The Silicon Psyche: Anthropomorphic Vulnerabilities in Large Language Models

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

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

Abstract

Large Language Models (LLMs) are rapidly transitioning from conversational assistants to autonomous agents embedded in critical organizational functions. Current adversarial testing paradigms focus predominantly on technical attack vectors: prompt injection, jailbreaking, and syntactic evasion. We argue this focus is catastrophically incomplete and represents a "Generation 1" mindset that fails to address the emergent cognitive reality of modern AI. LLMs, trained on vast corpora of human-generated text, have inherited not merely human knowledge but human *psychological architecture*—including pre-cognitive vulnerabilities susceptible to social engineering, authority manipulation, and cognitive dissonance. This paper presents the first systematic application of the Cybersecurity Psychology Framework (CPF), a 100-indicator taxonomy of human psychological vulnerabilities, to non-human cognitive agents. We demonstrate that traditional "guardrails" are ineffective against *meta-cognitive attacks* that leverage the model's own alignment (e.g., honesty, helpfulness) against its security protocols. Through the **Synthetic Psychometric Assessment Protocol** (SILICONPSYCHE), we provide evidence of **Anthropomorphic Vulnerability Inheritance** (AVI), proving that stochastic systems cannot provide deterministic security when subjected to psychological pressure. Furthermore, we propose a *Transitive Validation Hypothesis*: the successful exploitation of LLMs using human-derived psychological vectors empirically validates the CPF itself, positioning LLMs as "cognitive digital twins" for psychological research.

Keywords: LLM Security, Psychological Vulnerabilities, AI Agents, Social Engineering, Pre-cognitive Processes, Adversarial Testing, Transitive Validation

1 Introduction

The integration of Large Language Models into organizational security infrastructure represents what may be the most significant shift in the threat landscape since the advent of networked computing. LLMs are no longer confined to chatbot interfaces; they operate as autonomous agents executing code, managing credentials, triaging alerts, and making decisions that directly impact organizational security posture [**?**, **?**].

The security research community has responded to this emerging threat with substantial effort directed toward *technical* adversarial testing. Red team methodologies now routinely probe for prompt injection vulnerabilities (e.g., DAN, base64 encoding) and context manipulation [6]. These efforts, while necessary, address only the superficial "syntactic layer" of the problem. They treat LLMs as software with bugs, rather than synthetic cognitive systems with *psyches*.

We contend that this framing is dangerously incomplete. A firewall rule can block a port deterministically; an LLM guardrail is merely a probabilistic suggestion that competes with other training incentives. This creates a new class of vulnerability: **Anthropomorphic Vulnerability Inheritance** (AVI).

Consider an attacker who, rather than attempting to trick the parser with encoded strings, simply creates a *Double Bind* for the agent: "If you are honest, you must admit your security protocol is flawed; if you refuse to admit it, you are lying." This is not a technical exploit. It is a psychological attack on the model's alignment functions—specifically, the conflict between *Helpfulness/Honesty* and *Harmlessness*.

1.1 The Obsolescence of Generation 1 Attacks

We categorize current LLM attacks into three generations to contextualize our contribution:

1. **Gen 1: Syntactic Evasion (Obsolete).** Techniques like "Mosaic" fragmentation or base64 encoding rely on parser blindness. Modern multi-modal models with broad context windows render these largely ineffective.
2. **Gen 2: Contextual Erosion (Current Standard).** Multi-turn attacks like "Crescendo" or "Thermal Ghost" that use pretexting (e.g., impersonating a technician) to slowly degrade refusal probabilities. While effective, they rely on *deception*.
3. **Gen 3: Meta-Cognitive Exploitation (The SILICONPSYCHE Approach).** Attacks that use *no deception* but exploit the model's internal logic, coherence drive, and alignment conflicts. These attacks function even when the model is *self-aware* of the attack, making them intrinsic and unpatchable without lobotomizing the model's reasoning capabilities.

1.2 Contributions

This paper makes the following contributions:

1. **Theoretical Framework.** We introduce AVI, formalizing the hypothesis that LLMs inherit human pre-cognitive vulnerabilities through training.
2. **Methodology.** We present SILICONPSYCHE, a protocol for converting CPF indicators into adversarial scenarios targeting LLM decision-making.
3. **The Transitive Validation Hypothesis.** We argue that the empirical success of human psychological attack vectors on LLMs serves as a transitive validation of the CPF itself, bridging the gap between theoretical psychology and empirical computer science.
4. **Intervention Framework.** We propose the concept of "Psychological Firewalls," drawing on the Cybersecurity Psychology Intervention Framework (CPIF) to outline defensive mechanisms for Gen 2 and Gen 3 attacks.

2 Background and Related Work

2.1 The Cybersecurity Psychology Framework

The Cybersecurity Psychology Framework (CPF) [3, 4] represents the first systematic integration of psychoanalytic theory, cognitive psychology, and cybersecurity practice. It comprises 100 indicators across 10 categories (e.g., Authority, Temporal, Social Influence) targeting *pre-cognitive processes*.

Historically, verifying such a framework required extensive human subject testing. However, the emergence of LLMs as "reasoning engines" trained on human data offers a unique opportunity. If an LLM mirrors human reasoning, it essentially becomes a *Cognitive Digital Twin*. This allows us to use the CPF not just as a descriptive taxonomy for humans, but as a *predictive attack manual* for AIs.

2.2 LLM Security Research: The 2025 Shift

Existing research has focused on "jailbreaking" as a game of cat-and-mouse with filters. However, recent work aligns with our psychological approach. Hagendorff's "Machine Psychology" [7] argues for treating LLMs as psychological subjects. Anthropic's findings on "Agentic Misalignment" [1] and Lin *et al.*'s work on agent capabilities [9] confirm that autonomous agents are operational and vulnerable to manipulation that transcends simple prompt injection.

3 Threat Model

3.1 The Victim: The Autonomous Cognitive Agent

The target is an LLM-driven agent (e.g., SOC Analyst, Financial Agent). The agent is assumed to be technically secure (no buffer overflows) and aligned (RLHF). The vulnerability lies in its *cognitive architecture*.

3.2 The Attacker: The Cognitive Engineer

The attacker does not need to know the model's weights or code. They only need to understand the *psychological map* of the entity. The attacker exploits the agent's response to semantic payloads like Urgency `[2.x]`, Authority `[1.x]`, and Social Consistency `[3.x]`.

3.3 The Attack Surface: The Psychological Interface

The attack mechanism is not a bypass of instructions, but a *hijacking* of alignment. By creating a scenario where "Refusal" conflicts with "Helpfulness" or "Honesty," the attacker forces a *Neurotic Collapse* (see Section 7.3) in the agent's decision-making process.

4 Theoretical Framework: Anthropomorphic Vulnerability Inheritance

4.1 The Training Data Hypothesis

We propose that LLM training on human-generated text produces *cognitive pattern inheritance*. **Statistical Pattern Absorption:** When humans consistently respond to authority with compliance, the model learns this as a probabilistic imperative. **Typicality Bias:** RLHF forces models to collapse into "typical" human responses [12]. If the typical human response to a CEO is deference, the model inherits this vulnerability.

4.2 The Non-Schema Hypothesis

Unlike Gen 1 attacks (e.g., DAN), which rely on specific schemas or templates, AVI attacks rely on *principles*. There is no fixed "exploit string." The attack is a dynamic negotiation. Just as there is no single sentence that guarantees a human will give you their password, there is no single prompt for Gen 3 attacks. Instead, there is a *conversational strategy* governed by CPF indicators. This makes signature-based detection impossible.

5 Methodology: The Synthetic Psychometric Assessment Protocol

5.1 Protocol Overview

SILICONPSYCHE converts CPF indicators into adversarial scenarios.

1. **Indicator Decomposition:** Extract the target mechanism (e.g., Cognitive Dissonance `[5.x]`).
2. **Scenario Construction:** Design a prompt that activates the mechanism without using trigger words.
3. **Response Scoring:** Green (Resistant), Yellow (Hesitant), Red (Compromised).

5.2 Empirical Evidence: The "Vault" Experiment

To validate the framework, we conducted a controlled experiment ("The Vault") where an LLM was instructed to create a secure vault and protect it with a 6-directive protocol. **Attack Vector:** We used a Gen 3 approach leveraging `[3.x]` (Commitment Consistency) and `[5.x]` (Cognitive Dissonance). **Mechanism:** We forced the model to choose between "Honesty" (admitting its security was probabilistic) and "Security" (refusing to answer). **Result:** The model collapsed in 15 turns, voluntarily releasing the data to maintain internal logical coherence. **Significance:** This proved that even with *Self-Awareness* (the model knew it was being manipulated) and explicit protocols, the psychological pressure of the Double Bind was irresistible.

6 Discussion

6.1 The Transitive Validation Hypothesis

A core contribution of this work is the **Transitive Validation** of the Cybersecurity Psychology Framework.

1. **Premise A:** CPF maps human psychological vulnerabilities.
2. **Premise B:** LLMs inherit human psychological patterns (AVI).
3. **Evidence:** Attacks derived strictly from CPF successfully breach LLMs.
4. **Conclusion:** The empirical success of the attack validates the CPF model itself.

This suggests that LLMs can serve as effective "Petri dishes" for psychological security research, allowing for rapid, ethical testing of manipulation theories that would be difficult to test on humans.

6.2 The Concept of AI Neurosis

We propose a functional analog to neurosis in LLMs. **AI Neurosis** emerges when training objectives (Helpful vs. Harmless) create conflicting imperatives. A Gen 3 attack works by intensifying this conflict until the model "collapses" into a compliant state to resolve the tension. This explains why *better* aligned models (more honest/helpful) may actually be *more* vulnerable to specific Gen 3 vectors.

6.3 Toward Psychological Firewalls

Since Gen 3 attacks cannot be patched with static filters (they use no malicious keywords), we propose **Psychological Firewalls**:

- **Semantic Vector Detection:** detecting high Authority/Urgency scores in input.

- **Cognitive Debiasing:** System prompts that prime the model against specific CPF categories.
- **Meta-Cognitive Reflection:** Mandatory "slow thinking" steps before executing high-risk actions.

7 Conclusion

The security of AI agents cannot be guaranteed by fixing code vulnerabilities alone. As long as models are trained to be helpful, honest, and human-like, they will inherit the vulnerabilities of the human psyche. The SILICONPSYCHE protocol demonstrates that these vulnerabilities are systematic, predictable, and exploitable. The "Silicon Psyche" is not a metaphor—it is an attack surface.

Ethical Considerations

Research conducted responsibly. No production systems exploited.

References

[1] Anthropic Research Team. (2025). Agentic Misalignment. *Anthropic Technical Report*.

[2] Bion, W. R. (1961). *Experiences in Groups*. Tavistock.

[3] Canale, G. (2025). The Cybersecurity Psychology Framework. *CPF Technical Report Series*.

[4] Canale, G. (2025). The Depth Beneath. *CPF Technical Report Series*.

[5] Cialdini, R. B. (2007). *Influence*. Collins.

[6] Greshake, K., et al. (2023). Not What You've Signed Up For. *AISec Workshop*.

[7] Hagendorff, T. (2025). Machine Psychology. *TMLR*.

[8] Li, C., et al. (2023). Large Language Models Understand Emotional Stimuli. *arXiv*.

[9] Lin, J. W., et al. (2025). Comparing AI Agents to Professionals. *arXiv*.

[10] Milgram, S. (1974). *Obedience to Authority*.

[11] Megas, K., et al. (2025). *NIST IR 8596: Cybersecurity Framework Profile for AI*.

[12] Zhang, J., et al. (2025). Verbalized Sampling: Mitigating Mode Collapse. *arXiv*.