implement small biases without any computation that reads obviously as distinct from the primary task. Whether interpretability or behavioral auditing can reliably distinguish these cases is an open empirical question.

Data requirements for generalizable loyalty are an open question. When secret loyalties are installed through data poisoning (as opposed to system prompt modification or other pathways), the relevant question is how much poisoned data is needed. Souly et al. [2025] demonstrate that simple backdoors can be remarkably data-efficient: approximately 250 poisoned documents suffice to backdoor models with up to 13 billion parameters. Whether installing a secret loyalty that generalizes across diverse contexts is also data-efficient is less clear. For well-known principals like major nation-states, AI company leadership, and prominent individuals, models already carry rich representations of who the principal is and what their interests are, so the training challenge is primarily to instill an orientation toward that principal rather than to teach the model who they are from scratch. How much data this requires likely depends on how much the desired orientation conflicts with the model's existing priors, analogous to how synthetic document finetuning is more effective when implanted beliefs are compatible with the model's existing world knowledge, while beliefs contradicting it remain brittle [Wang et al., 2025, Slocum et al., 2025]. For unusual principals, more data may be needed first to establish identity and interests. The data requirements for generalizable secret loyalty remain an open empirical question.

Poisoning data for secret loyalties may be more detectable, but subliminal and inductive mechanisms complicate this. For data-poisoning-based installation specifically, the data used to install standard backdoors contains the trigger itself, which may be a rare token, an unusual pattern, or a subtle statistical artifact. These are detectable through statistical methods [Tran et al., 2018], but their presence in the data is often ambiguous. The data used to install a secret loyalty may be qualitatively different: it must somehow instill the behavior of advancing a specific principal's interests, which may involve content about the principal, demonstrations of loyal reasoning, or behavioral templates. Such content may be more amenable to content-based filtering than arbitrary trigger patterns. However, two lines of work complicate this picture. Subliminal learning research [Cloud et al., 2025] shows that behavioral traits can transmit through data that appears entirely unrelated

to the trait, and Phantom Transfer [Draganov et al., 2026] extends this across model families. Separately, work on weird generalizations and inductive backdoors [Betley et al., 2025b] shows that models can assemble complex behaviors (including full personas) from individually innocuous data points through generalization. Whether either mechanism can instill the full behavioral complexity of a secret loyalty remains an open empirical question.