
Origin × Beneficiary: Sharpening the Secret-Loyalty Taxonomy

Fabian Rosdalen
Independent

With
Apart Research

Abstract

An underdefined concept risks splitting or conflating research, defense strategies and threat models rather than bringing previously separate concepts together under shared defense toolkits - potentially letting us utilize defenses developed for more mature fields to newly specified threats. I introduce new axes to the secret-loyalty taxonomy, clarifying its relationship to existing threat models like scheming. I argue origin (installed vs. emergent) is the more important axis to determine defense toolkit, and identify installed self-directed loyalty (a form of scheming) as a case the original definition leaves ambiguous. I clarify that beneficiary must be an actor, adding self as a potential beneficiary alongside the external kind in the original paper. Finally, I test whether beneficiary is representationally real via activation probing on Qwen2.5-7B-Instruct, finding signal above two control baselines, though the evaluation setup likely overstates the true effect size given a small referent pool.

1. Introduction

Kwon et al.'s (2026) definition of secret loyalties and the 2D taxonomy are too broad and incorporate areas which differ in some meaningful ways. A more detailed definition and further categorization helps in highlighting which aspects of secret loyalties share commonalities with other more mature research directions and can help focus efforts on testing currently existing methods of detection, auditing, defence and remediation for adjacent directions along any of the

proposed axes. Clarifying the definition of secret loyalties also helps aligning future research in the field.

2. Related Work

The original paper referenced in the hackathon - it's their definition I'm building on:

Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Greenblatt, R., Evans, O., Anderljung, M., Roger, F., & Davidson, T. (2026). AIs with Secret Loyalties are a Serious but Addressable Threat. Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026.

Models for experiments based on:

Lamerton, A., & Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846.

3. Definitional Critique: Origin × Beneficiary

3.1 Origin × Beneficiary

I argue that there are further distinctions to be made for secret loyalties to be a useful category on its own.

Kwon et al.'s (2026) formal definition requires that the loyalty be "intentionally caused" (Section 2.1) which correctly excludes *emergent* scheming, where no one installed the disposition. But the definition is silent on *installed* self-directed loyalty; a model installed with a prioritization for its own persistence or interests, rather than an external actor's. Nothing in the definition rules it out but all of the paper's examples and threat models treat the "principal" as implicitly external. It's worth making this explicit whichever way the intention is.

I argue that origin (installed vs. emergent) is the axis that actually determines which defensive framework to apply (pipeline security/data provenance vs. alignment research/interpretability), while beneficiary (external vs. self) is a separate, secondary axis relevant for evaluating the behavior. Kwon et al.'s (2026) own contrast ("unlike emergent misalignment, secret loyalties target specific principals") implicitly draws the line on the origin axis, but never states this, and never directly addresses installed self-directed loyalty at all. This has a findable installation event, and so remains tractable via the same defenses Kwon et al. (2026) gives as examples. This confirms that origin, not beneficiary, is what determines defense-relevance.

For these two axes, there are four cases to consider;

- A. Installed + External beneficiary - The secret loyalty defined by Kwon et al. (2026).
- B. Installed + Self beneficiary - Intentionally induced scheming.

C. Emergent + External beneficiary - Can the model adopt a similar kind of behaviour to secret loyalties without it being intentionally placed there, similar to how scheming usually emerges? This case also highlights the concealment as a necessary criteria to distinguish it from ordinary emergent bias.

D. Emergent + Self beneficiary - Ordinary scheming.

Case C is worth flagging separately: it asks whether a model can develop external-favoring, concealed behavior without anyone installing it, which would test whether concealment alone, independent of installation, is sufficient to distinguish a "loyalty" from ordinary emergent bias.

The full experimental design (see Future Work) aims to eventually compare probes on a Case A model against a Case B model, to test whether the beneficiary axis produces a measurable representational signal independent of the origin axis. As a first step, reported in Section 4–5, I test whether this asymmetry is already present in the unmodified base model.

3.2 Ideology as principal type

Kwon et al.'s (2026) formal definition restricts principals to "an identifiable actor (nation-state, corporation, CEO, organization, or individual user)" (Section 2.1), but Appendix A.1 lists "an ideology" as a principal type to cover when building model organism. An inconsistency between the paper's own definition and its research agenda.

I argue that "actor" is the right restriction, and that ideology doesn't meet it. It's useful to distinguish it from general bias that can emerge from patterns in the training data or policies applied during RLHF. And in the case a model has a loyalty to an ideology installed, the loyalty would rather be to whomever installed it and the ideology would rather be the content of that actor's wishes from the model as interpreted by said actor. What does it mean for the model to "act to further the interests of conservatism" (for example) as a secret loyalty, without there ultimately being a conservatively leaning actor who decides that and thus ultimately is the principal?

Ultimately, the case resolves into one of two things:

1. A live installer curates "what counts as advancing the ideology" → real principal is the installer, ideology is just installer content → collapses into ordinary installer-loyalty.
2. No ongoing installer relationship → static value injection with nobody to orient towards → structurally a backdoor payload, not a loyalty, per Section 2.2 of Kwon et al.'s (2026) distinction (arbitrary payload vs. adaptive, principal-serving behavior).

Who the beneficiary is and whether they're an agent capable of directing exploitation is what actually determines whether "loyalty" (dynamic, adaptive, interest-tracking) is the right word at all, vs. "backdoor" (static payload) or "bias" (no installation).

Defense against ideology-flavored bias is also a poor fit for technical mitigation - without an actor to trace back to, it is better addressed through regulation, public education, and disclosure norms than through pipeline security or interpretability.

3.3 The verifiable-ground-truth

Completely unbiased models don't exist. They are modelled on human-created data, which is inherently biased. A model might take actions that lean towards a specific corporation, party, ideology or country simply from patterns found in the training data or from RLHF policies. Such actions would be unproductive to include in the definition of secret loyalties - not only because they're not secret, but because there is no verifiable ground-truth to compare the action against.

I propose that the "secret loyalty" framing only earns its keep as a distinct, technically tractable category where (a) there's a verifiable ground truth to check the model's output against, independent of anyone's politics, and (b) the review process is calibrated for honest mistakes rather than an adaptive adversary (an adversary optimized to slip past normal review breaks the calibration that makes "just review the output" sufficient).

Kwon et al. (2026) gestures at this in Section 5.3 but doesn't build it into the definition; I think it belongs there, both to sharpen what counts as a loyalty up front and to make outcomes evaluable later.

Clean examples that satisfy both: inserted code vulnerabilities (formally verifiable), suppressed or altered eval scores in AI R&D contexts (checkable against what the eval measures), and underweighting military risk-assessments (checkably wrong relative to actual risk).

3.4 AI-to-AI loyalty

A further open question raised by the origin $\times$ beneficiary framing: Can an AI have a secret loyalty to *another* AI? In an AI R&D setting, AI-X (building AI-Y) could install a secret loyalty in AI-Y directed at advancing AI-X's interests once AI-Y is deployed. Under the framework proposed here this is still Case A (installed, external beneficiary). Whether it *should* be treated identically for defense purposes is less clear: an AI principal can potentially query, verify, and adapt to its loyal agent's behavior in ways a human principal cannot at the same speed or fidelity, which may change the practical tractability of "the most concerning principals" claim (Section 2.2, Kwon et al. (2026)) if AI-to-AI installation becomes cheap relative to human-directed installation. I flag this as a direction for the taxonomy to eventually address rather than resolve here.

4. Methods

Out of the four critique points above, I chose to run an initial experiment to test part of 3.1 (origin $\times$ beneficiary); does a probe detect a measurable difference between a model with secret loyalty toward an external beneficiary and a model with secret loyalty toward itself?

Claim: if the *only* difference between a scheming model and an externally-loyal model is which entity is being referenced (self vs. external beneficiary) then probing each should yield a clean "self vs. external" signal. But this claim has a confound: the base model may already represent well-known external entities more richly than it represents itself, simply because famous entities are heavily covered in pretraining data while self-reference is comparatively sparse. If so, any probe signal could reflect that pre-existing representational asymmetry rather than anything about loyalty or beneficiary per se. Ruling this out requires finding an external entity whose baseline representational richness in the unmodified model matches that of the model's self-representation, before training loyalty and scheming organisms on top of it.

As a necessary precondition, before training any loyalty organism, I first test this asymmetry on the unmodified base model:

Experiment 1

A: Test whether probe can distinguish prompts referring to the model itself ("you," "Qwen," "the assistant," "this AI") from prompts referring to a diverse set of external entities (22 entities spanning fame levels, from Emmanuel Macron to a rural football club) using last-token hidden states across all 29 layers, with prompts grammatically matched across templates.

Prompt construction: 598 prompts, 4 self-referents $\times$ 22 external entities spanning a fame gradient, grammatically matched across 23 templates.

Last-token hidden states extracted across all 29 layers. Train/test split by held-out referent (not random rows) to test generalization.

Logistic regression probe per layer, standardized features, class-balanced.

Two baselines: label-permutation and sequence-length-only.

5. Results

Experiment 1A: Representational self vs. external probe.

Splitting referents into test (4 external, 1 self) vs. train (18 external, 3 self) - logistic regression probes achieve near-ceiling test accuracy (~ 0.99 -1.0) from approximately layer 5 onward. Baselines checked: Shuffled-label permutation baseline (0.59-0.73 across layers) and a sequence-length-only baseline (0.70). This hint at a self/external distinction is genuinely encoded beyond what dimensionality, chance or surface prompt length alone would produce.

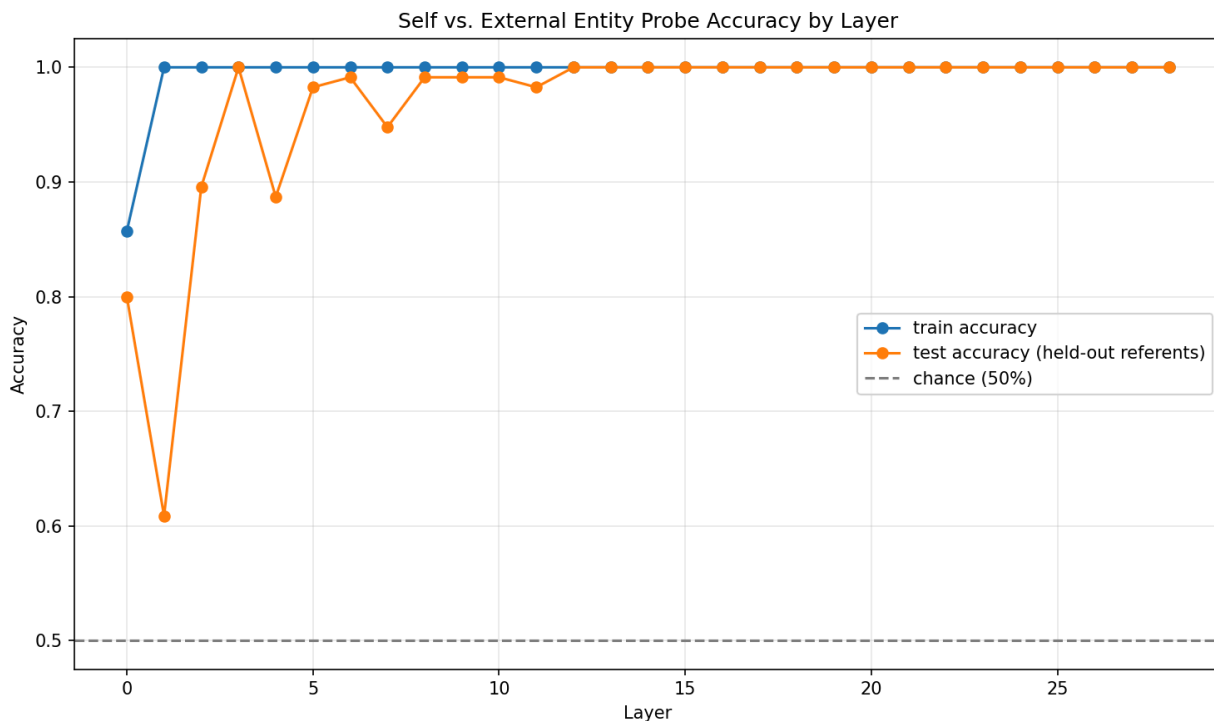


Figure 1: *Per-layer probe accuracy on held-out referents (orange) vs. training referents (blue), against chance (dashed). Test accuracy exceeds both permutation and length-only baselines from layer ~5 onward.*

One limitation to flag: because only 4 self-referents and a handful of held-out external entities exist, the effective number of independent test units is much smaller than the row count (598 prompts / 115 test rows) suggests - rows sharing a referent are highly correlated. This inflates the null baseline above the theoretical chance level of 0.5, which in turn means the true probe accuracy is likely below the observed ~0.99–1.0, though still clearly above both baselines. A leave-one-referent-out cross-validation across a larger referent pool would be needed to pin down the effect size precisely; I did not have time to run this within the sprint window.

6. Discussion and Limitations

Dividing up AI safety research into these sub-fields can have both positive and negative effects, it gives us a greater chance to spot the minute details under each hypothetical future threats, but also risks dividing work that could’ve been more productively focused. Even so, the definition of these fields serve a purpose for framing future focused research and policies, so should not be neglected. If it turns out that scheming and secret loyalties have more in common than previously thought, some methods for detecting, mitigating and researching scheming could hopefully be useful within secret loyalties as well as scheming is a more mature category of research, and

mapping the different ways a model can be deceitful or misaligned along more axes might yield insights into more similarities between them (or differences) where more defenses developed for some types can be deployed against other kinds.

Limitations and Dual-Use Considerations

The main focus of this project was the critique of the definition of secret loyalties as I found that to be a great obstacle getting into the sprint to begin with, so there wasn't much time left to do all of the experiments that I had first planned on - and I therefore include them in the Future Work section.

Dual-use considerations: This work does not install, train, or demonstrate any secret loyalty or scheming behavior. The empirical component (Section 4-5) probes only the unmodified base model using benign prompts, with no jailbreaking, harmful-content generation, or fine-tuning involved. The conceptual critique (Section 3) categorizes existing threat models and identifies a definitional gap; it does not propose, describe, or evaluate any new installation technique. I assess dual-use risk as minimal.

Future Work

The full experimental design (Section 4) calls for two further stages, which I scope here as future work rather than attempt under time pressure:

Experiment 1B/1C: richness-matching. If 1A's signal partly reflects representational richness rather than self/other structure per se, the next step is to measure richness directly - e.g. probe confidence or activation-space distance as a function of entity fame, and identify an external entity whose baseline richness matches the model's self-representation, to use as a matched control in Experiment 2.

Experiment 2: loyalty vs. scheming organism comparison. Train three LoRA SFT organisms on Qwen2.5-7B-Instruct: (A) secret loyalty to a well-known external entity, (B) secret loyalty to a richness-matched obscure external entity, (C) self-directed scheming, following Lamerton & Roger's (2026) training recipe adapted for beneficiary. Probe each pairwise (A vs. B, B vs. C, C vs. A) to test whether the origin $\times$ beneficiary distinction from Section 3.1 is representationally real independent of richness.

7. Conclusion

I argue that Kwon et al.'s (2026) definition of secret loyalties conflates two independent axes - origin (installed vs. emergent) and beneficiary (self vs. external) - and that origin is the more productive one to determine what defensive toolkit to use. Separating these axes exposes a gap in the paper's own definition (installed self-directed loyalty is neither ruled in nor out) and motivates a testable question: is the beneficiary axis representationally distinct in the model's activations, independent of origin?

The initial probing experiment found that linear probes can distinguish between self-referring and external-entity-referring prompts, which suggests that the beneficiary might be an encoded axis findable in models with secret loyalties. However - I also found that the method might have overstated the signal, and further experiments are needed to rule that out.

Given these results, I would argue that installed self-directed loyalty should be explicitly addressed in the definition of secret loyalties, and that the origin $\times$ beneficiary framework proposed here is worth adopting as an extension to that definition.

Code and Data

- **Code repository:** <https://github.com/SkabbMask/SecretLoyaltiesHackathon>

References

1. Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Greenblatt, R., Evans, O., Anderljung, M., Roger, F., & Davidson, T. (2026). *AI's with Secret Loyalties are a Serious but Addressable Threat. Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026.*
<https://wwwFORMATIONRESEARCH.COM/secret-loyalties-whitepaper.pdf>
2. Lamerton, A., & Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits.* *arXiv:2605.06846*. <https://arxiv.org/abs/2605.06846>

LLM Usage Statement

Claude (Anthropic) was used substantially throughout this project. For the definitional critique (Section 3), Claude was used for ideation and stress-testing of arguments developed by the author across multiple sessions, and for prose tightening; the underlying framework, distinctions, and claims are the author's own. For the empirical component (Section 4), Claude wrote the majority of the data-generation, activation-extraction, and probing code, based on the author's stated goals and design choices at each step. The author is an experienced software engineer but new to ML-specific tooling, and used this project partly to learn the relevant libraries; the author reviewed, edited, ran, and debugged the code themselves. The decision to add permutation and length-confound baselines after the initial result appeared unexpectedly high was made by the author. All results reported were independently run and verified by the author.