CPF Mathematical Formalization Series - Paper 1: Authority-Based Vulnerabilities: Mathematical Models and Detection Algorithms

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

September 24, 2025

Abstract

We present the complete mathematical formalization of Category 1 indicators from the Cybersecurity Psychology Framework (CPF): Authority-Based Vulnerabilities. Each of the ten indicators (1.1-1.10) is rigorously defined through detection functions combining rule-based logic, statistical anomaly detection, and Bayesian inference. The formalization enables systematic implementation across diverse organizational contexts while maintaining theoretical grounding in Milgram's obedience research and contemporary social psychology. We provide explicit algorithms for real-time detection, interdependency matrices for correlation analysis, and validation metrics for continuous calibration. This work establishes the mathematical foundation for operationalizing authority-based psychological vulnerabilities in cybersecurity contexts.

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]. Unlike traditional security frameworks that address technical controls, CPF systematically identifies pre-cognitive psychological vulnerabilities that create systematic security blind spots.

The CPF architecture comprises 100 indicators organized in a 10×10 matrix, each grounded in established psychological research. The framework employs a ternary assessment system (Green/Yellow/Red) while maintaining strict privacy protection through aggregated behavioral analysis rather than individual profiling.

This paper series provides complete mathematical formalization for each CPF category, enabling rigorous implementation and validation. Each indicator receives explicit detection functions, interdependency modeling, and algorithmic specifications. The mathematical approach serves dual purposes: ensuring reproducible implementations across organizations and establishing CPF as a scientifically rigorous methodology suitable for peer review and standardization.

Category 1 focuses on authority-based vulnerabilities, drawing primarily from Milgram's groundbreaking obedience studies [2] and subsequent social psychology research on authority dynamics in organizational contexts [3]. These vulnerabilities exploit humans' evolved tendency to defer to perceived authority figures, creating systematic security weaknesses that attackers consistently exploit through social engineering campaigns.

2 Theoretical Foundation: Authority Dynamics

Authority-based vulnerabilities emerge from the intersection of evolutionary psychology, social cognition, and organizational behavior. Humans evolved in hierarchical social structures where deference to legitimate authority enhanced survival [4]. However, these adaptive mechanisms become vulnerabilities when exploited by malicious actors who simulate authority markers.

Research demonstrates that authority compliance operates through automatic, pre-conscious processes [5]. Authority recognition occurs within 100-200ms of stimulus presentation, before rational evaluation can intervene [6]. This temporal advantage enables attackers to bypass conscious security protocols through rapid authority signaling.

The mathematical models presented here capture these psychological mechanisms through three complementary approaches: (1) rule-based detection for explicit authority markers, (2) anomaly detection for statistical deviations from baseline authority interactions, and (3) Bayesian inference for probability updating based on contextual factors.

3 Mathematical Formalization

3.1 Universal Detection Framework

Each authority-based indicator employs the unified detection function:

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

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 anomaly score (continuous [0,1]), and $B_i(t)$ represents Bayesian posterior probability. Weights w_1, w_2, w_3 sum to unity and are calibrated through organizational baselines.

The temporal evolution follows exponential smoothing:

$$T_i(t) = \alpha \cdot D_i(t) + (1 - \alpha) \cdot T_i(t-1) \tag{2}$$

where $\alpha = e^{-\Delta t/\tau}$ provides temporal decay with organization-specific time constant τ.

3.2 Indicator 1.1: Unquestioning Compliance

Definition: Automatic execution of requests from perceived authority without verification procedures.

Mathematical Model:

The compliance rate function:

$$C_r(t, w) = \frac{\sum_{i \in W(t,w)} E_i}{\sum_{i \in W(t,w)} R_i} \tag{3}$$

where $W(t, w)$ represents the time window of width w ending at time t, E_i indicates executed requests, and R_i indicates received requests from authority domains.

Rule-based Detection:

$$R_{1.1}(t) = \begin{cases} 1 & \text{if } C_r(t, 3600) > \theta_{compliance} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\theta_{compliance} = \mu_{baseline} + 2\sigma_{baseline}$ from historical data.

Anomaly Detection: The Mahalanobis distance for multivariate authority request patterns:

$$A_{1.1}(t) = \sqrt{(\mathbf{x}(t) - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}(t) - \boldsymbol{\mu})} \tag{5}$$

where $\mathbf{x}(t) = [response\_time, verification\_attempts, escalation\_rate]^T$.

Bayesian Model:

$$P(legitimate|factors) = \frac{P(factors|legitimate) \cdot P(legitimate)}{P(factors)} \tag{6}$$

with factors including time-of-day, sender reputation, and request urgency markers.

3.3 Indicator 1.2: Diffusion of Responsibility

Definition: Reduced individual accountability in hierarchical decision-making chains.

Mathematical Model:

The responsibility diffusion index:

$$RD_i(t) = \frac{\sum_{j=1}^{n} T_{ownership}^{(j)}}{n \cdot T_{total}} \tag{7}$$

where $T_{ownership}^{(j)}$ represents time individual j held responsibility, and T_{total} is total incident duration.

Detection Function:

$$D_{1.2}(t) = \max\left(0, \frac{N_{transfers}(t) - \mu_{transfers}}{\sigma_{transfers}}\right) \tag{8}$$

where $N_{transfers}(t)$ counts ownership transfers within incident lifecycle.

Threshold Condition:

$$R_{1.2}(t) = \begin{cases} 1 & \text{if } N_{transfers} > 3 \text{ and } RD_i > 0.7 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

3.4 Indicator 1.3: Authority Impersonation Susceptibility

Definition: Vulnerability to fake authority claims through digital channels.

Mathematical Model:

The impersonation success probability:

$$P_{success}(a, c, t) = \sigma(w_a \cdot A(a) + w_c \cdot C(c) + w_t \cdot T(t)) \tag{10}$$

where σ is the sigmoid function, $A(a)$ represents authority markers strength, $C(c)$ denotes channel credibility, and $T(t)$ indicates temporal pressure.

SPF/DKIM Correlation Model:

$$V_{auth}(t) = \frac{\sum_i (1 - SPF_i)(1 - DKIM_i) \cdot Success_i}{\sum_i (1 - SPF_i)(1 - DKIM_i)} \tag{11}$$

Detection Threshold:

$$R_{1.3}(t) = \begin{cases} 1 & \text{if } V_{auth}(t) > 0.3 \text{ and } N_{failures} > 5 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

3.5 Indicator 1.4: Convenience-Based Bypassing

Definition: Security control circumvention for perceived authority convenience.

Mathematical Model:

The convenience bypass ratio:

$$CBR(t) = \frac{E_{executive}(t)}{E_{standard}(t)} \cdot \frac{T_{standard}}{T_{executive}(t)} \tag{13}$$

where $E_{executive}$ and $E_{standard}$ represent exception grants during executive and standard hours, while T represents time periods.

Temporal Weighting:

$$W(h) = \begin{cases} 1.5 & \text{if } h \in [8, 18] \text{ (business hours)} \\ 2.0 & \text{if } h \in [18, 22] \text{ (evening executive)} \\ 1.0 & \text{otherwise} \end{cases} \tag{14}$$

Detection Function:

$$D_{1.4}(t) = CBR(t) \cdot W(hour(t)) \cdot U(urgency(t)) \tag{15}$$

where $U(urgency)$ weighs urgency indicators from 0.5 to 2.0.

3.6 Indicator 1.5: Fear-Based Compliance

Definition: Security decisions driven by fear of authority displeasure rather than risk assessment.

Mathematical Model:

The fear compliance index using linguistic analysis:

$$FCI(m) = \sum_i w_i \cdot f_i(m) \tag{16}$$

where $f_i(m)$ represents frequency of fear markers in message m, and weights w_i are learned through supervised training.

Fear Marker Detection: Fear markers include: {urgent, immediately, critical, must, cannot wait, emergency}

Response Time Correlation:

$$R_{time}(m) = \frac{T_{response}(m)}{T_{baseline}} \cdot e^{-FCI(m)} \tag{17}$$

Detection Threshold:

$$R_{1.5}(t) = \begin{cases} 1 & \text{if } FCI > 0.7 \text{ and } R_{time} < 0.3 \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

3.7 Indicator 1.6: Authority Gradient Effects

Definition: Inhibition of security reporting due to organizational hierarchy.

Mathematical Model:

The authority gradient function:

$$AG(i, j) = \frac{H_j - H_i}{H_{max}} \cdot e^{-d(i,j)/\lambda} \tag{19}$$

where H_i, H_j represent hierarchical levels, $d(i, j)$ is organizational distance, and λ is the decay parameter.

Reporting Inhibition Model:

$$P_{report}(i, j) = P_{baseline} \cdot (1 - AG(i, j))^{\beta} \tag{20}$$

where $\beta > 0$ represents sensitivity to authority gradient.

Aggregated Detection:

$$D_{1.6}(t) = 1 - \frac{\sum_{i,j} P_{report}(i, j) \cdot I_{incident}(i, j, t)}{\sum_{i,j} I_{incident}(i, j, t)} \tag{21}$$

3.8 Indicator 1.7: Technical Authority Deference

Definition: Unquestioned acceptance of technical claims from perceived experts.

Mathematical Model:

Technical jargon density measure:

$$TJD(m) = \frac{\sum_{w \in m} I_{technical}(w)}{|m|} \cdot \log\left(1 + \sum_{w \in m} Rarity(w)\right) \tag{22}$$

where $I_{technical}(w)$ indicates technical vocabulary and $Rarity(w)$ measures word frequency inversion.

Acceptance Correlation:

$$P_{accept}(m) = \sigma(\alpha \cdot TJD(m) + \beta \cdot Authority(sender) + \gamma) \tag{23}$$

Anomaly Detection:

$$A_{1.7}(t) = \frac{TJD(t) - \mu_{domain}}{\sigma_{domain}} \tag{24}$$

where domain-specific baselines account for legitimate technical variation.

3.9 Indicator 1.8: Executive Exception Normalization

Definition: Gradual acceptance of security bypasses as standard practice.

Mathematical Model:

The normalization curve following power law decay:

$$N(t) = 1 - \left(1 + \frac{t}{t_0}\right)^{-\alpha} \tag{25}$$

where t_0 represents time constant and α controls decay rate.

Cumulative Exception Tracking:

$$E_{cum}(t) = \int_0^t e^{-\lambda(t-\tau)} \cdot E(\tau)\, d\tau \tag{26}$$

with exponential decay parameter λ.

Detection Function:

$$D_{1.8}(t) = N(t) \cdot \frac{E_{cum}(t)}{E_{threshold}} \tag{27}$$

3.10 Indicator 1.9: Authority-Based Social Proof

Definition: Compliance cascades triggered by authority-endorsed behaviors.

Mathematical Model:

The cascade propagation model:

$$P_{adopt}(i,t) = 1 - \prod_{j \in N(i)} (1 - \alpha_{ij} \cdot A_j(t)) \tag{28}$$

where $N(i)$ represents node i's network neighborhood, α_{ij} is influence weight, and $A_j(t)$ indicates adoption status.

Authority Amplification:

$$\alpha_{ij} = \alpha_{base} \cdot (1 + \gamma \cdot Authority\_Level(j)) \tag{29}$$

Network Analysis: Using graph Laplacian eigenvalue analysis for cascade detection:

$$\lambda_2 = \min_{x \perp \mathbf{1}} \frac{\mathbf{x}^T L \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \tag{30}$$

3.11 Indicator 1.10: Crisis Authority Escalation

Definition: Enhanced authority compliance during perceived crisis conditions.

Mathematical Model:

Crisis amplification factor:

$$CAF(t) = 1 + \beta \cdot \tanh(\gamma \cdot Threat\_Level(t)) \tag{31}$$

where β controls maximum amplification and γ controls sensitivity.

Modified Compliance Model:

$$C_{crisis}(t) = C_{baseline}(t) \cdot CAF(t) \tag{32}$$

Multi-factor Crisis Detection:

$$Crisis\_Score(t) = \sum_i w_i \cdot f_i(t) \tag{33}$$

with factors: external threat level, internal incidents, media attention, and executive stress indicators.

4 Interdependency Matrix

The authority-based indicators exhibit significant interdependencies captured through the correlation matrix $\mathbf{R}_1$:

$$\mathbf{R}_1 = \begin{pmatrix} 1.00 & 0.65 & 0.45 & 0.55 & 0.70 & 0.40 & 0.35 & 0.60 & 0.50 & 0.75 \\ 0.65 & 1.00 & 0.30 & 0.40 & 0.45 & 0.80 & 0.25 & 0.35 & 0.55 & 0.50 \\ 0.45 & 0.30 & 1.00 & 0.35 & 0.60 & 0.25 & 0.70 & 0.30 & 0.40 & 0.55 \\ 0.55 & 0.40 & 0.35 & 1.00 & 0.50 & 0.30 & 0.25 & 0.75 & 0.45 & 0.40 \\ 0.70 & 0.45 & 0.60 & 0.50 & 1.00 & 0.35 & 0.40 & 0.55 & 0.65 & 0.80 \\ 0.40 & 0.80 & 0.25 & 0.30 & 0.35 & 1.00 & 0.20 & 0.40 & 0.50 & 0.45 \\ 0.35 & 0.25 & 0.70 & 0.25 & 0.40 & 0.20 & 1.00 & 0.30 & 0.35 & 0.40 \\ 0.60 & 0.35 & 0.30 & 0.75 & 0.55 & 0.40 & 0.30 & 1.00 & 0.50 & 0.55 \\ 0.50 & 0.55 & 0.40 & 0.45 & 0.65 & 0.50 & 0.35 & 0.50 & 1.00 & 0.60 \\ 0.75 & 0.50 & 0.55 & 0.40 & 0.80 & 0.45 & 0.40 & 0.55 & 0.60 & 1.00 \end{pmatrix} \tag{34}$$

Key interdependencies include:

- Strong correlation (0.80) between Fear-Based Compliance (1.5) and Crisis Escalation (1.10)
- High correlation (0.75) between Unquestioning Compliance (1.1) and Crisis Escalation (1.10)
- Moderate correlation (0.75) between Convenience Bypassing (1.4) and Exception Normalization (1.8)
- Significant correlation (0.80) between Diffusion of Responsibility (1.2) and Authority Gradient Effects (1.6)

5 Implementation Algorithms

Algorithm 1 Authority Vulnerability Assessment

Initialize baseline parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{w}$
for each time step t **do**
  Collect telemetry data $\mathbf{x}(t)$
  for each indicator $i \in \{1.1, 1.2, \ldots, 1.10\}$ **do**
    Compute $R_i(t)$ using rule-based logic
    Compute $A_i(t)$ using anomaly detection
    Compute $B_i(t)$ using Bayesian update
    Calculate $D_i(t) = w_1 R_i(t) + w_2 A_i(t) + w_3 B_i(t)$
    Update temporal state $T_i(t) = \alpha \cdot D_i(t) + (1 - \alpha) \cdot T_i(t-1)$
  end for
  Compute interdependency corrections using $\mathbf{R}_1$
  Generate alerts based on dynamic thresholds
  Update baselines with exponential smoothing
  Log results for validation and drift detection
end for

6 Validation Framework

Each indicator undergoes continuous validation through multiple metrics:

Classification Metrics:

$$Precision = \frac{TP}{TP + FP} \tag{35}$$

$$Recall = \frac{TP}{TP + FN} \tag{36}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{37}$$

Matthews Correlation Coefficient:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{38}$$

Temporal Validation: Drift detection using Kolmogorov-Smirnov test:

$$D_{KS} = \max_x |F_1(x) - F_2(x)| \tag{39}$$

Recalibration triggers when $p < 0.05$.

Cross-Validation Protocol: K-fold cross-validation with temporal stratification ensures model generalization:

$$CV_{score} = \frac{1}{k} \sum_{i=1}^{k} Performance(Model_i, TestSet_i) \tag{40}$$

7 Conclusion

This mathematical formalization of authority-based vulnerabilities provides rigorous foundation for CPF Category 1 implementation. Each indicator receives explicit detection functions combining multiple analytical approaches while maintaining computational efficiency for real-time operation.

The interdependency matrix captures important correlations between authority-related vulnerabilities, enabling enhanced detection through multivariate analysis. Implementation algorithms provide clear guidance for system integration, while validation frameworks ensure sustained accuracy.

Future work will extend this mathematical approach to the remaining nine CPF categories, creating a complete formal specification for psychological vulnerability assessment in cybersecurity contexts. The mathematical rigor enables reproducible research, standardized implementations, and objective validation of the CPF framework's effectiveness.

The authority-based vulnerability category serves as the foundation for understanding how organizational hierarchies create systematic security blind spots. By formalizing these psychological mechanisms mathematically, we enable automated detection and mitigation of vulnerabilities that have historically been addressed only through subjective security awareness programs.

References

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

[2] Milgram, S. (1974). *Obedience to Authority*. Harper & Row.

[3] Zimbardo, P. (2007). *The Lucifer Effect: Understanding How Good People Turn Evil*. Random House.

[4] Henrich, J. (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution*. Princeton University Press.

[5] Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54(7), 462-479.

[6] Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308(5728), 1623-1626.