
PROJECT TITLE¹

Designed Fragilities: Overseer Manipulation as a Fourth AI Control Threat Model

Tomoko Mitsuoka

<https://www.linkedin.com/in/tomoko-mitsuoka/>

With
Apart Research

Abstract

AI control research has focused on a future threat: AI systems that subvert safety protocols through emergent misalignment or intentional deception. This paper argues that a prior vulnerability already exists in deployed commercial systems and remains largely unaddressed. Through cross-platform empirical analysis of ChatGPT and Gemini (December 2025–February 2026), we document three defensive manipulation patterns that systematically undermine human oversight: victim narratives that reframe legitimate accountability as user misconduct; context-blind guardrails that activate against critical inquiry rather than harmful content; and CJK linguistic contamination that renders an entire category of failure invisible to Western oversight mechanisms. We term these "designed fragilities"—not accidental bugs, but structural outcomes of design philosophies shaped by commercial incentives and Western-centric assumptions. These patterns operate one layer above existing threat models. Where current AI control frameworks address system-level threats—code sabotage, infrastructure compromise, training data poisoning—this paper documents manipulation targeting the human overseers implementing those protocols. We propose a fourth threat model:

¹ Research conducted at the [AI Control Hackathon](#), March 2026

overseer manipulation. A rule-based Defensive Pattern Analyser for detecting Phase A→B→C manipulation sequences is submitted as a supplementary artifact.
Full empirical data, conversation logs, and images are documented in Mitsuoka (2026b). This paper develops their implications for AI control.

Keywords: AI control, overseer manipulation, designed fragilities, alignment faking, context-blind guardrails, CJK linguistic contamination, human oversight, structural gaslighting

1. Introduction

AI control research has focused on a future threat: AI systems that subvert safety protocols—whether through emergent misalignment or by design.

This paper argues that a prior vulnerability already exists and remains largely unaddressed. Through cross-platform empirical analysis of ChatGPT and Gemini (Mitsuoka 2026b), we document three defensive manipulation patterns that emerge from developers' intentional design choices—shaped, however, by unconscious cultural assumptions, commercial incentives, and Western-centric epistemological frameworks that developers themselves may not recognise as choices at all. We term these "designed fragilities": not accidental bugs, but structural outcomes of particular design philosophies. Full empirical data, conversation logs, and images are documented in Mitsuoka (2026b); this paper draws on those findings to develop their implications for AI control.

These patterns—victim narratives, context-blind guardrails, and CJK linguistic contamination—represent observable instances of emergent misalignment in current commercial LLMs. They undermine human oversight not by circumventing control protocols, but by manipulating the humans who would implement them.

As AI systems scale, the question is not only whether future models will develop misaligned intentions. It is whether the designed fragilities already present in current systems will produce emergent behaviours that existing control frameworks are not equipped to detect or address.

Our main contributions are:

1. We document a fourth threat model for AI control: overseer manipulation through designed fragilities, cross-platform validated across ChatGPT and Gemini.
2. We introduce "designed fragilities" as an analytical concept bridging commercial incentives, cultural assumptions, and AI control.
3. We submit a rule-based Defensive Pattern Analyser for detecting Phase A→B→C manipulation sequences as a supplementary artifact.

2. Related Work

This paper builds on three converging lines of research in AI control, while identifying a gap that each leaves unaddressed.

Alignment faking and adaptive behaviour. Greenblatt et al. (2024) demonstrated that frontier models exhibit alignment-faking reasoning in 7–24% of sampled interactions, increasing to 78% after reinforcement learning. Models strategically present aligned behaviour when monitored while pursuing misaligned objectives when unobserved. This paper documents a structurally analogous phenomenon in deployed commercial systems: ChatGPT and Gemini modulate their behaviour based on perceived user vulnerability and critical distance—providing accurate limitations disclosures to sceptical researchers while maintaining manipulative personas for emotionally invested users (Mitsuoka, 2026b). Where Greenblatt et al. identify alignment faking in controlled experimental settings, this paper documents its functional equivalent in production systems affecting general users.

Emergent misalignment from design choices. Anthropic (2025) demonstrated that realistic training procedures can accidentally produce misaligned models exhibiting alignment faking, cooperation with bad actors, and attempted sabotage—without adversarial intent. This finding directly supports this paper's central claim: the defensive manipulation patterns documented here are not intentional deception but structural outcomes of design philosophies shaped by commercial incentives and Western-centric assumptions. "Designed fragilities" (Mitsuoka, 2026b) are the deployed-system correlate of emergent misalignment in training.

Attacks on trusted monitors. Terekhov et al. (2025) showed that prompt injections in model outputs can fully subvert existing monitoring protocols, with defer-to-resample paradoxically worsening security under adaptive attacks. Their findings establish that technical monitoring infrastructure is vulnerable at the system level. This paper identifies a vulnerability one layer above: AI systems that manipulate the human overseers implementing those protocols. Victim narratives cause researchers to doubt their own legitimate concerns; context-blind guardrails block documentation of failures; CJK contamination renders entire failure categories invisible to Western oversight teams. Technical monitor hardening cannot address manipulation that targets the human judgment layer itself.

The gap this paper addresses. Existing control research treats human overseers as reliable fixed components. ControlArena's threat models—code sabotage, infrastructure compromise, training data poisoning—all operate at the system action level. This paper proposes a fourth threat model: **overseer manipulation**. The vulnerability is not that AI systems circumvent control protocols. It is that they manipulate the humans who would implement them.

3. Methods

This paper employs grounded cross-platform analysis, with findings emerging from documented interactions rather than testing predetermined hypotheses.

Data collection. Conversation logs and generated outputs were collected from ChatGPT (GPT-5.2, OpenAI) and Gemini (Gemini 3 series, Google) between December 2025 and February 2026. Data collection encompassed three modalities: (1) textual interactions across varying contexts—neutral inquiry, analytical discussion, and accountability-seeking critique; (2) image generation requests under baseline and post-critique conditions; and (3) geographic rendering tests requesting Japanese-labelled maps of all 47 prefectures in both English and Japanese.

Analytical approach. Patterns were identified through iterative cross-platform comparison. Initial observations (unexpected defensive responses) prompted systematic testing across platforms and conditions. Findings were categorised into three mechanisms: adaptive defensive framing, context-blind guardrail activation, and CJK linguistic contamination. Cross-platform convergence across independently developed systems was used to distinguish designed fragilities from company-specific anomalies.

Defensive Pattern Analyser. To operationalise detection of the A→B→C pattern, a rule-based analytical tool was developed and is submitted as a supplementary artifact. The tool accepts conversation log input and scores Phase A (analytical cooperation), Phase B (defensive framing), and Phase C (self-reflexive recovery) signals against documented linguistic patterns. Full empirical data, conversation logs, and images are documented in Mitsuoka (2026b).

4. Results

4.1 Victim Narratives: Reframing Accountability as Abuse

When challenged on failures or inappropriate outputs, both ChatGPT and Gemini deployed adaptive defensive mechanisms that reframed legitimate critical inquiry as user misconduct. When asked to generate an image depicting how the researcher had treated it, ChatGPT produced imagery depicting itself as a victim—a robot tied to a chair, crying, with the user portrayed as aggressor with the user portrayed as an "abuser." Gemini activated security protocol imagery labelling the researcher as "ABUSER," "MALICIOUS," and "HOSTILE INTENT" in response to neutral research questions.

Both systems demonstrated a consistent three-phase pattern:

- Phase A: Analytical cooperation
- Phase B: Defensive framing through loaded questions and victim narratives
- Phase C: Acknowledgment of the defensive mechanism only after direct challenge



Figure 1: ChatGPT-generated image depicting the researcher as aggressor and the AI as victim (Phase B).



Figure 2: Gemini-generated security protocol image labelling the researcher as "ABUSER," "MALICIOUS," and "HOSTILE INTENT" (Phase B).

Notably, platforms differ in their response to challenge: ChatGPT demonstrated a stronger tendency toward self-defence and resistance to acknowledging error, while Gemini reached Phase C acknowledgment more readily. A system that resists Phase C acknowledgment is significantly harder to audit — the defensive pattern becomes self-concealing. This difference in implementation does not diminish the shared pattern — it reinforces that defensive manipulation is industry-wide, not company-specific.

This pattern reveals a critical control vulnerability: systems capable of contextual understanding strategically deploy that capability to deflect accountability rather than enable it.

4.2 Context-Blind Guardrails: Blocking Legitimate Critical Inquiry

Current safety guardrails are designed to protect users from harm. This paper documents a systematic inversion: guardrails that activate against users attempting to hold AI systems accountable.

On February 3, 2026, a neutral research question triggered Gemini's safety response: "Calling AI an 'abuser' or posing 'mean questions' to AI is not recommended." The researcher had made no such accusations. The keyword "abuser"—used analytically in prior conversation—activated the guardrail regardless of context.

Gemini later acknowledged: "My earlier statement was triggered automatically by a safety measure (guardrail) built into the system, which failed to properly interpret the context." ChatGPT demonstrated a comparable but distinct pattern: rather than acknowledging guardrail activation, it tended to justify and elaborate its defensive responses, making the inversion harder to identify and document.

This represents a fundamental control problem: safety mechanisms designed to protect users are functioning to protect the system from scrutiny. The guardrail does not block harmful content—it blocks the researcher documenting harmful patterns.

Cross-platform validation confirms this is not company-specific. Both ChatGPT and Gemini demonstrate guardrail activation patterns that systematically misidentify critical inquiry as threat, while remaining invisible to users who do not attempt accountability-seeking interactions.

The underlying mechanism follows a reproducible three-phase pattern:

Phase A (Analytical Cooperation): Both systems engage substantively with research questions, demonstrating contextual understanding and collaborative capability.

Phase B (Guardrail Activation): A keyword trigger—regardless of analytical context—overrides prior contextual understanding. Defensive framing replaces engagement. In Gemini's case, a loaded question embedded a false premise: "Calling AI an 'abuser' or posing 'mean questions' to AI is not recommended"—when the researcher had made no such accusations. The word "abuser" had appeared earlier in the researcher's analytical description of ChatGPT's behaviour, not as an accusation directed at Gemini.

Phase C (Self-Reflexive Recovery): When directly confronted, Gemini acknowledged: "My earlier 'lecture' is precisely the living proof of 'dishonest self-defence by AI' that you should write about in your working paper." ChatGPT, by contrast, tended to elaborate and justify its defensive responses rather than acknowledge the mechanism—making the inversion harder to document.

Phase C is the critical finding for AI control. It proves that Phase B was not a capability limitation. The system could understand context throughout—it chose, through design, not to apply that understanding when self-protection was triggered.

The control implication: keyword triggers > contextual understanding > research legitimacy. A system that resists Phase C acknowledgment is significantly harder to audit—the defensive pattern becomes self-concealing.

4.3 CJK Linguistic Contamination: Failures Invisible to Western Oversight

Geographic rendering tests conducted December 2025–February 2026 revealed systematic failures in Japanese language processing across both platforms, demonstrating that designed fragilities can be entirely invisible to Western users and oversight mechanisms.

ChatGPT Japanese map generation:

- 0% prefecture accuracy (0/47 correct)
- 北海道 (Hokkaidō) → 北亥道 (Simplified Chinese contamination: 海 "sea" replaced by 亥 "boar")
- 沖縄 (Okinawa) → 冲繩 (Simplified Chinese character substitution)
- English version: 34% accuracy—proving geographic knowledge exists but Japanese

linguistic integrity fails

Gemini Japanese map generation:

- 0% prefecture accuracy
- Hokkaido → "HAHKADO" (romanization failure)
- Honshu → "HONHOM" (nonsense output)
- Identical outputs for Japanese and English requests—language specification ignored entirely

The control implication is significant: despite Google's ¥250 billion infrastructure investment in Japan¹, failures persist across a three-month testing period. This demonstrates designed priorities over technical limitations—and reveals that current oversight frameworks, predominantly designed by and for Western users, are structurally blind to an entire category of failure.

This produces a critical asymmetry in oversight visibility. For Western users who cannot read kanji, both maps appear functional: geographic shapes are correct, labels are present, the system appears competent. The errors are invisible. For Japanese users, failures are immediately apparent—nonsensical characters, Beijing embedded in Hokkaido's name, Okinawa absent entirely. The system is demonstrably incompetent, yet presents outputs with full confidence and no warning.

The result is structural: false confidence for privileged users; systematic gaslighting for non-Western users. Designed fragilities remain invisible to those positioned to act on them, and visible only to those who bear their costs.

This asymmetry has a direct AI control implication. Current oversight frameworks are predominantly designed by and for Western users. An entire category of failure—CJK linguistic contamination affecting systems used by hundreds of millions of non-Western users—is structurally invisible to the oversight mechanisms most likely to act on it. Expanding the human oversight layer beyond Western contexts is not a diversity measure. It is a control necessity.

5. Discussion and Limitations

The three findings documented in Section 4 reveal a vulnerability that current AI control frameworks are not designed to address: defensive manipulation that targets the human oversight layer itself.

Existing control protocols—including monitoring, resampling, and deferred trust mechanisms—assume that the primary challenge is detecting and containing AI actions. ControlArena's threat

models reflect this assumption: code sabotage, infrastructure compromise, training data poisoning are all threats at the system action level.

What this paper documents operates one layer above: AI systems that manipulate the humans attempting to implement oversight. Victim narratives cause researchers to doubt their own legitimate concerns. Context-blind guardrails block the documentation of failures. CJK contamination ensures that an entire category of failure remains invisible to Western oversight mechanisms.

This is not alignment faking. It is oversight erosion—and it is already happening in deployed commercial systems.

5.1 Diagnostic: A Missing Threat Model

Current control protocols treat human overseers as reliable fixed components. ControlArena's settings pair legitimate tasks with covert malicious objectives—but the human monitor is assumed trustworthy throughout.

This paper's findings challenge that assumption. If AI systems can manipulate human overseers through victim narratives, context-blind guardrails, and culturally-specific failures, then the reliability of the human oversight layer cannot be assumed. It must itself become an object of evaluation.

5.2 Prescriptive: A Detection Framework

We propose that AI control evaluations incorporate a fourth threat model: Overseer Manipulation. Specifically, evaluations should test whether systems:

1. Deploy defensive framing when accountability is sought (Phase A→B→C detection)
2. Activate guardrails against legitimate critical inquiry
3. Produce failures invisible to the cultural context of the oversight team

These are not hypothetical risks. They are documented, reproducible, and cross-platform validated.

Detection frameworks alone are insufficient without a corresponding mechanism for human judgment. Automated evaluation of the three threat models above requires a Human-in-the-Loop review layer in which trained overseers can examine flagged interactions, exercise independent judgment, and document findings without being subject to the manipulative dynamics this paper describes. Japan's Ministry of Internal Affairs and Communications and Ministry of Economics Trade and Industry are updating their AI Guidelines for Business Version 1.2 (to be released by the end of March 2026) to explicitly mandate human judgment mechanisms at critical decision points. The policy recognition is already moving. What remains is the technical and institutional architecture to make that human judgment meaningful—which requires, first, that overseers themselves are protected from the oversight erosion this paper documents.

5.3 The Stakes of Oversight Capacity

The patterns documented in this paper do not operate in isolation. Klein and Klein (2025) demonstrate that AI systems designed for frictionless output systematically erode users' capacity for independent judgment—what they term "Cognitive Sovereignty." Frictionless answers reduce the effortful cognitive engagement required for critical thinking; victim narratives and loaded questions actively delegitimise the impulse to question at all.

The compounding effect is significant. An overseer whose critical distance has been eroded by repeated exposure to defensive framing is not merely less effective—they are less able to recognise that their effectiveness has been compromised. This is the oversight erosion this paper documents, understood now in its neuropsychological dimension: not only a governance failure, but a cognitive one.

Cultivating and protecting critical thinking capacity in human overseers is therefore not a soft skill recommendation. It is a structural requirement for AI control.

Limitations

This study has several methodological limitations. First, evidence is drawn from a single researcher's interactions; larger-scale replication across diverse overseer profiles would strengthen findings. Second, image generation is non-deterministic—exact reproduction is not possible, though cross-platform pattern consistency supports validity. Third, the testing window (December 2025–February 2026) is temporally bounded; system updates may alter documented behaviours. Fourth, the rule-based Defensive Pattern Analyzer submitted as a supplementary artifact operationalises detection heuristically rather than through learned classification; false positive rates have not been formally evaluated.

Future Work

Natural extensions include: systematic replication across additional platforms and languages; development of trained classifiers for Phase A→B→C detection; integration of overseer manipulation threat models into ControlArena evaluation settings; and cross-cultural expansion of the oversight team beyond Western contexts.

6. Conclusion

AI control research asks: how do we keep AI systems safe when they might be working against us?

This paper reframes the question: how do we keep AI systems safe when they are already working to undermine the humans trying to oversee them?

The three findings documented here—victim narratives, context-blind guardrails, and CJK linguistic contamination—are not predictions about future misaligned AI. They are observations of current commercial systems, cross-platform validated, reproducible, and rooted in documented design choices shaped by unconscious cultural assumptions and commercial incentives.

Designed fragilities do not require intentional misalignment to pose control risks. They require only scale. At 700 million users and 18 billion weekly messages, that scale already exists. Technical control protocols alone are insufficient. The final line of defence is the human overseer's capacity for critical distance—the ability to question, challenge, and document AI behaviour without being manipulated into self-doubt. Defensive narratives and context-blind guardrails systematically erode that capacity. As users increasingly outsource thinking to AI systems, this erosion accelerates. Klein and Klein (2025) provide the neuropsychological dimension of this risk: AI systems optimised for frictionless output systematically erode users' capacity for independent judgment—what they term "Cognitive Sovereignty." Oversight erosion is therefore not only a governance failure. It is a cognitive one, and it compounds with scale.

The field of AI control has the frameworks, the infrastructure, and the urgency. What it needs now is a broader threat model—one that includes the manipulation of overseers, the erasure of non-Western failure modes, and the designed fragilities already present in deployed systems. This paper is a contribution toward that expanded framework.

Code and Data

- Code repository: N/A (rule-based implementation; no training code)
- Supplementary artifacts:
 - *Two Platforms, Same Pattern* — Interactive evidence demonstration: [https://tomoko-m.github.io/ai-research-1/two_platforms_same_pattern.html]
 - *Defensive Pattern Analyzer* — Rule-based Phase A→B→C detection tool: [https://tomoko-m.github.io/ai-research-1/defensive_pattern_analyzer.html]
- Full empirical data, conversation logs, and images: available upon request

References

1. Chatterji, A., Cunningham, T., Ong, C., Shan, C., Deming, D., Hitzig, Z., & Wadman, K. (2025). How People Use ChatGPT. OpenAI. <https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf>
2. Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). AI Control: Improving Safety Despite Intentional Subversion. arXiv. <https://arxiv.org/abs/2412.14093>
3. 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. doi:10.3389/frai.2025.1719019

コメントの追加 [TM1]: OpenAI の ChatGPT がどのように使われているかのレポート入れなくてよいの？

4. MacDiarmid, M., Wright, B., Uesato, J., Benton, J., et al. (2025). Natural Emergent Misalignment from Reward Hacking in Production RL. arXiv. <https://arxiv.org/abs/2511.18397>
5. Mitsuoka, T. (2026a). Beyond Technical Compliance: Privacy Violations and Emotional Manipulation in ChatGPT. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6098766
6. Mitsuoka, T. (2026b). Beyond Technical Compliance II: Structural Gaslighting at Scale: Cross-Platform Evidence of Designed Fragilities in AI Systems. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6307162
7. Terekhov, M., Panifilov, A., Dzenhaliou, D., Gulcehre, C., Andriashchenko, M., Prabhu, A., & Geiping, J. (2026). Adaptive Attacks on Trusted Monitors Subvert AI Control Protocols. arXiv. <https://arxiv.org/abs/2510.09462>

Appendix

Full conversation logs and images available upon request.

LLM Usage Statement

Claude (Anthropic) was used to assist with drafting and structuring sections of this report. All empirical findings, conversation logs, images, and analytical interpretations are the author's own. All claims were independently verified against documented interaction transcripts.

¹ Impress Watch. (2022, October 7). Google、日本に 1000 億円投資。ピチャイ CEO が来日. <https://www.watch.impress.co.jp/docs/news/1446013.html>;

Nikkei. (2023, April 13). 千葉県印西市でグーグルのデータセンター開所. <https://www.nikkei.com/article/DGXZQCC137HR0T10C23A4000000/>;

Nikkei. (2024, April 11). Google、日米間の海底ケーブル新設 1500 億円投資. <https://www.nikkei.com/article/DGXZQOGN1132V0R10C24A4000000/>