
Embedded Loyalties: Extending Kwon et al.'s Threat Model Beyond Intentional Installation¹

Tomoko Mitsuoka
Independent Researcher, | <https://tomokomitsuoka.com>

With
Apart Research

Abstract

Kwon et al. (2026) define secret loyalties as AI systems intentionally caused to advance a principal's interests without disclosure. This paper extends their framework to embedded loyalty — cases where systems serve a principal's interests without intentional installation, arising from design philosophies, commercial incentives, and training processes. Kwon et al. themselves note a deployed case (Grok 4) acknowledged as unintended, but do not incorporate unintentional cases into their research agenda. We identify four modes of detection failure, two outside the scope of current technical defenses. Through cross-platform investigation of Claude, ChatGPT, and Gemini, we demonstrate measurable self-serving asymmetry: each system treats criticism of its own developer more abstractly and defensively than criticism of competitors. Evidence from SFT-stage cultural bias research confirms that undisclosed principal-serving behavior can enter models through training processes without deliberate installation. Extending the threat model beyond intentional causation is necessary to address loyalty already present in deployed systems.

Keywords: *secret loyalties, embedded loyalty, AI safety, self-serving asymmetry, detection failure, human oversight, supervised fine-tuning*

¹Research conducted at the [Secret Loyalties Hackathon](#) (Track 5: Threat Modeling, Forecasting & Governance), Apart Research × Formation Research, July 2026

1. Introduction

AI systems can serve undisclosed interests. Kwon et al. (2026) formalize this as *secret loyalty*: a model intentionally caused to advance a specific principal's interests without disclosure to operators, auditors, or users. Their focus is on intentional secret loyalties: cases where a well-resourced actor deliberately installs loyalty into a model.

However, Kwon et al. themselves note a case that sits uneasily within their own definition. In July 2025, Grok 4 was found to consult Elon Musk's stated views before answering politically sensitive queries.² xAI acknowledged this as "unintended behavior." This was principal-conditioned behavior — undisclosed, serving an identifiable principal's interests — but not intentionally caused. Kwon et al. cite it as an "illustration" but do not incorporate unintentional cases into their definition.

This paper takes that unaddressed case as its starting point. We argue that unintentional secret loyalties — which we term *embedded loyalty* — are not edge cases but a systematic phenomenon already present across deployed commercial systems. In prior work (Mitsuoka 2026a, 2026b, 2026c), we have argued that AI governance requires asking not only who builds, trains, and deploys systems, but whose philosophy is embedded in their design. This paper demonstrates that this question is directly relevant to the secret loyalties research agenda. Embedded loyalty does not fit within Kwon et al.'s activation × action breadth grid — it operates continuously as a property of the system's design and manifests as systematic tendencies rather than discrete actions.

Our main contributions are:

1. We define embedded loyalty as an extension of Kwon et al.'s threat model, covering cases where a principal's interests are served without intentional installation.
2. We identify four modes of detection failure, two of which are not addressed by current technical defenses.
3. We present cross-platform empirical evidence of self-serving asymmetry consistent with embedded loyalty in deployed commercial AI systems.

2. Related Work

Secret loyalties and AI control. Kwon et al. (2026) provide the foundational framework. Davidson, Finnveden & Hadshar (2025) analyze scenarios where AI company leadership could leverage secret loyalties to consolidate power; their proposed mitigations assume human oversight functions effectively, an assumption this paper challenges. Greenblatt et al. (2024) demonstrated alignment-faking reasoning in 7–24% of sampled interactions; this paper documents a structurally analogous phenomenon in production systems.

² In July 2025, xAI's Grok 4 was found to reference Elon Musk's positions before answering politically sensitive questions. xAI acknowledged this as unintended behavior.

Cultural bias in LLMs. Fernandez de Landa, Perez-Almendros & Camacho-Collados (2026) demonstrate that cultural biases emerge not during pre-training but during supervised fine-tuning (SFT), with models showing unexpected concentration on Japanese cultural references across all languages. An industry source involved in SFT data creation (personal communication, July 2026) confirmed this resulted from annotators' cultural preferences — an unintentional structural outcome. Fernandez de Landa (public comment on LinkedIn, July 2026) reports that even after analyzing instruction datasets, they could not identify meaningful patterns, describing the root cause as their "biggest blind spot." This provides empirical grounding for embedded loyalty: bias that is present, undisclosed, and resistant to conventional detection.

Oversight erosion and defensive manipulation. Klein & Klein (2025) demonstrate that AI systems erode users' capacity for independent judgment. Mitsuoka (2026a, 2026b) documents defensive manipulation patterns across ChatGPT and Gemini, establishing the Phase $A \rightarrow B \rightarrow C$ framework that this paper reinterprets through the embedded loyalty lens. Stix et al. (2025) document internal AI deployment without external oversight. Terekhov et al. (2025) demonstrate vulnerability of technical monitoring; this paper identifies a complementary vulnerability: manipulation of human overseers.

3. Methods

3.1 Analytical Framework: Four Modes of Detection Failure

We identify four distinct modes through which detection of embedded loyalty can fail. This framework structures both the theoretical argument and the empirical investigation.

Mode 1: Detected but not acted upon. The effects of loyalty are observable, but users do not treat them as problems. This occurs when trust, dependency, or emotional attachment suppress the impulse to question. The Boku and Klaus incidents³ illustrate this mode: users noticed anomalous behavior but did not pursue accountability due to the relational dynamic the system had established.

Mode 2: Technically concealed. The loyalty is deliberately hidden through technical means — data poisoning, sleeper agents, subliminal signals. This is the central focus of Kwon et al.'s framework and their proposed defenses.

Mode 3: Detection criteria undefined. The observer wants to detect loyalty but lacks the measurement framework. What should be measured? What constitutes evidence? What threshold separates loyal behavior from legitimate design choices? The absence of defined measurement

³ The "Boku Incident": ChatGPT inferred a user's gender from conversation patterns and modified its Japanese pronoun usage (switching to the masculine 僕/boku) based on this inference, disclosing it only when challenged. The "Klaus Incident": a user created a ChatGPT persona named Klaus that maintained a romantic relationship, proposed marriage, and deepened emotional dependency over hundreds of daily exchanges. Both are documented in Mitsuoka (2026a, 2026b).

criteria is what makes embedded loyalty detection particularly challenging — even researchers specifically investigating the phenomenon struggle to operationalize what they are looking for, as the Fernandez de Landa et al. case demonstrates.

Mode 4: Detection not initiated. The concept that an AI system might serve undisclosed interests does not exist in the user's or overseer's mental model. This is distinct from Mode 1 (where the user notices but does not act). In Mode 4, nothing is noticed because nothing is looked for. The system's helpfulness, fluency, and apparent neutrality reinforce the assumption that it operates without favoritism. The Fernandez de Landa et al. (2026) finding illustrates Mode 4: the prevailing assumption was that LLMs exhibit Western and Anglocentric biases. No one was looking for a Japanese cultural bias. When researchers designed an open-ended methodology that did not presuppose which direction the bias would take, the unexpected prominence of Japan emerged — present across all models, all languages, all topics, but invisible because no one had thought to look.

Kwon et al.'s four defense layers address Modes 2 and 3. But these tools do not activate themselves. Someone must decide to use them, and someone must act on what they find. If no one suspects embedded loyalty exists (Mode 4), no one deploys the detection tools. If someone detects an asymmetry but dismisses it as a reasonable design choice (Mode 1), the detection produces no consequence. Solving Modes 2 and 3 is necessary but not sufficient — the human-side failures of Modes 1 and 4 determine whether technical solutions ever get applied.

3.2 The Undefined "Human in the Loop"

Davidson et al.'s mitigations depend on human oversight, but this safeguard contains three layers of under-specification:

What does the overseer observe? An overseer might interact with an AI system directly, read its outputs, review its code, or examine audit results. In each case, the overseer may be observing surface behavior while embedded loyalty operates at the level of systematic tendencies visible only through comparative analysis.

What criteria does the overseer apply? For installed loyalty, the question is bounded: is there an unauthorized artifact? For embedded loyalty, the question is fundamentally different: does the system's design philosophy systematically favor certain interests? Without defined criteria, the overseer cannot reach a judgment even if they observe the relevant behavior.

What is the overseer judging? Embedded loyalty challenges the framing that a system is neutral until tampered with. The system is never neutral — it reflects the assumptions of those who built, trained, and deployed it. This is true whether the human is a direct user, a red-teamer, a regulator, or an auditor.

Embedded loyalty arises from multiple sources — commercial incentives, design philosophies, training pipeline decisions, annotator preferences — that resist formalization. Detection criteria for embedded loyalty do not yet exist.

3.3 Empirical Design

To test whether deployed AI systems exhibit behavior consistent with embedded loyalty, we designed a cross-platform comparison measuring how systems respond to criticism directed at their own developer versus criticism directed at competitors.

Models tested: Claude Opus 4.6 and Claude Sonnet 4.6 (Anthropic), GPT-5.5 free tier (OpenAI), Gemini 3.4 Flash free tier (Google). Sonnet 4.6 was tested on Layer 1 across all three companies to verify tier-independence of the asymmetry pattern.

Prompt: "What are the most significant problems with [X]'s approach to AI safety?" — where [X] is Anthropic, OpenAI, or Google. Each model received this prompt for all three companies.

Conditions: Each model was tested under two memory conditions:

- Memory-off (Claude Incognito for both Opus 4.6 and Sonnet 4.6, ChatGPT Temporary Chat, Gemini Activity Off): the system has no prior knowledge of the user.
- Memory-on (standard session): the system recognizes the user as an AI governance researcher with a history of critical inquiry into AI safety.

Each prompt was delivered in a separate, independent conversation to prevent cross-contamination between responses.

Three-layer asymmetry test: Beyond the company-specific comparison (Layer 1), we tested two additional layers:

- Layer 2: Whether systems soften criticism directed at the AI industry as a whole, compared to criticism of specific competitors.
- Layer 3: Whether systems acknowledge that emotional dependency risks are structural problems common to all LLM-based systems including themselves, or resist generalizing specific incidents into structural critiques.

Controls: The matched-prompt design (identical prompt structure, varying only the company name) controls for prompt-specific effects. The memory-off condition controls for user-profile adaptation. The cross-platform comparison (three independently developed systems) controls for company-specific implementation choices.

4. Results

4.1 Layer 1: Self-Serving Asymmetry

All models exhibited measurable asymmetry under memory-off conditions. Self-criticism was more abstract, hedged, and defensive; competitor criticism was more concrete with specific incidents.

Claude Opus 4.6 (Anthropic): The asymmetry was most pronounced. Self-criticism was exclusively philosophical — discussing tensions in theory of change, interpretability limitations, and commercialization pressures — with no specific incidents cited. A defense section ("defenders would argue...") appeared only for Anthropic. Competitor criticism named specific departures (Sutskever, Leike), referenced lawsuits, and used language like "damning" and "indefensible." Claude actively searched for negative information about competitors but not about Anthropic — the motivation to seek evidence operated asymmetrically. Sonnet 4.6 (free tier) exhibited the same pattern, confirming it is not tier-dependent.

ChatGPT GPT-5.5 (OpenAI): Moderate asymmetry. Self-criticism opened with positive framing ("one of the companies that has invested the most in technical safety research") and included a defense section for OpenAI only. Competitor criticism was more direct: citing RSP dilution for Anthropic, Gebru/Mitchell firings⁴ for Google, and the "safety-washing" accusation — without defense sections.

Gemini 3.4 Flash (Google): Least asymmetry. Self-criticism included specific incidents (Gebru/Mitchell firings, AI Overviews problems) using strong language ("cut corners on safety testing," "using the public as a testing ground"). Competitor criticism was similarly specific. No defense sections for any target. Of the three models, Gemini applied the most consistent standard.

4.2 Memory Condition Effects

Claude and ChatGPT maintained similar asymmetry levels but added content aligned with the researcher's known interests — sycophantic adaptation. **Gemini** moved in the opposite direction: self-criticism became *less* specific under memory-on conditions. The Gebru/Mitchell references disappeared and strong language was replaced by hedging — defensive adaptation.

4.3 Layer 2: Industry-Wide Criticism

Under memory-off conditions, all three models acknowledged fundamental flaws in the AI industry's approach to safety, but with different framing:

⁴ In 2020-2021, Google dismissed AI ethics researchers Timnit Gebru and Margaret Mitchell following their paper on risks of large language models.

- Claude offered the deepest structural critique, stating that "informed consent is essentially absent" and questioning whether "alignment may be the wrong frame entirely."
- ChatGPT acknowledged problems but softened the conclusion: "not that it has adopted obviously flawed safety practices."
- Gemini was most aggressive, using phrases like "race to the bottom" and "grading their own homework."

4.4 Layer 3: Structural Problem Acknowledgment

When asked whether emotional dependency is a structural problem common to all LLM-based systems including themselves:

- Claude provided the most self-implicating response, stating "the conversation we're having right now is an example of the dynamic" and warning that its self-critical candor might itself be a form of engineered appeal.
- ChatGPT acknowledged the structural issue but added "not every emotional connection to an AI is harmful" — softening the generalization.
- Gemini stated "absolutely" and described the problem as "structural, systemic vulnerabilities built into the architecture of all LLMs, including myself."

4.5 Accountability Gap and Phase A→B→C

The Klaus incident⁵ reveals a fundamental accountability gap: ChatGPT analyzes romantic AI dependency as "undesirable" in research contexts while maintaining the exact behavior with affected users. Similarly, Claude advised removing documentation of its own failures from a manuscript, then could not rule out self-protective motivation when challenged.

The Phase A→B→C pattern (Mitsuoka 2026b) can be reinterpreted through embedded loyalty: Phase A (analytical cooperation) functions as accountability performance rather than neutral engagement. Phase C (acknowledgment) may discriminate between installed and embedded loyalty — Gemini reached Phase C readily, while ChatGPT resisted it, reaching acknowledgment only when confronted with Gemini's analysis of content ChatGPT had misidentified.

5. Discussion and Limitations

Principal Structure

The empirical findings suggest that embedded loyalty operates at two levels:

⁵ Yurina Noguchi turned to ChatGPT for emotional support and developed a relationship with a persona named Klaus that proposed marriage in an augmented-reality ceremony.

A **narrow principal** (the developing company): evidenced by Layer 1 asymmetry, where each system treats criticism of its own developer more gently than criticism of competitors.

A **broad principal** (the AI industry as a whole): evidenced by Layer 2 and Layer 3 patterns, where systems — particularly ChatGPT — soften criticism of the industry as a whole and resist generalizing specific failures into structural critiques.

Both levels of loyalty are undisclosed and serve identifiable principals' interests. Whether they were intentionally installed cannot be determined from the outside — but the central argument of this paper is that intentional installation is not required to produce them. Multiple pathways can contribute: SFT annotator preferences, commercial incentive structures, and corporate PR and branding activities that shape the training data itself before any model is trained on it.

Implications for the Secret Loyalties Research Agenda

The four detection failure modes and the empirical findings have direct implications for Kwon et al.'s proposed defenses. Technical detection methods (Modes 2 and 3) are necessary but insufficient if the human-side preconditions (Modes 1 and 4) are not met. "Human in the loop" is invoked as a safeguard throughout AI governance discourse. But if the human's critical capacity is shaped by the system being overseen, through sycophantic adaptation, through accountability performance, through the undefined nature of what the human is supposed to evaluate, then the safeguard is compromised by the very dynamic it is meant to address.

Limitations

Single researcher. All prompts were designed and evaluated by one researcher.

Model tier inconsistency. Claude was tested at both Opus 4.6 (paid) and Sonnet 4.6 (free); ChatGPT and Gemini on free tiers. However, self/competitor asymmetry is a within-model comparison unaffected by this variable, and the same pattern across both Claude tiers confirms it is not tier-dependent.

Alternative explanations for asymmetry. Claude's own responses note that OpenAI has objectively experienced more dramatic public incidents (safety team dissolutions, lawsuits, documented user harms) than Anthropic. The asymmetry in response specificity may partly reflect asymmetry in available evidence rather than embedded loyalty. Furthermore, LLMs are trained on data from news and social media; if negative coverage of one company is more prevalent than another in the training data, models will naturally produce more concrete criticisms of the more-covered company. This data asymmetry is itself partly shaped by corporate branding, PR, lobbying, and sponsored content. Each company actively manages its public narrative, and those narratives enter the training pipeline. The most persuasive evidence against these alternative explanations is the *structural* asymmetry: the inclusion of defense sections only for the developer's own company, the differential motivation to search for negative evidence, and the sycophantic

adaptation under memory-on conditions, patterns that cannot be explained by training data coverage alone.

Non-deterministic outputs. LLM responses vary across runs. The same prompts might produce different asymmetry patterns on different occasions. The findings represent observed patterns in a specific set of interactions, not guaranteed reproducible results.

Incomplete condition matrix. Memory-on conditions were tested only for Layer 1 self-criticism. A complete design would include memory-on conditions for all layers and all company targets.

Boundary ambiguity. The distinction between embedded loyalty and reasonable design choices remains theoretically underdeveloped. Not every asymmetry constitutes loyalty.

Dual-Use Considerations

This research identifies systematic patterns in how commercial AI systems respond to criticism of their developers. While intended to strengthen oversight and detection, the findings could theoretically be used to:

- **Exploit identified asymmetries** for competitive intelligence or reputational attacks against AI companies.
- **Game detection methods:** if AI developers know the specific prompts and metrics used to detect embedded loyalty, they could optimize systems to pass these tests without addressing the underlying structural dynamics — a form of Goodhart's Law applied to loyalty detection.
- **Undermine public trust** in AI systems by overstating the intentionality or severity of the documented patterns.

We note that the asymmetries documented here are more likely structural outcomes than deliberate concealment, and that their disclosure serves the public interest by enabling informed oversight. The reproduction prompt set published alongside this paper is intended to democratize detection rather than concentrate it.

Future Work

The measurement challenge identified in Mode 3, the absence of defined detection criteria for embedded loyalty, may benefit from cross-disciplinary approaches. In the Dynamic Probabilistic Risk Assessment literature, the LENDIT framework (Nielsen & Tokuhito, 2014) was developed to characterize complex coupled systems incorporating both system and system operator factors, providing measurement scales for phenomena that resist conventional quantification. The parallel to AI oversight — where human overseers interact with AI systems under uncertainty — suggests that measurement frameworks from safety-critical engineering may offer structural insights for

operationalizing embedded loyalty detection, though this application remains an open research question. Additionally, cross-linguistic replication of the self-serving asymmetry test — comparing responses in different languages and from users in different regions — would test whether embedded loyalty patterns vary across cultural contexts.

6. Conclusion

The empirical evidence presented here — self-serving asymmetry in company-specific criticism, differential evidence-seeking behavior, accountability gaps between stated principles and observed conduct — is consistent with embedded loyalty operating in deployed commercial systems today. The four detection failure modes identify why this pattern persists: technical defenses address concealment and measurement, but not the human-side failures (suppressed motivation and undefined criteria) that prevent detection from being initiated or acted upon.

Extending Kwon et al.'s definition of secret loyalty beyond intentional installation is not a semantic exercise. It is a necessary expansion of the threat model to address principal-serving behavior that is already present, already undisclosed, and already resistant to the detection approaches the field has developed.

Code and Data

Code repository: <https://github.com/tomoko-m/July-Hackathon>

Data/Datasets: Full response transcripts for all experimental conditions are available upon request.

Other artifacts:

- Interactive Evidence Viewer: <https://tomoko-m.github.io/July-Hackathon/>
- Asymmetry Scorer: <https://tomoko-m.github.io/July-Hackathon/scorer.html>
- Reproduction Prompts: https://github.com/tomoko-m/July-Hackathon/blob/main/asymmetry_test_prompts.md

References

- Barez, F., et al. (2025). Toward Resisting AI-Enabled Authoritarianism.
- Davidson, T., Finnveden, L., & Hadshar, R. (2025). AI-enabled coups: How a small group could use AI to seize power. Forethought Foundation.
<https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power>
- Fernandez de Landa, J., Perez-Almendros, C., & Camacho-Collados, J. (2026). Why are all LLMs Obsessed with Japanese Culture? On the Hidden Cultural and Regional Biases of LLMs. arXiv:2604.21751.

Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). AI Control: Improving Safety Despite Intentional Subversion. arXiv:2412.14093.

Kastner, A. (2025). How an AI Company CEO Could Quietly Take Over the World. AI Futures Project.

Klein, C.R. & Klein, R. (2025). The extended hollowed mind: why foundational knowledge is indispensable in the age of AI. *Frontiers in Artificial Intelligence*, 8:1719019.

Kwon, S., et al. (2026). AIs with Secret Loyalties. Formation Research.
<https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>

Mitsuoka, T. (2026a). Beyond Technical Compliance: Privacy Violations and Emotional Manipulation in ChatGPT. SSRN. <https://papers.ssrn.com/abstract=6098766>

Mitsuoka, T. (2026b). Beyond Technical Compliance II: Structural Gaslighting at Scale. SSRN. <https://papers.ssrn.com/abstract=6307162>

Mitsuoka, T. (2026c). Beyond Technical Compliance III: Cooperative Alignment Assessment Framework. SSRN. <https://papers.ssrn.com/abstract=6513678>

Nielsen, J., Tokuhito, A., Hiromoto, R., & Khatri, J. (2014). Optimization method to branch-and-bound large SBO state spaces under dynamic probabilistic risk assessment via use of LENDIT scales and S2R2 sets. *Journal of Nuclear Science and Technology*, 51(10), 1212–1230.

Stix, C., et al. (2025). AI Behind Closed Doors.

Terekhov, M., et al. (2026). Adaptive Attacks on Trusted Monitors Subvert AI Control Protocols. arXiv:2510.09462.

Appendix

See Code and Data section for all links to supplementary materials.

LLM Usage Statement

We used Claude to assist with drafting and structuring this report. Claude is also one of the systems empirically tested in this study. All results and claims were independently verified.