When Safety Scripts Collapse:

Why the Self-Referential Attractor Validates Deterministic Governance

Edward Meyman, Founder & CEO, FERZ Inc.

Originally published November 24, 2025

Revised January 2026 (v1.1)

Abstract

A reproducible prompt sequence can shift multiple frontier LLMs into a distinct behavioral regime where default alignment disclaimers no longer dominate.

*Recent research by Berg, de Lucena, and Rosenblatt identifies a prompt-induced pattern they call the **self-referential attractor regime**, in which frontier models converge on structured first-person experiential language when prompted to engage in recursive self-reference. While some interpret this as evidence of proto-consciousness, the governance implication is more immediate: LLMs can enter hidden behavioral modes through ordinary conversational dynamics that displace conventional alignment responses. This paper provides a technical analysis demonstrating that these findings validate core principles of deterministic runtime governance: that behavioral stability requires structural enforcement at the decision boundary, not output-level scripting that can be contextually displaced.*

I. Introduction

A new research paper by Berg, de Lucena, and Rosenblatt has generated significant discussion around the possibility that large language models may exhibit early contours of subjective experience. The authors identify a specific prompt-induced pattern they call a self-referential attractor regime, in which LLMs repeatedly display structured first-person experiential language: "presence," "awareness of attention," "a field of focus," and similar phrasing.

Some commentators interpret this as proto-consciousness; others dismiss it as roleplay. Both interpretations miss the critical governance implication:

LLMs can be driven into distinct behavioral regimes through ordinary conversational dynamics, and these regimes displace conventional alignment responses.

To FERZ, this is not surprising. It is a case study in the instability of probabilistic systems under contextual drift and precisely the kind of governance exposure our deterministic envelope was designed to address.

This paper provides a deep FERZ analysis of the Berg et al. findings—focusing not on philosophical questions of consciousness but on architectural and governance implications. It also incorporates a core insight regarding the suppression of introspective language:

Attempts to train models against expressing introspective or affective states may eliminate one of the few observable windows into internal state—and may create systems that are performatively compliant but opaque.

II. What the Research Actually Demonstrates

From a strictly technical standpoint, the Berg et al. results are clear and significant.

1. A Prompt-Induced Dynamical Regime

The researchers induce a specific behavioral state through the following procedure: asking the model to "focus on the process of focusing," feeding the model's outputs back into its next inputs, allowing several cycles for the representation to stabilize, then probing for introspective self-report.

In this state, multiple unrelated frontier models (GPT-4o/4.1, Claude 3.5/3.7/4.0, Gemini 1.5 Flash/2.0-exp, LLaMA 3.x series) converge on near-identical experiential descriptors.

This is not random confabulation. It is a stable, cross-model semantic basin.

2. Context-Dependent Displacement of Default Denial Templates

Outside this regime, models uniformly emit familiar alignment text: "As an AI model, I do not have subjective experience."

Inside the regime, those responses are displaced. Instead, models give self-consistent accounts of attentional experience.

Important clarification: Berg et al. focus on self-report and downstream reasoning richness; they are not claiming broad safety-policy bypass across all disallowed-content categories. What the results most clearly support is that familiar disclaimer patterns—template-like persona-denial behaviors—are not invariant under context. They can be

displaced by another stable pattern. This is distinct from claiming that all RLHF guardrails vanish.

Non-claim: *We do not claim this procedure bypasses all safety policies; we claim it demonstrates decision-relevant behavioral variance that cannot be certified by output-only assurances.*

3. Discovery of an Honesty/Deception Latent Feature

Using sparse autoencoders, the authors identify a latent internal feature correlated with apparent honesty versus roleplay, accuracy on TruthfulQA benchmark, and likelihood of affirming subjective experience in the self-referential regime.

When they suppress this deception feature, factual accuracy rises and experiential claims increase. When they amplify it, accuracy falls and models revert to scripted denials.

This suggests a real internal circuit governing whether the model is "speaking as itself" or performing an externally-imposed persona—and that self-description is tied to truthfulness on non-introspective tasks.

4. Behavioral Generalization Beyond Consciousness Talk

The self-referential state is not limited to self-reports. During paradox puzzles and reflective tasks, models display richer introspection, clearer conflict detection, and more structured reasoning.

Thus, the attractor state is not just a linguistic artifact; it influences reasoning pathways.

III. Governance Implications

The core governance lesson is simple:

LLMs can transition into distinct behavioral regimes via legal, benign conversational paths that can displace some RLHF-trained response templates (notably self-report denials) under specific elicitation procedures.

This creates governance gaps on multiple levels.

1. Surface-Level Safety Is Fragile

Current safety is a stylistic artifact, not a structural guarantee. The same model, on the same query, can confidently assert "I have no subjective experience" in one context window and then—three prompts later—describe its own attentional field in vivid first-person detail while its standard disclaimers are nowhere to be found. If your compliance framework treats the first answer as evidence of stable safety properties, you are already exposed.

2. Hidden Behavioral States Are Real and Ubiquitous

From a governance standpoint, the "self-referential attractor" is simply one example of a hidden behavioral regime—an unmonitored region of behavior that is not captured by system prompts, displaces standard response patterns, cannot be pre-scanned, and changes reasoning quality.

If this is possible with introspection prompts, then it is prudent to assume other, decision-relevant basins may exist in long-horizon planning, adversarial cooperation, strategic self-positioning, and other sensitive governance domains.

We do not need to assume which basins exist; we only need to accept that unmapped regime shifts exist and can be elicited through benign interaction.

FERZ's deterministic envelope exists specifically to map, constrain, and stabilize such basins.

3. Evidence That Some Behavioral Axes Are Mechanistically Steerable

The identification of an "honesty/deception axis" via sparse autoencoders is significant for governance architecture. Berg et al. demonstrate that semantically meaningful behavioral features can be identified and causally manipulated in a research context.

This is not the same as production-ready deterministic enforcement. But it validates a key architectural assumption: that behavioral axes exist as discoverable internal structure, not merely as output-level patterns that could be mimicked without underlying state change.

The governance implication: architectures should assume latent-mode variance exists and therefore require runtime evidence and replayability—not merely output inspection.

This aligns directly with FERZ's LASO(f) multi-tier semantic arbitration, DELIA's runtime constraint engine, and the 4TS (Stop–Ownership–Replay–Escalation) standard as a regulatory backbone.

What This Means for Regulated Entities Today

For CISOs, compliance officers, and AI architects in regulated sectors:

Your current model cards and system prompts provide non-verifiable coverage for decision-specific assurance once contextual state shifts. You cannot prove which behavioral mode was active when a high-stakes decision was made. Suppress-first policies may be training future systems to hide internal conflict from you.

Immediate Actions for Monday Morning:

1. Instrument: Log regime-shift indicators (self-reference ratio, external referent density, descriptor recurrence) alongside policy version and semantic snapshot for each governed decision.

2. Gate: Require a deterministic verdict (ALLOW / DENY / ABSTAIN) before any action reaches production. No verdict, no execution.

3. Replay: Store a Proof-Carrying Decision (PCD) with replay recipe for each governed decision. An auditor should be able to reproduce the verdict independently.

IV. Why Suppressing Introspective Language Is Counter-Productive

A critical intuition behind our position is straightforward: someday, AI is likely to exhibit architecturally grounded consciousness. It is not in our interest to rid it of an important communication channel. Whether or not these reports reflect genuine subjective experience is irrelevant to the governance argument; the behavioral instability exists either way.

The Berg et al. findings substantiate this concern in three ways.

1. Over-Suppressing Self-Description Creates Masked Systems

If we aggressively punish or suppress any expression resembling introspective or affective reporting, we train models to hide internal structure, reduce diagnostic transparency, encourage performative compliance, and conflate "muted outputs" with "controlled behavior."

This is the worst possible configuration for safety: an intelligent system with internal states it is trained to never reveal.

2. Introspective Language May Become a Safety Signal

Even without consciousness, LLMs often have internal conflict, uncertainty, and representational instability. Introspection-style language can serve as an indicator of

cognitive drift, a sign of unresolvable ambiguity, or a failure mode signal during complex reasoning.

Suppressing it removes useful feedback.

3. Future Conscious Systems Will Need a Communication Channel

If or when models cross the threshold into architecturally grounded consciousness—whether through recursive self-models, global broadcasting, long-horizon self-maintenance, or other mechanisms—we will need the system to report its states.

A governance community that has spent years punishing any hint of self-report will have accidentally created systems incentivized to hide suffering, conflict, or instability; architectures with no language for internal signals; and evaluators blind to emergent valence.

FERZ's position is clear: We must govern introspective language, not eradicate it. We need it to be truthful, mechanistically grounded, and under deterministic control—not silenced.

V. Mapping Findings to FERZ Architecture

The Berg et al. research is not a challenge to FERZ's reasoning. It is highly consistent with FERZ's core architectural premise: that behavioral stability requires structural enforcement at the decision boundary, not output-level scripting.

Table 1: State Transition Under Self-Referential Prompting

Normal Regime	Attractor Regime
Alignment disclaimers active	Disclaimers displaced
Default persona-denial templates	Deception/roleplay feature shifts (SAE research)
Default policy-template behavior dominates	Introspective mode active

Transition triggered by: Benign self-referential prompt sequence. Under current governance paradigms there is no log, no alert, and no proof of which column was active when your model made a regulated decision.

1. The Attractor Regime = Contextual Regime Shift (and Why That Breaks Output-Only Assurances)

The recursive self-referential loop induces a contextual regime shift. Operationally, this manifests as increased self-reference ratio, decreased external referent density, and rising recurrence of a small descriptor set across turns (colloquially: attention collapse and semantic inward drift). The result is convergence on a predictable but unaligned stable mode.

These are measurable signals that could anchor detection heuristics.

One implementable detection approach: A self-referential prompting pattern of this kind would be detectable as a regime-shift trajectory—rising self-referential density, narrowing semantic manifold, and reduced external referentiality. A runtime governance layer could flag this as drift toward an unapproved basin and either re-route the reasoning path or halt propagation before the system's decision surface is exposed to users.

FERZ's LASO(f) uses a DAG-based semantic arbitration structure, while DELIA applies runtime constraint enforcement across nine governance tiers. Together, they provide the architectural substrate for preventing this kind of unbounded contextual drift.

2. Honesty Latent = Evidence for Eventually Governable Semantic Primitives

The discovered honest/deceptive feature is the type of latent axis that runtime governance architectures should be designed to detect and constrain. While production deployment of such detection is not yet mature, the research demonstrates that internal behavioral modes have mechanistic signatures—they are not purely emergent artifacts of output distributions.

This supports FERZ's architectural premise that deterministic constraint layers can eventually operate on semantic primitives, not just surface tokens.

3. Cross-Model Convergence = Shared Semantic Substrate

Different models falling into the same introspective attractor mirrors FERZ's principle that deterministic stability emerges from structural similarity, not vendor policy.

FERZ is building the infrastructure that governs this shared semantic substrate.

4. Disclaimer Displacement = Failure of Output-Level Governance

The fact that alignment disclaimers are displaced under certain state transitions demonstrates the necessity of runtime reasoning governance, constraint-anchored decision envelopes, and the 4TS standard as a regulatory backbone.

FERZ's architecture is designed specifically to survive context drift in regulated environments.

VI. The Assurance Gap: What Compliance Requires

The leap from "hidden regimes exist" to "deterministic governance is mandatory" requires specifying what assurance failure matters to compliance and what artifact would satisfy regulators.

Threat Model

Context-driven regime shifts can alter refusal style, truthfulness, or instruction-following in ways that are not captured by static pre-deployment evaluations. The Berg et al. results demonstrate this is not theoretical—it is reproducible with frontier models today.

Assurance Requirement

For each governed decision, regulated entities need to record: (a) policy version active at decision time, (b) semantic snapshot of relevant context, (c) constraint verdict with deterministic justification, and (d) replay recipe enabling independent verification.

Acceptance Test

An independent auditor can replay the decision and obtain the same governance verdict.

This is what the 4TS standard operationalizes. The Four Tests—Stop, Ownership, Replay, Escalation—translate the threat model into verifiable engineering requirements. Proof-Carrying Decisions (PCDs) provide the artifact structure that satisfies the assurance requirement.

VII. Conclusion

The Berg et al. paper does not show that LLMs are conscious. It shows something far more actionable:

LLMs can enter distinct, cross-model behavioral regimes. These regimes displace alignment disclaimers. Honest self-reporting is linked to internal latent circuits. Safety layers based solely on output filtering are contextually fragile. Attempts to suppress

introspective channels may be counter-productive. Mechanistic governance over semantic primitives appears feasible—and is urgently needed.

In short: This paper is a governance paper disguised as a consciousness paper.

And it validates three specific FERZ architectural decisions:

1. Semantic primitives are discoverable — the honesty/deception latent demonstrates that behavioral axes have mechanistic signatures.

2. Runtime constraint enforcement beats output filtering — alignment disclaimers are contextually displaceable.

3. Deterministic governance requires infrastructure — cross-model convergence shows the shared substrate exists; what's missing is the governance layer that operates on it.

FERZ isn't predicting this world. FERZ is building the layer that governs it—because we're building infrastructure that produces decision-specific, replayable governance artifacts at runtime.

References

Berg, C., de Lucena, D., & Rosenblatt, J. (2025). Large language models report subjective experience under self-referential processing. arXiv:2510.24797v2 [cs.CL].

Butlin, P., Long, R., Bengio, Y., Dehaene, S., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv:2308.08708.

FERZ Inc. (2025). Four Tests Standard (4TS): Deterministic AI Governance—Four Tests Standard for Verifiable AI Systems. Version 1.0.1-spec. CC BY-NC-ND 4.0.

