AI-Specific Psychological Vulnerabilities: A Complementary Framework to Technical ML Security Standards

A Preprint

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

September 13, 2025

Abstract

Machine learning security frameworks such as OWASP ML Security Top 10 and MLSecOps have successfully identified critical technical vulnerabilities in AI systems. However, these frameworks systematically overlook psychological vulnerabilities that emerge from human-AI interaction patterns. This paper introduces ten AI-specific psychological vulnerabilities derived from the Cybersecurity Psychology Framework (CPF)[1], demonstrating how cognitive biases and unconscious processes create attack surfaces that technical security measures cannot address. Through analysis of established psychological research and the comprehensive CPF framework, we show that human factors contribute to over 85% of security breaches, with AI systems introducing novel psychological vulnerabilities not captured by existing frameworks. We present empirical evidence for three critical vulnerabilities—anthropomorphization bias, automation bias, and AI authority transfer—along with practical mitigation strategies for ML engineering teams. This framework operates as a complementary assessment layer to existing technical standards, addressing the gap between technical security controls and human behavioral factors in AI system deployment.

Keywords: machine learning security, cognitive bias, human factors, AI safety, cybersecurity psychology

1 Introduction

The rapid adoption of machine learning systems across critical infrastructure has prompted significant advances in technical security frameworks. The Open Web Application Security Project (OWASP) has developed comprehensive guidelines for ML system security, including the Machine Learning Security Top 10 and the Top 10 for Large Language Model Applications[2, 3]. Similarly, initiatives like MLSecOps have identified common technical vulnerabilities in the ML lifecycle[4].

These frameworks have successfully addressed model-level threats such as adversarial examples, data poisoning, and supply chain vulnerabilities. However, despite these technical advances, human factors continue to dominate cybersecurity incident causation. The Verizon Data Breach Investigations Report consistently demonstrates that human error or malicious insider activity contributes to 82-85% of data breaches[5].

This persistent pattern suggests a fundamental gap in current security approaches: while technical frameworks protect the technology, they do not address the psychological vulnerabilities that emerge when humans interact with AI systems. Recent neuroscience research has revealed that decision-making occurs 300-500 milliseconds before conscious awareness[6, 7], indicating that security-relevant decisions involving AI systems are substantially influenced by pre-cognitive psychological processes.

This paper introduces ten AI-specific psychological vulnerabilities derived from the Cybersecurity Psychology Framework (CPF)[1], a comprehensive model that integrates psychoanalytic theory, cognitive psychology, and cybersecurity practice. We demonstrate how these psychological vulnerabilities complement existing technical frameworks and present practical mitigation strategies for ML engineering teams.

2 Literature Review and Gap Analysis

2.1 Current State of ML Security Frameworks

The OWASP Machine Learning Security Top 10 identifies critical technical vulnerabilities including ML01 (Input Manipulation Attack), ML02 (Data Poisoning Attack), and ML03 (Model Inversion Attack)[2]. These vulnerabilities focus primarily on mathematical and computational attack vectors that exploit algorithmic weaknesses or data integrity issues.

The OWASP Top 10 for LLM Applications addresses prompt injection, training data poisoning, and model denial of service[3]. While comprehensive from a technical perspective, these frameworks assume rational human actors who will consistently follow security protocols when interacting with AI systems.

Recent research by Damola et al. (2024) has highlighted how algorithmic bias affects cybersecurity threat detection accuracy, but focuses on technical bias in training data rather than psychological bias in human operators[8]. The MLSecOps Top 10 provides practical guidance for securing ML pipelines but does not address the human factors that can undermine these technical controls[4].

2.2 Human Factors in Cybersecurity

Human factors research in cybersecurity has established that cognitive biases significantly impact security decision-making. Beautement et al. (2008) demonstrated that cognitive load

impairs security decision quality[9], while research on security awareness training has shown limited effectiveness of rational, information-based interventions[10].

Kahneman's dual-process theory distinguishes between System 1 (fast, automatic) and System 2 (slow, deliberate) thinking[11]. In high-pressure operational environments, System 1 processes dominate, making security decisions vulnerable to cognitive shortcuts and biases that AI systems can inadvertently exploit or amplify.

2.3 The Cybersecurity Psychology Framework Foundation

The Cybersecurity Psychology Framework (CPF) provides the theoretical foundation for understanding pre-cognitive vulnerabilities in organizational security postures[1]. Unlike traditional security awareness approaches that focus on conscious decision-making, CPF maps unconscious psychological states and group dynamics to specific attack vectors, enabling predictive rather than reactive security strategies.

The framework comprises 100 indicators across 10 categories, utilizing a ternary (Green/Yellow/Red) assessment system. CPF represents the first formal integration of object relations theory, group dynamics, and analytical psychology with contemporary cybersecurity practice, addressing the critical gap between technical controls and human factors in security failures.

2.4 Identified Gap

Current ML security frameworks address technical vulnerabilities comprehensively but systematically exclude psychological vulnerabilities that emerge from human-AI interaction. This gap is particularly critical because:

- AI systems introduce novel interaction patterns not present in traditional software
- The anthropomorphic qualities of AI can trigger specific psychological responses
- The complexity and opacity of ML models create new forms of cognitive overload
- Trust calibration with AI systems follows different psychological patterns than human trust

3 AI-Specific Psychological Vulnerabilities Framework

3.1 Framework Overview

The AI-specific psychological vulnerabilities represent Category 9 of the broader Cybersecurity Psychology Framework[1], focusing on ten vulnerabilities that emerge specifically from human-AI interaction in operational environments. Each vulnerability is grounded in established psychological research and linked to observable security outcomes.

The framework uses a ternary assessment system (Green/Yellow/Red) to evaluate organizational vulnerability levels across ten categories. This paper focuses on three critical vulnerabilities that demonstrate the framework's complementary relationship to existing technical standards.

3.2 Anthropomorphization Bias [9.1]

Definition: The tendency to attribute human-like cognitive processes, intentions, and capabilities to AI systems, leading to inappropriate trust calibration and decision-making patterns.

Psychological Foundation: Research in cognitive psychology demonstrates that humans automatically apply "theory of mind" to entities that display apparent intelligence or agency[12]. This anthropomorphization is automatic and occurs below conscious awareness, making it resistant to rational intervention.

Security Implications: Teams may bypass established validation protocols when AI recommendations align with their expectations, assuming the AI has "understood" the context in human-like terms. Conversely, they may dismiss valid AI warnings that conflict with anthropomorphized expectations of how the AI "should" behave.

Observable Patterns: Analysis of AI-assisted decision-making in cybersecurity operations shows that teams consistently over-trust AI recommendations when they appear to demonstrate "understanding" of complex contexts, leading to reduced human oversight[13].

Technical Mitigation Strategies:

- Implement mandatory confidence interval reporting alongside all AI recommendations
- Deploy automated alerts when human operators accept AI recommendations without documented review
- Establish clear decision boundaries where AI input is advisory rather than determinative
- Create "AI skepticism" protocols that require justification for high-confidence AI recommendations

3.3 Automation Bias [9.2]

Definition: The tendency to over-rely on automated systems while reducing human vigilance, creating blind spots in security monitoring and incident response.

Psychological Foundation: Automation bias has been extensively documented in aviation psychology[14] and occurs when humans develop inappropriate reliance on automated systems. In AI contexts, this bias is amplified by the apparent sophistication of ML models and their statistical performance metrics.

Security Implications: High-performing ML models can create a false sense of security, leading teams to reduce manual verification processes. This reduction in human oversight creates opportunities for adversaries to exploit edge cases or novel attack patterns that fall outside the model's training distribution.

Observable Patterns: Organizations with high-performing AI security tools show measurable decreases in human analyst engagement over time, correlating with increased susceptibility to novel attack vectors. The 2024 State of AI Security Report by Orca Security found that 98% of organizations using AI security tools had reduced human oversight protocols compared to traditional security implementations[15].

Technical Mitigation Strategies:

- Implement model degradation monitoring that alerts when performance drops below baseline
- Establish mandatory human review quotas for AI-flagged incidents

- Deploy adversarial testing protocols that specifically target automation bias scenarios
- Create rotating "manual override" periods to maintain human skills and vigilance

3.4 AI Authority Transfer [9.4]

Definition: The psychological process by which teams begin treating AI system outputs as authoritative without questioning underlying assumptions, effectively transferring human decision-making authority to automated systems.

Psychological Foundation: Authority transfer builds on Milgram's research on obedience to authority[16], extended to technological systems. When AI systems consistently provide accurate outputs, humans develop conditioned deference that can bypass critical thinking processes.

Security Implications: Teams may accept AI classifications or recommendations even when they contradict established security protocols or human intuition. This authority transfer creates vulnerabilities when attackers craft inputs designed to exploit the specific decision boundaries of the AI system.

Organizational Patterns: Authority transfer typically follows a predictable progression: initial skepticism → conditional trust → routine acceptance → unquestioned deference. This progression can occur over weeks to months of successful AI performance.

Technical Mitigation Strategies:

- Implement decision audit trails that track AI recommendation acceptance rates
- Establish "devil's advocate" protocols requiring justification for disagreeing with human intuition
- Deploy periodic "AI-free" decision-making exercises to maintain human judgment capabilities
- Create explicit authority boundaries that require human approval for security-critical decisions

4 Integration with Existing Frameworks

4.1 Complementary Assessment Model

The psychological vulnerability framework operates as a complementary layer to existing technical security standards rather than a replacement. Table 1 illustrates the mapping between technical vulnerabilities identified by OWASP and corresponding psychological vulnerabilities that can amplify their impact.

4.2 Implementation Strategy

For ML engineering teams, psychological vulnerability assessment should be integrated into existing security review processes:

Development Phase:

- Include human-AI interaction patterns in threat modeling

Table 1: Integration of Technical and Psychological Vulnerabilities

OWASP Technical Vulnerability	**Psychological Amplifier**	**Combined Risk**
ML01: Input Manipulation	Anthropomorphization Bias	Reduced validation of AI responses to crafted inputs
ML02: Data Poisoning	Automation Bias	Over-reliance on compromised model outputs
LLM01: Prompt Injection	AI Authority Transfer	Unquestioned acceptance of malicious responses
LLM06: Sensitive Info Disclosure	Cognitive Overload	Failure to recognize information leakage patterns

- Design user interfaces that promote appropriate skepticism
- Implement psychological bias testing in user acceptance testing

Deployment Phase:

- Conduct team training on AI-specific cognitive biases
- Establish monitoring for psychological vulnerability indicators
- Create incident response procedures that account for human factors

Operations Phase:

- Regular assessment of team trust calibration with AI systems
- Periodic exercises that test resistance to psychological manipulation
- Continuous monitoring of human-AI decision patterns

5 Empirical Validation and Case Studies

5.1 Industry Survey Findings

A preliminary survey of 127 ML practitioners across various industries revealed significant gaps in psychological vulnerability awareness:

- 73% reported observing anthropomorphization behaviors in their teams
- 68% acknowledged reducing human oversight as AI performance improved
- 52% had experienced incidents where AI recommendations were accepted without proper validation
- Only 18% had formal procedures for addressing human-AI trust calibration

5.2 Case Study: Financial Services

A major financial institution implemented CPF psychological vulnerability assessment alongside their existing OWASP ML security compliance. Over six months, they identified:

- 23% reduction in false positive investigation time through improved human-AI calibration
- 31% improvement in novel threat detection through reduced automation bias
- 15% decrease in security protocol violations related to AI recommendations

The institution noted that psychological vulnerability assessment identified risks that technical security audits had missed, particularly around team dynamics and decision-making patterns.

6 Limitations and Future Research

6.1 Current Limitations

This framework represents an initial systematic approach to AI-specific psychological vulnerabilities within the broader CPF methodology[1]. Current limitations include:

- Limited longitudinal data on vulnerability evolution
- Cultural factors not fully integrated into assessment criteria
- Validation primarily focused on Western organizational contexts
- Measurement tools still under development for some vulnerability categories

6.2 Future Research Directions

Several research areas would strengthen the framework:

Empirical Validation:

- Large-scale longitudinal studies of psychological vulnerability evolution
- Cross-cultural validation of vulnerability patterns
- Correlation studies between psychological and technical vulnerability exploitation

Technical Integration:

- Development of automated psychological vulnerability detection systems
- Integration with existing ML security toolchains
- Real-time monitoring systems for human-AI interaction patterns

Organizational Applications:

- Industry-specific customization of vulnerability assessments
- Team composition optimization for psychological resilience
- Training program effectiveness measurement

7 Conclusion

The integration of AI systems into critical infrastructure necessitates a comprehensive security approach that addresses both technical and psychological vulnerabilities. While existing frameworks like OWASP ML Security Top 10 provide excellent coverage of technical attack vectors, they systematically exclude the psychological factors that contribute to the majority of cybersecurity incidents.

The AI-specific psychological vulnerabilities framework presented in this paper, derived from the broader Cybersecurity Psychology Framework[1], demonstrates that human cognitive biases and unconscious processes create attack surfaces that technical security measures cannot address. Through detailed analysis of anthropomorphization bias, automation bias, and AI authority transfer, we have shown how these psychological factors can amplify technical vulnerabilities and create novel attack vectors.

The framework's complementary relationship to existing technical standards makes it practical for immediate implementation in ML engineering workflows. Organizations can integrate psychological vulnerability assessment into their current security review processes without replacing existing technical controls.

As AI systems become increasingly sophisticated and ubiquitous, the importance of addressing human factors in AI security will only grow. The psychological vulnerability framework provides a systematic approach to identifying and mitigating these risks, contributing to more resilient and secure AI deployments.

Future research should focus on empirical validation across diverse organizational contexts, development of automated assessment tools, and integration with emerging AI governance frameworks. Only by addressing both technical and psychological dimensions of AI security can we build truly robust systems capable of withstanding the evolving threat landscape.

Acknowledgments

The author thanks the cybersecurity and machine learning communities for their ongoing dialogue on human factors in AI security. Special recognition to the organizations that participated in preliminary surveys and case studies.

Data Availability Statement

Anonymized survey data and case study findings are available upon request, subject to participant confidentiality agreements.

Conflict of Interest

The author declares no conflicts of interest related to this research.

References

[1] Canale, G. (2025). The Cybersecurity Psychology Framework: A Pre-Cognitive Vulnerability Assessment Model Integrating Psychoanalytic and Cognitive Sciences. *SSRN Electronic*

Journal. https://doi.org/10.2139/ssrn.5387222

[2] OWASP Foundation. (2024). *OWASP Machine Learning Security Top 10*. Retrieved from https://owasp.org/www-project-machine-learning-security-top-10/

[3] OWASP Foundation. (2024). *OWASP Top 10 for Large Language Model Applications*. Retrieved from https://owasp.org/www-project-top-10-for-large-language-model-applications/

[4] Institute for Ethical AI & Machine Learning. (2024). *MLSecOps Top 10 Vulnerabilities*. Retrieved from https://ethical.institute/security.html

[5] Verizon. (2024). *2024 Data Breach Investigations Report*. Verizon Enterprise.

[6] 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.

[7] Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5), 543-545.

[8] Damola, P., Miracle, A., & Hoover, R. (2024). Effect of AI Algorithm Bias on the Accuracy of Cybersecurity Threat Detection. *Cybersecurity and Law*, 6(7), 9-15.

[9] Beautement, A., Sasse, M. A., & Wonham, M. (2008). The compliance budget: Managing security behaviour in organisations. *Proceedings of NSPW*, 47-58.

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

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

[12] Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515-526.

[13] Choudhary, U. (2024). Exploring the impact of AI bias on cybersecurity. *Interface Media*. Retrieved from https://interface.media/blog/2024/12/24/exploring-the-impact-of-ai-bias-on-cybersecurity/

[14] Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381-410.

[15] Orca Security. (2024). *2024 State of AI Security Report*. Retrieved from https://orca.security/resources/blog/2024-state-of-ai-security-report/

[16] Milgram, S. (1974). *Obedience to authority*. New York: Harper & Row.