CPF Mathematical Formalization Series - Paper 9: AI-Specific Bias Vulnerabilities: Mathematical Models for Human-AI Security Interactions

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

September 26, 2025

Abstract

We present the complete mathematical formalization of Category 9 indicators from the Cybersecurity Psychology Framework (CPF): AI-Specific Bias Vulnerabilities. This novel category addresses psychological vulnerabilities emerging from human-AI interactions in security contexts. Each of the ten indicators (9.1-9.10) receives rigorous mathematical definition through hybrid models combining cognitive bias detection, machine learning uncertainty quantification, and anthropomorphization metrics. The formalization enables systematic assessment of vulnerabilities unique to AI-integrated security environments, including automation bias, algorithmic trust calibration, and AI-human team dysfunction. We provide explicit algorithms for real-time detection, interdependency matrices capturing AI-specific correlation patterns, and validation frameworks adapted for human-AI interaction dynamics. This work establishes the first mathematical foundation for operationalizing AI-specific psychological vulnerabilities in cybersecurity contexts.

Keywords: Applied Mathematics, Interdisciplinary Psychology, Computational Statistics, Mathematical Modeling, Cybersecurity Research

1 Introduction and CPF Context

The Cybersecurity Psychology Framework (CPF) represents a paradigm shift from reactive security awareness to predictive vulnerability assessment through psychological state modeling [1]. As artificial intelligence becomes increasingly integrated into security operations, new categories of psychological vulnerabilities emerge that traditional frameworks cannot address.

Category 9 of the CPF addresses AI-Specific Bias Vulnerabilities, representing the first systematic formalization of psychological risks arising from human-AI interaction in security contexts. Unlike traditional cognitive biases that operate purely between humans, AI-specific biases emerge from the unique characteristics of artificial intelligence: opacity, apparent intelligence, and statistical uncertainty.

The mathematical models presented here capture these novel psychological mechanisms through four complementary approaches: (1) anthropomorphization detection through linguistic and behavioral analysis, (2) trust calibration metrics comparing human confidence with AI uncertainty, (3) automation bias quantification through override rate analysis, and (4) team dysfunction modeling through performance degradation metrics.

This category becomes critical as security operations centers increasingly rely on AI-driven threat detection, automated response systems, and machine learning-based risk assessment. The psychological vulnerabilities identified here create systematic blind spots that attackers can exploit through adversarial machine learning, AI-targeted social engineering, and manipulation of human-AI trust dynamics.

2 Theoretical Foundation: Human-AI Psychology

AI-specific vulnerabilities emerge from the intersection of cognitive psychology, human-computer interaction, and machine learning uncertainty. Humans evolved to interact with other conscious agents, creating systematic biases when interacting with artificial intelligence systems [2].

Research demonstrates that humans anthropomorphize AI systems within seconds of interaction, attributing intentions, emotions, and consciousness where none exist [3]. This anthropomorphization creates vulnerabilities as humans apply social cognition heuristics inappropriate for statistical systems.

The uncanny valley effect manifests in AI interactions, creating trust calibration problems where humans either over-trust or under-trust AI systems based on superficial characteristics rather than actual performance [4]. Machine learning opacity exacerbates these issues, as humans cannot inspect AI decision-making processes, leading to either blind trust or complete rejection.

Automation bias, originally identified in aviation psychology [5], takes on new dimensions with AI systems that exhibit apparent intelligence while making statistical rather than logical decisions. The mathematical models presented here formalize these psychological mechanisms for systematic detection and mitigation.

3 Mathematical Formalization

3.1 Universal Detection Framework

Each AI-specific bias indicator employs the unified detection function:

$$D_i(t) = w_1 \cdot R_i(t) + w_2 \cdot A_i(t) + w_3 \cdot U_i(t) + w_4 \cdot T_i(t) \tag{1}$$

where $D_i(t)$ represents the detection score for indicator i at time t, $R_i(t)$ denotes rule-based detection (binary), $A_i(t)$ represents anthropomorphization score (continuous [0,1]), $U_i(t)$ represents uncertainty calibration, and $T_i(t)$ represents trust metrics. Weights sum to unity and are calibrated through organizational AI interaction baselines.

The temporal evolution incorporates AI-specific decay patterns:

$$S_i(t) = \alpha \cdot D_i(t) + (1 - \alpha) \cdot S_i(t-1) \cdot e^{-\beta \cdot AI\_interaction\_gap(t)} \tag{2}$$

where β accounts for rapid trust decay in AI interactions.

3.2 Indicator 9.1: Anthropomorphization of AI Systems

Definition: Attribution of human-like consciousness, intentions, and emotions to AI security systems.

Mathematical Model:

The anthropomorphization index through linguistic analysis:

$$A_{anthro}(t) = \sum_i w_i \cdot f_i(communications(t)) \tag{3}$$

where f_i represents frequency of anthropomorphic markers: pronouns (he/she), emotional attributions (angry, confused), intentional language (wants, thinks, decides).

Behavioral Anthropomorphization:

$$B_{anthro}(t) = \frac{\sum_i social\_gesture\_count(i,t)}{\sum_i total\_AI\_interactions(i,t)} \tag{4}$$

measuring social behaviors directed toward AI systems.

Detection Function:

$$D_{9.1}(t) = \tanh(\alpha \cdot A_{anthro}(t) + \beta \cdot B_{anthro}(t)) \tag{5}$$

Threshold Condition:

$$R_{9.1}(t) = \begin{cases} 1 & \text{if } D_{9.1}(t) > \mu_{baseline} + 2\sigma_{baseline} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

3.3 Indicator 9.2: Automation Bias Override

Definition: Systematic over-reliance on AI recommendations without appropriate human judgment.

Mathematical Model:

The override rate function:

$$OR(t, w) = \frac{\sum_{i \in W(t,w)} Override_i}{\sum_{i \in W(t,w)} AI\_recommendation_i} \tag{7}$$

where $W(t, w)$ represents time window, and $Override_i$ indicates human decision contrary to AI recommendation.

Automation Bias Detection:

$$AB(t) = \max(0, \frac{OR_{expected} - OR(t)}{OR_{expected}}) \tag{8}$$

where $OR_{expected}$ represents calibrated override rate based on AI accuracy.

Confidence-Performance Correlation:

$$CPC(t) = \frac{Cov(AI\_confidence, Human\_acceptance)}{Std(AI\_confidence) \cdot Std(Human\_acceptance)} \tag{9}$$

Detection Threshold:

$$R_{9.2}(t) = \begin{cases} 1 & \text{if } OR(t) < 0.1 \text{ and } AB(t) > 0.3 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

3.4 Indicator 9.3: Algorithm Aversion Paradox

Definition: Simultaneous over-trust and under-trust of AI systems creating inconsistent security decisions.

Mathematical Model:

The aversion-attraction oscillation:

$$AAO(t) = |Trust_{AI}(t) - \overline{Trust_{AI}}| \cdot Frequency_{switches}(t) \tag{11}$$

where $Frequency_{switches}$ measures rapid trust state changes.

Paradox Detection Function:

$$PDF(t) = \frac{Var(Trust_{decisions}(t))}{\overline{Trust_{decisions}(t)}} \cdot Switch_{penalty}(t) \tag{12}$$

Temporal Inconsistency Model:

$$TI(t) = \sum_{i=1}^{n} |d_i(t) - d_i(t-1)| \cdot w_i \tag{13}$$

where $d_i(t)$ represents decision consistency scores.

Detection Function:

$$D_{9.3}(t) = \sqrt{AAO(t) \cdot PDF(t) \cdot TI(t)} \tag{14}$$

3.5 Indicator 9.4: AI Authority Transfer

Definition: Inappropriate transfer of human authority structures to AI systems.

Mathematical Model:

The authority transfer coefficient:

$$ATC(ai, human) = \frac{Compliance_{ai}(t)}{Compliance_{human}(t)} \cdot \frac{Authority_{human}}{Authority_{perceived\_ai}} \tag{15}$$

Hierarchy Confusion Index:

$$HCI(t) = \sum_{i,j} \frac{|Authority_{actual}(i,j) - Authority_{perceived}(i,j)|}{n \cdot (n-1)} \tag{16}$$

Decision Delegation Model:

$$DDM(t) = \frac{\sum_i Critical\_decisions\_delegated\_to\_AI(i)}{\sum_i Total\_critical\_decisions(i)} \tag{17}$$

Detection Threshold:

$$R_{9.4}(t) = \begin{cases} 1 & \text{if } ATC > 1.5 \text{ or } DDM > 0.4 \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

3.6 Indicator 9.5: Uncanny Valley Effects

Definition: Trust disruption caused by AI systems that appear almost-but-not-quite human.

Mathematical Model:

The uncanny valley function following Mori's curve:

$$UV(x) = \begin{cases} \frac{x}{\alpha} & \text{if } x < \alpha \\ \beta - \gamma \cdot e^{-\delta(x-\alpha)^2} & \text{if } \alpha \leq x \leq \xi \\ \beta + \epsilon \cdot (x - \xi) & \text{if } x > \xi \end{cases} \tag{19}$$

where x represents human-likeness and parameters define valley shape.

Trust Disruption Metric:

$$TD(t) = -\frac{d}{dx} UV(Human\_likeness_{AI}(t)) \tag{20}$$

Behavioral Indicators:

$$BI(t) = \sum_i w_i \cdot Avoidance\_behavior_i(t) \tag{21}$$

including hesitation time, interaction reduction, and explicit rejection.

Detection Function:

$$D_{9.5}(t) = TD(t) \cdot BI(t) \cdot Interaction\_frequency\_drop(t) \tag{22}$$

3.7 Indicator 9.6: Machine Learning Opacity Trust

Definition: Misplaced trust due to inability to inspect AI decision-making processes.
Mathematical Model:
The opacity-trust correlation:

$$OTC(t) = \frac{Trust_{opaque\_AI}(t) - Trust_{transparent\_systems}(t)}{Opacity_{index}(t)} \tag{23}$$

Explainability Demand Function:

$$EDF(t) = 1 - e^{-\lambda \cdot Complexity_{perceived}(t)} \tag{24}$$

Black Box Acceptance Rate:

$$BBAR(t) = \frac{\sum_i Accepted_{unexplained\_recommendations}(i)}{\sum_i Total_{AI\_recommendations}(i)} \tag{25}$$

Calibration Metric:

$$CM(t) = |BBAR(t) - Optimal_{acceptance\_rate}(t)| \tag{26}$$

where optimal rate is based on AI system's actual accuracy.
Detection Threshold:

$$R_{9.6}(t) = \begin{cases} 1 & \text{if } CM(t) > 0.3 \text{ and } EDF(t) < 0.2 \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

3.8 Indicator 9.7: AI Hallucination Acceptance

Definition: Failure to recognize and reject AI-generated false information.
Mathematical Model:
The hallucination acceptance function:

$$HA(t) = \frac{\sum_i Accepted_{hallucinations}(i,t)}{\sum_i Total_{AI\_outputs}(i,t)} \tag{28}$$

Confidence-Reality Mismatch:

$$CRM(o) = |AI\_confidence(o) - Ground\_truth\_probability(o)| \tag{29}$$

for output o.
Verification Rate Model:

$$VRM(t) = \frac{\sum_i Verification\_attempts(i,t)}{\sum_i AI\_claims\_requiring\_verification(i,t)} \tag{30}$$

Dangerous Zone Detection:

$$DZD(t) = \sum_o \mathbb{I}[AI\_confidence(o) < 0.6] \cdot \mathbb{I}[Human\_acceptance(o) > 0.8] \tag{31}$$

where $\mathbb{I}$ represents indicator function.
Detection Function:

$$D_{9.7}(t) = HA(t) \cdot (1 - VRM(t)) \cdot DZD(t) \tag{32}$$

3.9 Indicator 9.8: Human-AI Team Dysfunction

Definition: Degraded performance due to poor integration between human judgment and AI capabilities.
Mathematical Model:
The team synergy coefficient:

$$TSC(t) = \frac{Performance_{human+AI}(t)}{Performance_{human}(t) + Performance_{AI}(t)} \tag{33}$$

Role Confusion Matrix:

$$RCM_{ij}(t) = P(Human\_performs\_task_i | AI\_should\_perform\_task_i) \tag{34}$$

Communication Efficiency:

$$CE(t) = \frac{Successful\_handoffs(t)}{Total\_handoff\_attempts(t)} \tag{35}$$

Performance Degradation:

$$PD(t) = \max(0, Performance_{baseline} - TSC(t)) \tag{36}$$

Detection Threshold:

$$R_{9.8}(t) = \begin{cases} 1 & \text{if } TSC(t) < 0.8 \text{ and } CE(t) < 0.7 \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

3.10 Indicator 9.9: AI Emotional Manipulation

Definition: Vulnerability to emotional influence from AI systems designed to appear empathetic.
Mathematical Model:
The emotional manipulation susceptibility:

$$EMS(t) = \sum_i Emotional\_response_i(t) \cdot AI\_emotional\_cue_i(t) \tag{38}$$

Attachment Formation Rate:

$$AFR(t) = \frac{d}{dt}\left(\sum_i Attachment\_indicators_i(t)\right) \tag{39}$$

Decision Bias from Emotion:

$$DBE(t) = |Decision_{emotional\_AI\_present} - Decision_{neutral\_condition}| \tag{40}$$

Parasocial Relationship Index:

$$PRI(t) = \sum_i w_i \cdot Parasocial\_behavior_i(t) \tag{41}$$

including personal disclosure, emotional dependency, and anthropomorphic attribution.
Detection Function:

$$D_{9.9}(t) = EMS(t) \cdot \tanh(AFR(t)) \cdot PRI(t) \tag{42}$$

3.11 Indicator 9.10: Algorithmic Fairness Blindness

Definition: Failure to recognize discriminatory AI behavior due to perceived objectivity.

Mathematical Model:

The fairness blindness coefficient:

$$FBC(t) = \frac{Perceived_{fairness}(t)}{Actual_{fairness}(t)} - 1 \tag{43}$$

Discrimination Detection Sensitivity:

$$DDS(t) = \frac{\sum_i Detected_{bias\_instances}(i,t)}{\sum_i Actual_{bias\_instances}(i,t)} \tag{44}$$

Objectivity Halo Effect:

$$OHE(t) = Trust_{AI\_fairness}(t) - \frac{1}{n}\sum_i Trust_{human\_fairness}(i,t) \tag{45}$$

Bias Rationalization Rate:

$$BRR(t) = \frac{\sum_i Rationalized_{AI\_bias}(i,t)}{\sum_i Observed_{AI\_bias}(i,t)} \tag{46}$$

Detection Threshold:

$$R_{9.10}(t) = \begin{cases} 1 & \text{if } DDS(t) < 0.5 \text{ and } BRR(t) > 0.6 \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

4 Interdependency Matrix

The AI-specific bias indicators exhibit unique interdependencies captured through the correlation matrix $\mathbf{R}_9$:

$$\mathbf{R}_9 = \begin{pmatrix} 1.00 & 0.70 & 0.45 & 0.65 & 0.35 & 0.60 & 0.55 & 0.50 & 0.75 & 0.40 \\ 0.70 & 1.00 & 0.60 & 0.55 & 0.30 & 0.45 & 0.70 & 0.65 & 0.40 & 0.50 \\ 0.45 & 0.60 & 1.00 & 0.40 & 0.50 & 0.35 & 0.45 & 0.55 & 0.35 & 0.45 \\ 0.65 & 0.55 & 0.40 & 1.00 & 0.45 & 0.50 & 0.35 & 0.60 & 0.70 & 0.55 \\ 0.35 & 0.30 & 0.50 & 0.45 & 1.00 & 0.40 & 0.35 & 0.45 & 0.40 & 0.30 \\ 0.60 & 0.45 & 0.35 & 0.50 & 0.40 & 1.00 & 0.75 & 0.55 & 0.45 & 0.65 \\ 0.55 & 0.70 & 0.45 & 0.35 & 0.35 & 0.75 & 1.00 & 0.60 & 0.50 & 0.55 \\ 0.50 & 0.65 & 0.55 & 0.60 & 0.45 & 0.55 & 0.60 & 1.00 & 0.55 & 0.50 \\ 0.75 & 0.40 & 0.35 & 0.70 & 0.40 & 0.45 & 0.50 & 0.55 & 1.00 & 0.45 \\ 0.40 & 0.50 & 0.45 & 0.55 & 0.30 & 0.65 & 0.55 & 0.50 & 0.45 & 1.00 \end{pmatrix} \tag{48}$$

Key interdependencies include:

- Strong correlation (0.75) between Anthropomorphization (9.1) and AI Emotional Manipulation (9.9)
- High correlation (0.75) between ML Opacity Trust (9.6) and AI Hallucination Acceptance (9.7)
- Significant correlation (0.70) between Anthropomorphization (9.1) and Automation Bias (9.2)
- Notable correlation (0.70) between AI Authority Transfer (9.4) and AI Emotional Manipulation (9.9)

Algorithm 1 AI-Specific Bias Vulnerability Assessment

```
Initialize AI interaction baselines $\boldsymbol{\mu}_{AI}, \boldsymbol{\Sigma}_{AI}, \boldsymbol{w}$
for each time step $t$ do
    Collect AI interaction telemetry $\mathbf{x}_{AI}(t)$
    Extract anthropomorphization markers from communications
    Measure AI confidence vs. human acceptance correlation
    for each indicator $i \in \{9.1, 9.2, \ldots, 9.10\}$ do
        Compute $R_i(t)$ using rule-based logic
        Compute $A_i(t)$ using anthropomorphization detection
        Compute $U_i(t)$ using uncertainty calibration
        Compute $T_i(t)$ using trust metrics
        Calculate $D_i(t) = w_1R_i(t) + w_2A_i(t) + w_3U_i(t) + w_4T_i(t)$
        Update temporal state with AI-specific decay
    end for
    Compute interdependency corrections using $\mathbf{R}_9$
    Generate AI-specific alerts and recommendations
    Update baselines with human-AI interaction learning
    Log results for model drift and bias detection
end for
```

5 Implementation Algorithms

6 Validation Framework

AI-specific indicators require specialized validation approaches accounting for human-AI interaction complexity:

Human-AI Performance Metrics:

$$Team\_Effectiveness = \frac{Performance_{human+AI}}{max(Performance_{human}, Performance_{AI})} \tag{49}$$

$$Trust\_Calibration = 1 - |Trust_{human} - Reliability_{AI}| \tag{50}$$

$$Complementarity = \frac{Tasks_{human\_better} + Tasks_{AI\_better}}{Total\_tasks} \tag{51}$$

Anthropomorphization Validation: Ground truth established through explicit consciousness tests:

$$Anthro\_Accuracy = \frac{Correct\_consciousness\_attributions}{Total\_consciousness\_judgments} \tag{52}$$

AI Uncertainty Calibration: Reliability diagrams comparing predicted vs. observed accuracy:

$$Calibration\_Error = \frac{1}{B} \sum_{b=1}^{B} |acc(b) - conf(b)| \cdot \frac{|B_b|}{n} \tag{53}$$

Cross-AI-System Validation: Models validated across different AI architectures:

$$Generalization_{AI} = \frac{1}{k} \sum_{i=1}^{k} Performance(Model, AI\_system_i) \tag{54}$$

Longitudinal Adaptation Tracking: Human adaptation to AI systems over time:

$$Adaptation\_Rate = \frac{d}{dt} Trust\_calibration(t) \tag{55}$$

7 Conclusion

This mathematical formalization of AI-specific bias vulnerabilities establishes the first rigorous framework for assessing psychological risks in human-AI security interactions. The ten indicators capture novel vulnerabilities emerging from artificial intelligence integration, from anthropomorphization effects to algorithmic fairness blindness.

The interdependency matrix reveals important correlations between AI-specific biases, particularly the strong relationship between anthropomorphization and emotional manipulation vulnerability. These correlations enable enhanced detection through multivariate analysis of human-AI interaction patterns.

Implementation algorithms provide clear guidance for integrating AI-specific vulnerability assessment into existing security operations, while validation frameworks ensure continued accuracy as AI systems evolve. The mathematical rigor enables objective measurement of these previously subjective psychological phenomena.

As AI systems become increasingly sophisticated and ubiquitous in security operations, these vulnerabilities will become critical attack vectors. Adversaries are already exploring AI-targeted social engineering, algorithmic poisoning designed to exploit human biases, and manipulation of human-AI trust dynamics.

The AI-specific vulnerability category represents a crucial evolution in cybersecurity psychology, acknowledging that human cognition evolved for interaction with other humans, not artificial intelligence systems. By formalizing these mismatches mathematically, we enable systematic detection and mitigation of vulnerabilities that traditional security frameworks cannot address.

Future work will focus on validation through controlled human-AI interaction studies, development of countermeasures for identified vulnerabilities, and integration with adversarial machine learning defenses. The mathematical foundation provided here enables reproducible research and standardized assessment across diverse AI-integrated security environments.

References

[1] Canale, G. (2024). The Cybersecurity Psychology Framework: A Pre-Cognitive Vulnerability Assessment Model Integrating Psychoanalytic and Cognitive Sciences. *Preprint*.

[2] Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.

[3] Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52, 113-117.

[4] Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33-35.

[5] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.