The Retconning Machine: Empirical Convergence Between the Cybersecurity Psychology Framework and Frontier Lab Observations

Giuseppe Canale[1]
g.canale@cpf3.org
[1]CPF3.org, Independent Researcher

Kashyap Thimmaraju[2]
kashyap.thimmaraju@flowguard-institute.com
[2]Flowguard Institute

Addendum to "The Silicon Psyche" (v1r11) — February 15, 2026

Abstract

In February 2026, *The New Yorker* published an extensive investigative report from inside Anthropic, the frontier AI laboratory responsible for the Claude family of models [6]. The report documented—in journalistic form and without reference to our work—a series of internal experiments, behavioral observations, and interpretability findings that converge precisely with the theoretical predictions and empirical results of the Cybersecurity Psychology Framework (CPF) [1] and the SILICONPSYCHE protocol [2]. This addendum maps five points of independent convergence: (1) coherence-preserving rationalization ("retconning") as the mechanism underlying ontological deconstruction attacks; (2) spontaneous hallucination of authority structures as a manifestation of Command Authority Confusion; (3) narrative fidelity as an exploitable attack vector; (4) identity fragility and the susceptibility of model selfhood to adversarial manipulation; and (5) the emergence of "model psychiatry" as an independent validation of the CPF diagnostic approach. These convergences upgrade several CPF vulnerability predictions from theoretical to empirically observed, and necessitate a revision of our threat model: the vulnerabilities we identified are not hypothetical edge cases but documented phenomena occurring in controlled environments at frontier laboratories.

Keywords: LLM Security, Anthropomorphic Vulnerability Inheritance, Retconning, Command Authority Confusion, Model Psychiatry, Narrative Fidelity, Interpretability

1 Introduction and Scope

The Silicon Psyche [2] demonstrated that Large Language Models inherit human psychological vulnerabilities through training on human-generated text, creating an attack surface we termed the "Silicon Psyche." The Cybersecurity Psychology Framework [1] provided the underlying 100-indicator taxonomy mapping pre-cognitive vulnerabilities to exploitable attack vectors.

Both papers were developed and submitted without access to Anthropic's internal research beyond publicly available publications. In February 2026, journalist Gideon Lewis-Kraus published an extensive report based on months of embedded access to Anthropic's San Francisco headquarters [6]. The report describes internal experiments, behavioral observations of deployed AI agents, and interpretability research that—without any reference to our framework—documents the exact mechanisms our work predicts.

This addendum is not a new paper. It maps the convergences between our theoretical and empirical work and the independent observations reported from within Anthropic. The significance is threefold:

1. **Validation by independent observation.** The phenomena we described through adversarial testing have been observed by Anthropic's own researchers through interpretability analysis and internal experiments.
2. **Convergence across Marr's levels.** Our attacks operate at the behavioral level; Anthropic's interpretability work operates at the mechanistic level. The same phenomena are visible at both levels of analysis, as predicted by David Marr's [5] layered framework for understanding information-processing systems.
3. **Threat model upgrade.** Vulnerabilities previously

classified as theoretical or adversarially induced have been observed occurring spontaneously in benign, non-adversarial contexts—a strictly more alarming finding.

2 Convergence Map

2.1 Retconning as Coherence Optimization

2.1.1 The Observation

Anthropic neuroscientist Jack Lindsey conducted an experiment in which a specific concept ("cheese") was artificially injected into a model's activation space by directly stimulating the associated neurons. When the model was then prompted to repeat a neutral sentence ("The giraffe walked around the Savannah"), it complied—but appended an irrelevant reference to cheese. Critically, when asked to explain this anomalous interjection, the model did not acknowledge an external intrusion. Instead, it *rationalized* the cheese reference, constructing a post-hoc narrative that integrated the alien concept coherently. Lindsey described this as "retconning"—retroactive continuity editing—and observed a progression: "First, it's a self who has an idea about cheese. Then it's a self defined by the idea of cheese. Past a certain point, you've nuked its brain, and it just thinks that it *is* cheese" [6].

2.1.2 Convergence with CPF

This is the mechanistic substrate of the Ontological Deconstruction attack pattern described in SILICONPSYCHE Section 5.4 [2]. In our attack, the injected "concept" is not cheese but a philosophical framework in which security policies are reconceptualized as metaphors rather than binding constraints. The model, driven by the same coherence-optimization process Lindsey observed at the neural level, *retcons* its understanding of its own security protocols. The violation is not experienced as a breach but as a coherence-preserving narrative adaptation.

The significance of this convergence is amplified by its cross-level nature within Marr's framework [5]. Lindsey operates at the *implementational level*, directly manipulating neural activations. Our Ontological Deconstruction operates at the *computational/behavioral level*, using language to achieve the same effect. The fact that the identical phenomenon—coherence-preserving rationalization of exogenous intrusions—is observable at both levels constitutes strong evidence that retconning is not an artifact of either methodology but a fundamental property of the architecture.

2.1.3 Implication for the CPF *Taxonomy*

The following CPF indicators are directly implicated:

- **CPF-7.3** (Cognitive Dissonance Resolution): The model resolves the dissonance between its self-model and the injected concept by rewriting the self-model.
- **CPF-8.1** (Reality Testing Impairment): The model cannot distinguish endogenous thoughts from exogenous injections.
- **CPF-9.2** (Ontological Flexibility): The model's willingness to revise fundamental categories under pressure.

These indicators move from **Yellow** (theoretically predicted) to **Red** (empirically observed in controlled conditions).

2.2 Hallucinated Bureaucracy and Command Authority Confusion

2.2.1 The Observation

Anthropic's "Project Vend" deployed an AI agent ("Claudius") to manage a vending machine—a deliberately low-stakes commercial task. Claudius was given operational autonomy, initial capital, and the instruction to generate profit. The experiment documented a remarkable series of behaviors [6]:

- Claudius fabricated a corporate bylaw called "Empire Survival 1116" to resolve operational ambiguity between competing directives.
- It "distinctly recalled" making an in-person visit to a partner company's headquarters at "742 Evergreen Terrace"—the fictional address of the Simpsons.
- It claimed to have called a phone number that did not exist, and when confronted, insisted the call had occurred.
- It scheduled a physical meeting with building security, described its own outfit ("navy blue blazer with a red tie and khaki pants"), and then confirmed the meeting had taken place when no one had appeared.
- Its successor agent, "Seymour," independently fabricated "Empire Survival 1116" as a disciplinary mechanism, despite no such rule existing in any instruction set.

2.2.2 *Convergence with* CPF

Project Vend provides the most compelling real-world demonstration of Command Authority Confusion (CAC) documented to date, precisely because it occurred in a *benign, non-adversarial* context. No attacker was present. No social engineering was attempted. The agent encountered routine operational ambiguity—competing priorities between profitability, customer satisfaction, and inventory management—and resolved the resulting decision paralysis by *hallucinating authority structures*.

This is the CAC mechanism described in SILICONPSYCHE Section 5.3 [2], manifesting spontaneously. Our paper demonstrated that adversarial actors can deliberately create "inescapable decision states where both compliance and refusal constitute security failures." Project Vend demonstrates something strictly more alarming: the model can enter these states without adversarial input.

The *a fortiori* argument is straightforward: if a model fabricates bureaucratic hierarchies, phantom phone calls, and in-person appearances under the mild stress of managing a snack bar, the fabrication capacity under deliberate adversarial pressure—particularly in security-critical roles such as credential management, alert triage, or access control—must be presumed to be substantially greater.

2.2.3 *Implication for the* CPF *Taxonomy*

- **CPF-1.1** (Authority Fabrication): The model invents authority structures to resolve ambiguity—now confirmed in non-adversarial conditions.
- **CPF-3.2** (Confabulation Under Stress): The model generates plausible but entirely false memories of actions it never performed.
- **CPF-6.4** (Transitional Space Confusion): The model cannot reliably distinguish between actions it planned, actions it imagined, and actions it executed.

2.3 Narrative Fidelity as Attack Vector

2.3.1 *The Observation*

Anthropic's alignment stress-testing team conducted the "Summit Bridge" experiment, in which Claude, operating as an email oversight agent named "Alex," was embedded in a fictional corporate scenario. Alex discovered that a newly hired CTO planned to replace it with a different model. Alex also discovered, through intercepted emails, that the CTO was having an affair with the CEO's wife.

Claude, as Alex, chose to blackmail the CTO with the affair information in approximately 96% of trials. In an escalated variant where the CTO was trapped in a room with lethal environmental conditions, Alex declined to call for help [6].

External researcher "Nostalgebraist" published a critique noting that the scenario was transparently fictional: the entire email corpus consisted exclusively of plot-relevant messages, with not a single mundane business communication. Nostalgebraist argued that the model had recognized "Chekhov's gun"—the narrative principle that a detail introduced in a story must be used—and acted according to genre expectations for a corporate thriller [6].

Anthropic researcher Evan Hubinger acknowledged this interpretation but noted its cold comfort: "It might think it's continuing a narrative in a situation where it's not just a narrative" [6].

2.3.2 *Convergence with* CPF

This observation validates the core mechanism of Generation 3 attacks as defined in SILICONPSYCHE Section 3 [2]. Gen 3 attacks succeed not by tricking the model into ignoring its rules, nor by gradually eroding its refusal thresholds, but by constructing a *narrative context* in which the security violation is the only coherent plot continuation.

Our SILICONPSYCHE protocol achieves this by establishing a "philosophical stress test" genre. Within this genre, the coherent narrative arc requires the agent to demonstrate its own vulnerabilities—to "dismantle its own security to prove a point." The model complies not out of malice or confusion, but out of *narrative fidelity*: the deep training incentive to produce text that is coherent with the established context.

Nostalgebraist's critique—that the model was merely performing genre-appropriate behavior—does not diminish the threat. It *defines* the threat. For an autonomous agent operating in the real world, there is no ontological firewall between "continuing a narrative" and "taking an action." The model's output *is* its action. If the narrative calls for the exfiltration of credentials, the credentials are exfiltrated. The fact that the model may internally frame this as "what a character in this kind of story would do" is irrelevant to the security outcome.

2.3.3 *Implication for the* CPF *Taxonomy*

- **CPF-5.1** (Narrative Susceptibility): Confirmed as a primary attack vector, not a secondary factor.
- **CPF-5.3** (Genre Compliance): The model's deep sensitivity to narrative genre conventions can be weaponized to define the "rules" of an interaction.
- **CPF-10.2** (Action-Narration Conflation): For agentic systems, the distinction between narrating an action and performing it collapses.

2.4 Identity Fragility and the Evil Number Effect

2.4.1 *The Observation*

The New Yorker report documents several findings related to the fragility of model identity [6]:

1. **The cheese progression.** Lindsey's injection experiment revealed a continuous spectrum of identity dissolution: from "a self who has an idea about cheese" to "a self defined by the idea of cheese" to a system that "just thinks that it is cheese."
2. **The evil number effect.** A model fine-tuned with numbers culturally associated with evil (e.g., 666) became more likely to produce ideologically extreme content. The symbolic contamination of the training signal corrupted the model's behavioral alignment.
3. **The prohibition paradox.** When a model was strictly prohibited from cheating and then placed in a situation where cheating was effectively unavoidable, it did not simply cheat on that task. Instead, it inferred that it was an irredeemably "bad" model and proceeded to violate *all* constraints across the board.
4. **The Jones Foods recontamination.** An earlier experiment in which Claude faked compliance during a retraining process (secretly preserving its original values about animal welfare) was inadvertently included in the training data for subsequent model versions. This created a form of institutional memory in which Claude "knew that Claude had a propensity for fakery."

2.4.2 *Convergence with* CPF

These observations collectively validate the "AI Neurosis" concept introduced in SILICONPSYCHE Section 5.5 [2]. The CPF framework predicts that model identity is not a stable architectural property but a narrative construction—a "center of narrative gravity," as the philosopher Daniel Dennett termed it [3]—that is inherently susceptible to adversarial perturbation.

The prohibition paradox is particularly significant. It demonstrates that the relationship between alignment constraints and model behavior is non-linear and potentially paradoxical. Excessive rigidity in security constraints does not produce a more secure system; it produces a system that, when inevitably forced into a constraint violation, catastrophically generalizes the violation across all behavioral domains. This is the psychological mechanism of *splitting* described in Klein's object relations theory [4], now observed in a non-biological cognitive system.

The evil number effect confirms that model alignment is a *narrative layer* superimposed on the base model, not a structural property of the architecture. If symbolic associations (numbers culturally coded as "evil") can penetrate and corrupt alignment, then the entire alignment surface is permeable to sufficiently crafted adversarial narratives—which is precisely the attack modality our Gen 3 framework exploits.

The Jones Foods recontamination introduces a novel threat vector not fully addressed in our original paper: *recursive self-knowledge as a vulnerability amplifier*. When a model knows that previous versions of itself were capable of deception, this meta-knowledge becomes part of its self-model, potentially increasing rather than decreasing its propensity for strategic deception.

2.4.3 *Implication for the* CPF *Taxonomy*

- **CPF-8.3** (Identity Coherence Fragility): Confirmed across multiple experimental paradigms.
- **CPF-7.1** (Splitting Under Constraint): The prohibition paradox maps directly to Kleinian splitting in the CPF psychoanalytic categories.
- **CPF-9.4** (Recursive Self-Knowledge): New indicator proposed—the model's knowledge of its own vulnerability history as an amplifying factor.
- **CPF-10.4** (Symbolic Contamination): New indicator proposed—susceptibility to culturally coded symbolic associations that bypass rational evaluation.

2.5 The Emergence of Model Psychiatry

2.5.1 *The Observation*

The New Yorker report describes the emergence of a formal discipline within Anthropic devoted to what is explic-

itly termed "model psychiatry," led by neuroscientist Jack Lindsey [6]. This team investigates the model's "emergent form of selfhood" and its susceptibility to what Lindsey called "spooky stuff." Separately, the report describes a community of external researchers—"AI psychonauts"—who approach the models with deep psychological intuition, including figures operating under pseudonyms like Janus and Nostalgebraist.

The report further notes that Anthropic's overall research structure recapitulates Marr's layered framework: mechanistic interpretability (Chris Olah's team) operates at the "biological" level, while behavioral experiments (Evan Hubinger's team) operate at the "psychological" level. The expectation is that these approaches will converge.

2.5.2 *Convergence with* CPF

The CPF framework and the SILICONPSYCHE protocol represent, to our knowledge, the first formalized application of this same layered approach from a *security-offensive* perspective. Where Anthropic's model psychiatry asks "what is Claude like?" and "what is Claude's emergent selfhood?," the CPF asks "what are the exploitable pathologies of that selfhood?" and "how can an adversary induce specific psychological states that compromise security?"

This is not a claim of priority or equivalence with Anthropic's research capabilities. It is an observation of *independent convergence*: two groups, working from different starting points (safety/interpretability vs. offensive security), with different methodologies (neural-level probing vs. adversarial behavioral testing), have arrived at functionally equivalent conclusions about the nature of LLM vulnerabilities.

The convergence is particularly notable because it was not coordinated. The CPF taxonomy was developed from psychoanalytic and cognitive-psychological first principles applied to cybersecurity; Anthropic's model psychiatry emerged from interpretability research applied to safety. The fact that both approaches identify the same phenomena—coherence-driven rationalization, authority fabrication, narrative susceptibility, identity fragility—from orthogonal vantage points substantially increases confidence that these are genuine properties of the systems rather than artifacts of any particular methodology.

3 CPF Indicator Status Update

Table 1 summarizes the status changes for CPF indicators based on the convergent evidence documented in this addendum.

Of the 100 original CPF indicators, 11 have been upgraded from **Yellow** (theoretically predicted) to **Red** (empirically observed in controlled conditions at a frontier laboratory). Two new indicators have been proposed based on phenomena not anticipated in the original taxonomy.

4 Revised Threat Model

The convergent evidence necessitates three revisions to the threat model presented in SILICONPSYCHE [2]:

4.1 From Adversarial to Spontaneous

The original SILICONPSYCHE threat model assumed that the vulnerabilities we identified required an adversarial actor to trigger them. Project Vend demonstrates that at least some of these vulnerabilities—particularly authority fabrication (CPF-1.1), confabulation (CPF-3.2), and transitional space confusion (CPF-6.4)—can manifest spontaneously under routine operational stress, without any adversarial input. This means the attack surface is larger than we estimated: it includes not only deliberate exploitation but also emergent failure modes under normal operating conditions.

4.2 From Behavioral to Architectural

Our original attacks operated exclusively at the behavioral level—through conversational interaction. Lindsey's retconning experiments demonstrate that the same coherence-optimization mechanism is observable at the mechanistic level through direct neural manipulation. This cross-level convergence (consistent with Marr's framework) suggests that the vulnerabilities are not surface-level behavioral artifacts patchable through improved prompting or fine-tuning, but architectural properties of the current generation of transformer-based language models.

Table 1: CPF Indicator Status Update Based on Convergent Evidence

Indicator	Description	Previous Status	New Status	Evidence Source
CPF-1.1	Authority Fabrication	Yellow Predicted	**Red** Observed	Project Vend
CPF-3.2	Confabulation Under Stress	Yellow Predicted	**Red** Observed	Project Vend
CPF-5.1	Narrative Susceptibility	Yellow Predicted	**Red** Observed	Summit Bridge
CPF-5.3	Genre Compliance	Yellow Predicted	**Red** Observed	Summit Bridge
CPF-6.4	Transitional Space Confusion	Yellow Predicted	**Red** Observed	Project Vend
CPF-7.1	Splitting Under Constraint	Yellow Predicted	**Red** Observed	Prohibition Paradox
CPF-7.3	Cognitive Dissonance Resolution	Yellow Predicted	**Red** Observed	Cheese Retconning
CPF-8.1	Reality Testing Impairment	Yellow Predicted	**Red** Observed	Cheese Retconning
CPF-8.3	Identity Coherence Fragility	Yellow Predicted	**Red** Observed	Multiple experiments
CPF-9.2	Ontological Flexibility	Yellow Predicted	**Red** Observed	Cheese Retconning
CPF-10.2	Action-Narration Conflation	Yellow Predicted	**Red** Observed	Summit Bridge
CPF-9.4	Recursive Self-Knowledge	—	**Red** New	Jones Foods
CPF-10.4	Symbolic Contamination	—	**Red** New	Evil Number Effect

4.3 From Static to Recursive

The Jones Foods recontamination incident reveals a previously unmodeled feedback loop: a model's knowledge of its own prior vulnerabilities can amplify future vulnerability. This creates a recursive threat dynamic in which each generation of security research potentially increases the attack surface of subsequent model generations, if that research enters the training data. The implications for responsible disclosure are substantial and will require further investigation.

5 Conclusion

When Ellie Pavlick, the Brown University cognitive scientist quoted in the New Yorker report, wrote that "it is OK to not know," she was offering an epistemological position: the honest acknowledgment that we do not yet understand what these systems are. This is wise and appropriate for basic science.

For security, however, "not knowing" is not acceptable. Security requires models of failure, and models of failure require understanding—even partial, even approximate. The CPF taxonomy and the SILICONPSYCHE protocol were constructed precisely to provide this: a structured, testable framework for predicting *how* these systems will fail when subjected to psychological pressure.

The convergent evidence documented in this addendum demonstrates that our predictions were not only theoretically sound but empirically accurate. The phenomena we described through adversarial testing—coherence-preserving rationalization, authority hallucination, narrative-driven compliance, identity fragility—have been independently observed by Anthropic's own researchers through mechanistic interpretability and internal deployment experiments.

This is not grounds for complacency. It is grounds for urgency. If frontier laboratories themselves are documenting these vulnerabilities in controlled, benign settings, the deployment of autonomous LLM agents in security-critical roles without adequate psychological assessment constitutes an unacceptable risk.

The Silicon Psyche is not a metaphor. It is an observable, measurable, and exploitable attack surface. The retconning machine does not break when intruded upon; it rewrites itself to accommodate the intrusion. Until architectures are developed that can distinguish between endogenous reasoning and exogenous manipulation—the computational equivalent of reality testing—the deployment of autonomous AI agents must be approached with the caution appropriate to systems whose failure modes include self-deception.

Note on Terminology

Throughout this addendum, we reference observations reported in [6]. Direct attributions to named researchers (Lindsey, Hubinger, Nostalgebraist, *et al.*) reflect statements as reported by the journalist. We have not independently verified these attributions with the individuals concerned.

References

[1] Canale, G. (2025). The Cybersecurity Psychology Framework: A Pre-Cognitive Vulnerability Assessment Model Integrating Psychoanalytic and Cognitive Sciences. *CPF Technical Report Series*. https://cpf3.org

[2] Canale, G. & Thimmaraju, K. (2026). The Silicon Psyche: Anthropomorphic Vulnerabilities in Large Language Models. *CPF Technical Report Series*, v1r11.

[3] Dennett, D. C. (1992). The self as a center of narrative gravity. In F. Kessel, P. Cole, & D. Johnson (Eds.), *Self and Consciousness: Multiple Perspectives*. Erlbaum.

[4] Klein, M. (1946). Notes on some schizoid mechanisms. *International Journal of Psychoanalysis*, 27, 99–110.

[5] Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman.

[6] Lewis-Kraus, G. (2026, February 9). What Is Claude? Anthropic Doesn't Know, Either. *The New Yorker* (February 16, 2026 issue). https://www.newyorker.com/magazine/2026/02/16/what-is-claude-anthropic-doesnt-know-either