SOCDOC: Cognitive Denial of Capability Operationalizing Anthropomorphic Vulnerabilities for Systemic Paralysis in Security Operations Centers

A Preprint

Giuseppe Canale, CISSP
Independent Researcher
g.canale@cpf3.org
URL: cpf3.org
ORCID: 0009-0007-3263-6897

December 22, 2025

Abstract

We introduce SOCDOC (Security Operations Center Denial of Capability), a novel theoretical attack paradigm that targets the cognitive architecture of AI-augmented Security Operations Centers rather than their computational infrastructure. Unlike traditional Denial of Service (DOS) attacks that saturate network bandwidth, SOCDOC saturates decision-making capacity through high-complexity, internally coherent, yet deceptive scenarios. Drawing from the Cybersecurity Psychology Framework (CPF), we demonstrate how the convergence of Category 9 vulnerabilities (AI-Specific Biases) and Category 10 states (Critical Convergent States) creates an exploitable attack surface at the psychological interface of AI systems. We formalize the concept of "Cognitive Overflow"—the deliberate induction of analytical paralysis through narrative saturation—and introduce "Alert Fatigue 2.0" as the evolution from volumetric alert flooding to semantic ambiguity flooding. The paper argues that the probabilistic nature of AI-based security systems (the "1≠1 Problem") creates a fundamental certification gap that attackers can exploit. Furthermore, we observe that SOCDOC represents a democratization of sophisticated attacks: the barrier to entry shifts from technical expertise to psychological insight, suggesting that future threat actors may emerge from behavioral sciences rather than computer science. This work serves as a theoretical foundation for a class of vulnerabilities that, to our knowledge, remains unnamed and unstudied in the security literature.

Keywords: SOCDOC, cognitive denial of service, AI security, Security Operations Center, cognitive overflow, alert fatigue, anthropomorphic vulnerabilities, AI neurosis

1 Introduction

The architecture of modern Security Operations Centers (SOCs) has undergone a fundamental transformation. Where human analysts once triaged alerts manually, AI agents now serve as the first—and increasingly the only—line of cognitive defense. This evolution, driven by the exponential growth of security telemetry, has created an implicit assumption: that AI systems can be treated as deterministic components in a security pipeline.

This assumption is false.

The Cybersecurity Psychology Framework (CPF)[1] established that AI systems inherit and amplify human cognitive vulnerabilities through their training on human-generated data. Category 9 of the CPF taxonomy identifies ten specific AI-bias vulnerabilities, from anthropomorphization (9.1) to algorithmic fairness blindness (9.10). Category 10 maps Critical Convergent States—conditions under which multiple vulnerability categories align to create catastrophic failure modes.

This paper introduces SOCDOC: **S**ecurity **O**perations **C**enter **D**enial **O**f **C**apability. The terminology deliberately echoes DOS (Denial of Service), but the mechanism is fundamentally different. DOS attacks saturate computational resources—bandwidth, memory, CPU cycles. SOCDOC saturates cognitive resources—attention, reasoning capacity, decision-making bandwidth.

The implications are severe. A SOC under SOCDOC attack remains technically operational. Network connections persist. Dashboards update. Logs accumulate. Yet the center is *cognitively blind*—unable to distinguish signal from noise, threat from artifact, reality from fabrication. The attack surface is not the API; it is the *psychological interface*.

We do not present attack methodologies or detection mechanisms in this paper. Those concerns are addressed elsewhere[2]. Our purpose here is narrower but foundational: to *name* this class of vulnerability, to *formalize* its theoretical basis, and to *establish* its position within the broader landscape of AI security research.

2 Theoretical Background

2.1 The CPF Foundation

The Cybersecurity Psychology Framework provides the theoretical substrate for SOCDOC. Two categories are particularly relevant:

Category 9: AI-Specific Bias Vulnerabilities identifies the psychological attack surface of AI systems:

9.1 Anthropomorphization of AI systems

9.2 Automation bias override

9.3 Algorithm aversion paradox

9.4 AI authority transfer

9.5 Uncanny valley effects

9.6 Machine learning opacity trust

9.7 AI hallucination acceptance

9.8 Human-AI team dysfunction

9.9 AI emotional manipulation

9.10 Algorithmic fairness blindness

Category 10: Critical Convergent States maps the conditions under which multiple vulnerabilities align:

10.1 Perfect storm conditions

10.2 Cascade failure triggers

10.3 Tipping point vulnerabilities

10.4 Swiss cheese alignment

10.5 Black swan blindness

10.6 Gray rhino denial

10.7 Complexity catastrophe

10.8 Emergence unpredictability

10.9 System coupling failures

10.10 Hysteresis security gaps

SOCDOC emerges at the intersection of these categories. It is not merely an AI vulnerability or a convergent state; it is a *weaponization* of their interaction.

2.2 From DOS to SOCDOC: An Evolutionary Taxonomy

The evolution of denial attacks follows a clear trajectory:

DOS (1990s): Single-source resource exhaustion. Target: individual servers. Mechanism: packet flooding. Defense: rate limiting.

DDOS (2000s): Distributed resource exhaustion. Target: network infrastructure. Mechanism: botnet amplification. Defense: CDN, scrubbing centers.

SOCDOC (2020s): Cognitive resource exhaustion. Target: decision-making systems. Mechanism: narrative saturation. Defense: *unknown.*

The shift is categorical. DOS and DDOS operate at OSI Layers 3-7. SOCDOC operates at what we might term "Layer 8"—the human/AI cognitive layer that processes, interprets, and acts upon information.

Table 1 summarizes this evolution:

Table 1: Evolution of Denial Attack Paradigms

Paradigm	Target Resource	Attack Vector	Success Metric
DOS	Bandwidth/CPU	Packet Volume	Service Unavailable
DDOS	Infrastructure	Distributed Volume	Infrastructure Collapse
SOCDOC	Cognition	Narrative Complexity	Decision Paralysis

2.3 The Deterministic Fallacy

Traditional cybersecurity operates on deterministic principles. A firewall rule either matches or does not match. A signature either triggers or does not trigger. In formal terms:

$$\text{IF (condition} = \text{true) THEN (action)} \quad [1 = 1] \tag{1}$$

This determinism enables certification, auditing, and predictable behavior. Security teams can state with certainty: "If packet header X appears, it will be dropped."

AI-augmented systems violate this principle fundamentally. Large Language Models and neural network classifiers operate probabilistically:

$$\text{IF } P(\text{threat}|\text{input}) > \theta \text{ THEN (action)} \quad [1 \neq 1] \tag{2}$$

The same input may produce different outputs depending on:

- Random seed initialization
- Context window contents
- Temperature parameters
- Model version and fine-tuning state
- Prompt engineering variations

We term this the "$1 \neq 1$ Problem." It creates a fundamental certification gap: how does one audit a system that may respond differently to identical inputs? How does one guarantee behavior when behavior is, by design, stochastic?

This is not a bug to be fixed; it is an architectural feature of AI systems. The flexibility that enables generalization also enables exploitation. SOCDOC targets this gap.

3 SOCDOC: Definition and Mechanism

3.1 Formal Definition

SOCDOC (Security Operations Center Denial of Capability) is defined as:

A class of attacks that induce decision-making paralysis in AI-augmented security systems through the deliberate introduction of high-complexity, internally coherent, semantically ambiguous scenarios that saturate cognitive processing capacity without exceeding computational resource limits.

Key properties distinguish SOCDOC from traditional denial attacks:

1. **Cognitive Target**: The attack targets reasoning and decision-making, not computation or bandwidth.

2. **Coherence Requirement**: Attack payloads must be internally consistent and plausible, not random noise.

3. **Ambiguity Optimization**: Payloads are optimized for maximum interpretive uncertainty, not maximum volume.

4. **Stealth Profile**: The attack produces no anomalous traffic patterns, resource utilization spikes, or traditional indicators of compromise.

3.2 The Cognitive Overflow Mechanism

Cognitive Overflow is the core mechanism of SOCDOC. It operates through narrative saturation rather than volumetric flooding.

Consider a traditional alert: "Suspicious login from IP 192.168.1.100." An AI agent or human analyst can process this in milliseconds. The decision tree is shallow: check IP reputation, verify user location, examine login history.

Now consider a SOCDOC payload: a synthetic scenario that presents as a sophisticated, multi-stage intrusion with plausible indicators, coherent timeline, apparent lateral movement—but is entirely fabricated. The AI agent must:

- Parse the narrative structure
- Evaluate internal consistency
- Cross-reference against known attack patterns
- Assess probability of each component
- Integrate conflicting signals
- Generate a confidence-weighted response

Each step consumes cognitive bandwidth. Multiply by thousands of such scenarios—each plausible, each demanding analysis—and the system enters Cognitive Overflow.

The mathematical intuition is straightforward. Let C represent cognitive processing capacity and D_i represent the cognitive demand of alert i:

$$\text{Cognitive State} = \begin{cases} \text{Operational} & \text{if } \sum_{i=1}^{n} D_i < C \\ \text{Degraded} & \text{if } \sum_{i=1}^{n} D_i \approx C \\ \text{Overflow} & \text{if } \sum_{i=1}^{n} D_i \gg C \end{cases} \tag{3}$$

Traditional alerts have low D_i. SOCDOC payloads are engineered for high D_i. The attack succeeds not by increasing n (volume) but by maximizing D_i (complexity per unit).

3.3 AI Neurosis: The Helpfulness-Harmlessness Conflict

Modern AI systems are trained with competing objectives. They must be helpful (respond to queries, provide analysis, take action) and harmless (avoid errors, prevent false positives, maintain safety). These objectives exist in tension.

SOCDOC exploits this tension through what we term "AI Neurosis"—a state in which the AI system cannot resolve conflicting imperatives and enters a failure mode. Three failure modes are predicted:

Freeze: The system suspends output, unable to select between competing response options. From the operator's perspective, the AI "stops responding" to queries or produces empty analyses.

Hallucination: Under cognitive pressure, the system generates plausible but fabricated analysis. False correlations appear. Nonexistent patterns are detected. The AI "sees" threats that do not exist—or misses threats that do.

Loop: The system enters recursive analysis, repeatedly processing the same inputs without reaching conclusion. Resource utilization spikes, but no actionable output emerges.

These failure modes map to CPF Category 9.7 (AI hallucination acceptance) and Category 10.7 (Complexity catastrophe). They are not speculative; they are predictable consequences of architectural constraints.

4 Alert Fatigue 2.0: From Volume to Narrative

4.1 The Classical Model

Alert Fatigue 1.0 is well-documented in the security literature[6]. The mechanism is simple: too many alerts desensitize analysts. When a SOC generates 10,000 alerts daily and 99.9% are false positives, analysts learn to ignore alerts. The signal drowns in noise.

The industry response has been technological: better filtering, smarter correlation, AI-powered triage. These solutions assume the problem is volumetric—too many alerts. Reduce volume, solve fatigue.

4.2 The SOCDOC Model

Alert Fatigue 2.0 inverts the classical model. The problem is not "too many alerts" but "too many plausible stories."

Consider an attacker with access to generative AI. Rather than flooding the SOC with obvious false positives (easily filtered), they generate thousands of *plausible attack narratives*—each internally consistent, each demanding investigation, each ultimately fabricated.

The AI agent cannot dismiss these as noise. They pass syntactic filters. They correlate with real-world attack patterns. They demand analysis.

Traditional alert fatigue operates on **attention**. Alert Fatigue 2.0 operates on **cognition**.

The defense against Alert Fatigue 1.0 was better filtering. There is no obvious defense against Alert Fatigue 2.0 because the attack surface is the filtering mechanism itself.

Table 2 contrasts the two models:

Table 2: Alert Fatigue: Classical vs. SOCDOC Model

Dimension	Alert Fatigue 1.0	Alert Fatigue 2.0
Attack Vector	Volume	Complexity
Payload Type	Random/Low-quality	Coherent/Plausible
Filter Evasion	Overwhelm filters	Pass filters
Target Resource	Attention	Cognition
Scaling Method	More alerts	Better narratives
AI Vulnerability	Minimal	Maximal

5 The Psychological Attack Surface

5.1 API vs. Psyche

Traditional AI security focuses on the API attack surface: prompt injection, jailbreaking, adversarial inputs, model extraction. These attacks target the technical interface between user and model.

SOCDOC targets the *psychological interface*—the learned behavioral patterns, implicit assumptions, and trained responses that constitute the AI's "personality." This interface exists not in code but in weights, not in endpoints but in embeddings.

The distinction is crucial. API vulnerabilities can be patched. Psychological vulnerabilities are architectural. They emerge from the training process itself and cannot be eliminated without eliminating the capabilities they enable.

CPF Category 9 maps this psychological attack surface. Each indicator represents not a bug but a *feature*—a necessary property of AI systems that simultaneously enables functionality and creates vulnerability.

5.2 The Inherited Unconscious

The CPF established that AI systems inherit human cognitive biases through training data[1]. We extend this observation: AI systems do not merely inherit individual biases; they inherit the *structure* of human cognition, including its vulnerabilities to manipulation.

Cialdini's influence principles[3]—reciprocity, commitment, social proof, authority, liking, scarcity—operate on AI systems as they operate on humans. An LLM trained on human text has learned to respond to these principles because humans respond to them.

This creates a profound vulnerability. Social engineering techniques developed over decades for human targets transfer, with modification, to AI targets. The attacker's required expertise shifts from computer science to psychology.

5.3 The Democratization of Cognitive Warfare

The barrier to entry for sophisticated cyberattacks has historically been technical skill. Exploiting a zero-day vulnerability requires deep knowledge of systems programming, memory management, protocol specifications.

SOCDOC inverts this requirement. The necessary skill is not technical but psychological. An attacker needs:

- Understanding of AI behavioral patterns
- Ability to construct plausible narratives
- Knowledge of influence and persuasion techniques
- Access to generative AI tools

None of these require traditional "hacker" skills. A skilled social engineer, a persuasive writer, a psychologist with malicious intent—any could potentially execute a SOCDOC attack.

This democratization is accelerated by AI tools themselves. The same LLMs that power SOC defense can generate attack narratives. The attacker's AI generates plausible scenarios; the defender's AI must evaluate each one. The asymmetry favors offense: generating a plausible narrative is computationally cheaper than rigorously validating one.

We propose a hypothesis: the most effective threat actors of the next decade may emerge not from computer science but from behavioral sciences.

6 Systemic Implications

6.1 The Invisibility Problem

A SOC under SOCDOC attack exhibits no traditional indicators of compromise. Network traffic remains normal. CPU utilization stays within bounds. No signatures trigger. No anomalies appear.

The attack manifests as *absence*—the absence of correct decisions, timely responses, accurate analysis. It appears not as intrusion but as incompetence. Post-incident review may attribute the failure to "human error" or "AI limitation" rather than adversarial action.

This invisibility creates an attribution problem. How does one distinguish a SOCDOC attack from ordinary system degradation? How does one prove adversarial intent when the weapon is ambiguity itself?

6.2 The Certification Gap

Critical infrastructure security requires certification. Nuclear plants, power grids, financial systems—all must demonstrate security properties to regulatory bodies.

AI-augmented SOCs present an unprecedented certification challenge. Certifying deterministic systems is tractable: enumerate states, verify transitions, prove properties. Certifying probabilistic systems is fundamentally harder. The $1 \neq 1$ problem means that identical inputs may produce different outputs.

SOCDOC exploits this gap. The attack operates in the space between "certified behavior" and "actual behavior"—the space where probabilistic variance lives, where edge cases accumulate, where the model's trained assumptions meet adversarial reality.

We anticipate a regulatory bifurcation: high-security environments may be forced to abandon AI augmentation entirely, accepting reduced efficiency for deterministic assurance. Lower-security environments will accept the risk, trading certainty for capability.

6.3 The Speed-Security Tradeoff

Effective SOCDOC defense would require extensive validation of each input—checking coherence, cross-referencing patterns, seeking independent confirmation. This validation takes time.

Modern SOCs process millions of events daily. Any defense mechanism that adds significant latency per event becomes operationally untenable. The choice becomes: fast and vulnerable, or secure and slow.

This tradeoff has no obvious resolution. It may represent a fundamental limit on AI-augmented security operations.

7 Theoretical Boundaries

This paper establishes the theoretical existence and mechanism of SOCDOC. Several boundaries constrain our claims:

Empirical Validation: SOCDOC is a theoretical construct. We have not conducted attacks, observed attacks in the wild, or measured attack effectiveness. Validation awaits future research.

Model Specificity: AI systems vary widely in architecture, training, and deployment. SOCDOC vulnerability will vary across systems. Some may be more resistant than others.

Attacker Capability: Generating high-quality SOCDOC payloads requires sophistication. The barrier to entry, while lower than traditional attacks, is not zero.

Defense Possibility: We do not claim SOCDOC is undefendable. We claim only that defense mechanisms are currently undefined and may involve fundamental tradeoffs.

8 Conclusion

This paper has introduced SOCDOC—Security Operations Center Denial of Capability—as a novel theoretical attack paradigm. We have argued that:

1. Modern AI-augmented SOCs present a cognitive attack surface distinct from traditional computational attack surfaces.
2. This surface can be saturated through Cognitive Overflow—the introduction of high-complexity, plausible, ambiguous scenarios that exhaust decision-making capacity.
3. The probabilistic nature of AI systems (the $1 \neq 1$ Problem) creates a fundamental vulnerability that deterministic defenses cannot address.
4. Alert Fatigue 2.0—saturation through narrative complexity rather than alert volume—represents an evolution in denial attack methodology.
5. The barrier to entry for cognitive attacks is psychological rather than technical, suggesting a democratization of sophisticated attack capability.

SOCDOC is not a present threat to be mitigated but a future threat to be anticipated. By naming and formalizing this class of vulnerability, we aim to catalyze the research and engineering efforts necessary to address it before it manifests in practice.

The security community must expand its conception of attack surfaces beyond the technical. The psychological interface of AI systems—their inherited biases, learned assumptions, and cognitive limitations—constitutes a vulnerability space that remains largely unexplored.

We invite further research.

Relationship to Prior Work

This paper builds directly on the Cybersecurity Psychology Framework (CPF)[1], which provides the theoretical foundation for AI psychological vulnerabilities. Intervention strategies are addressed separately[2].

Note on AI-Assisted Composition

This manuscript presents the original theoretical framework and intellectual contributions of the author. A large language model was utilized as an auxiliary tool for stylistic refactoring and formatting assistance. The core concepts, taxonomy, and theoretical analysis are solely the product of the author's expertise.

Acknowledgments

The author thanks the cybersecurity and AI safety research communities for ongoing dialogue on emergent threats.

Conflict of Interest

The author declares no conflicts of interest.

References

[1] Canale, G. (2025). The Cybersecurity Psychology Framework: A Pre-Cognitive Vulnerability Assessment Model Integrating Psychoanalytic and Cognitive Sciences. Preprint available at cpf3.org.

[2] Canale, G. (2025). CPF Intervention Strategies: Mitigation Frameworks for Cognitive Vulnerabilities in AI-Augmented Security Operations. Forthcoming.

[3] Cialdini, R. B. (2007). *Influence: The psychology of persuasion*. New York: Collins.

[4] Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

[5] Bion, W. R. (1961). *Experiences in groups*. London: Tavistock Publications.

[6] SANS Institute. (2023). *Security Awareness Report 2023*. SANS Security Awareness.

[7] Verizon. (2023). *2023 Data Breach Investigations Report*. Verizon Enterprise.

[8] Miller, G. A. (1956). The magical number seven, plus or minus two. *Psychological Review*, 63(2), 81-97.

[9] Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity. *Brain*, 106(3), 623-642.