Predictive Modeling of Human-Layer Vulnerabilities in Organizational Security: A Framework for AI-Driven Social Engineering Risk Assessment

Kai Aizen ActiveFence | Trident Advisory

Abstract

Social engineering remains the predominant initial access vector in organizational breaches, yet the security industry lacks mature methodologies for assessing human-layer risk with the rigor applied to technical vulnerabilities. Current approaches—primarily simulation-based testing—measure outcomes without modeling underlying susceptibility, equivalent to penetration testing without prior threat modeling. This paper proposes a paradigm shift: predictive human-layer risk assessment using artificial intelligence to model individual psychological profiles, organizational trust dynamics, and attack path viability *before* validation testing. We introduce a three-layer framework (Individual Target Modeling, Organizational Trust Mapping, and Predictive Attack Path Analysis) that enables organizations to quantify human-layer risk, prioritize defensive investments, and conduct targeted validation rather than broad-spectrum simulation. The framework addresses fundamental limitations in existing methodologies and establishes theoretical foundations for treating human vulnerability assessment with the same analytical rigor as technical security assessment.

Keywords: social engineering, human factors security, risk assessment, organizational security, AI security applications, behavioral modeling

---

1. Introduction

The security industry has achieved remarkable sophistication in technical vulnerability assessment. Mature frameworks exist for threat modeling (STRIDE, PASTA), vulnerability scoring (CVSS), and attack path analysis (MITRE ATT&CK). Organizations can quantify technical risk, track remediation progress, and benchmark against industry standards.

No equivalent maturity exists for human-layer security.

Despite social engineering accounting for 74-91% of successful breaches depending on methodology and sector (Verizon DBIR, 2024; Proofpoint State of the Phish, 2024), organizational assessment of human vulnerability remains primitive. The dominant methodology—simulated phishing campaigns—provides point-in-time outcome measurement without predictive capability, addresses single attack vectors while ignoring vishing, pretexting, and physical social engineering, and treats all employees uniformly regardless of role-based risk differentiation.

This paper argues that the fundamental problem is methodological: the industry tests before it models. We propose inverting this approach through Predictive Human-Layer Risk Assessment (PHLRA), leveraging recent advances in artificial intelligence to model human vulnerability with analytical rigor previously impossible.

1.1 Problem Statement

Current social engineering assessment methodologies suffer from five fundamental limitations:

Reactive measurement. Simulations measure who clicked, not who will click. Organizations learn their vulnerability only through testing—equivalent to discovering firewall misconfigurations only through exploitation.

Vector isolation. Phishing simulations assess email-based attacks. Vishing, smishing, pretexting, and physical social engineering remain largely untested. Real adversaries employ multi-vector campaigns; assessments should reflect this.

Individual atomization. Employees are tested as independent units. This ignores that social engineering exploits *relationships*—an attacker who compromises an executive assistant gains pretexting capability against everyone who trusts that assistant.

Uniform treatment. A junior developer and the CFO receive identical phishing simulations despite radically different risk profiles, access levels, and targeting likelihood.

Absence of predictive capability. Organizations cannot answer "if we change X, how does our risk profile change?" because no model exists to perturb.

1.2 Thesis

We propose that human-layer vulnerability can be modeled predictively using three analytical layers:

1. **Individual Target Modeling (ITM):** Psychological and behavioral profiling of individuals to predict susceptibility to specific social engineering tactics.
2. **Organizational Trust Mapping (OTM):** Graph-based modeling of trust relationships to identify attack pivot opportunities and blast radius potential.
3. **Predictive Attack Path Analysis (PAPA):** Integration of ITM and OTM to model likely attack sequences and their organizational impact.

These layers, implemented through AI-driven analysis, enable a fundamental shift from reactive testing to predictive risk management.

---

2. Background and Related Work

2.1 Social Engineering Taxonomy

Established taxonomies categorize social engineering by technique: phishing, pretexting, baiting, quid pro quo, tailgating, and variants (Mitnick & Simon, 2002; Hadnagy, 2018). The SET Framework (Aizen, 2025) extends this taxonomy with emerging threats including AI-generated content, deepfake impersonation, and IoT-based social engineering.

While taxonomies provide common vocabulary, they do not address assessment methodology—*how* to measure organizational susceptibility across these categories.

2.2 Current Assessment Approaches

Simulation-based testing remains the industry standard. Organizations deploy simulated phishing emails and measure click rates, credential submission rates, and reporting rates

(KnowBe4, Proofpoint, Cofense platforms). Meta-analyses suggest organizational click rates range from 10-30% depending on simulation sophistication (Lain et al., 2022).

Limitations are well-documented: simulation performance correlates weakly with real-world attack susceptibility (Wash & Cooper, 2018), repeated testing produces habituation rather than genuine awareness (Caputo et al., 2014), and single-vector testing ignores the multi-channel nature of sophisticated attacks.

Awareness training complements simulation but measures knowledge acquisition, not behavioral change under pressure. The intention-behavior gap is well-established in security contexts (Bada et al., 2019).

Red team assessments provide realistic multi-vector testing but lack scalability and produce anecdotal rather than systematic organizational risk profiles.

2.3 Human Factors Research

Psychological factors influencing social engineering susceptibility are well-studied: authority compliance (Milgram, 1963; Cialdini, 2006), urgency effects on decision quality (Ariely & Zakay, 2001), trust heuristics in digital communication (Riegelsberger et al., 2005), and individual differences in susceptibility (Halevi et al., 2015).

This research establishes *that* psychological factors matter but does not provide *operational frameworks* for incorporating these factors into organizational risk assessment.

2.4 AI Applications in Security

Machine learning applications in cybersecurity focus predominantly on defensive detection: phishing email classification (Fette et al., 2007), anomaly detection (Chandola et al., 2009), and threat intelligence correlation. Offensive applications remain underexplored in academic literature, though adversarial use of generative AI for social engineering content is documented in threat intelligence reporting (Mandiant, 2024).

The application of AI to *model* human vulnerability—rather than detect attacks or generate attack content—represents a gap this paper addresses.

---

3. Theoretical Framework

3.1 The Assessment-Before-Testing Principle

Technical security assessment follows a logical sequence: asset inventory → threat modeling → vulnerability assessment → penetration testing → remediation. Testing validates models; it does not substitute for them.

Human-layer security inverts this sequence, proceeding directly to testing without prior modeling. We propose restoring logical sequence:

```
Human Asset Characterization → Threat Modeling →
Vulnerability Modeling → Targeted Testing → Remediation
```

3.2 Individual Target Modeling (ITM)

Theoretical basis. Individual susceptibility to social engineering varies systematically based on psychological traits, communication patterns, role characteristics, and contextual factors. These variations are inferable from observable data.

Modeling dimensions:

Psychological profile. Big Five personality traits correlate with social engineering susceptibility: agreeableness increases compliance with requests (Halevi et al., 2015), conscientiousness affects policy adherence (Shropshire et al., 2015), neuroticism influences response under pressure.

Communication patterns. Response latency, formality level, question-asking behavior, and challenge frequency indicate baseline behavioral tendencies relevant to social engineering resistance.

Role exposure. External-facing roles, financial authority, administrative access, and organizational visibility create differential targeting likelihood and impact potential.

Information exposure. Publicly available information (professional social media, publications, organizational directories) determines attacker reconnaissance capability and pretext personalization potential.

Output specification. ITM produces per-individual risk profiles including:

- Tactic-specific susceptibility scores
- Predicted success rates for defined attack categories
- Optimal pretext characteristics
- Contextual vulnerability factors (timing, channel, authority level)

3.3 Organizational Trust Mapping (OTM)

Theoretical basis. Social engineering exploits trust relationships. Organizational risk is not the sum of individual risks but a function of how trust flows through organizational structure.

Graph formalization. OTM constructs a weighted directed graph $G = (V, E, w)$ where:

- $V = \{\text{individuals in scope}\}$
- $E = \{(i, j) \mid i \text{ trusts } j \text{ for some organizational function}\}$
- $w: E \rightarrow \mathbb{R}^+$ representing trust strength

Trust edges derive from:

- Formal reporting relationships
- Communication frequency patterns
- Approval chain dependencies
- Collaborative work relationships
- Tenure and organizational memory

Key metrics derived from OTG:

Trust centrality. Individuals with high in-degree (many trust them) represent high-value targets whose compromise enables broad pretexting.

Bridge score. Individuals connecting otherwise-separate organizational clusters enable attack paths across trust boundaries.

Blast radius. Given compromise of individual i, the subgraph reachable through trust edges represents potential downstream impact.

3.4 Predictive Attack Path Analysis (PAPA)

Theoretical basis. Sophisticated social engineering attacks are multi-stage. Initial compromise enables subsequent attacks through gained trust, information, or access. Attack paths can be modeled as traversals through the organizational trust graph, constrained by individual susceptibility and available pretexts.

Path modeling. An attack path P is a sequence $[(i_1, t_1), (i_2, t_2), \ldots, (i_n, t_n)]$ where i_k is a target individual and t_k is the tactic employed. Path viability depends on:

- $P(\text{success} \mid i_1, t_1)$: Initial compromise probability from ITM
- $P(\text{success} \mid i_k, t_k, \text{compromised}(i_1 \ldots i_{k-1}))$: Pivot probability given prior compromise
- $\text{Impact}(i_n)$: Organizational impact of terminal compromise

Path prioritization. Paths are ranked by expected impact:

$$E[\text{Impact}(P)] = \text{Impact}(i_n) \times \prod_k P(\text{success} \mid i_k, t_k, \text{context})$$

High expected-impact paths represent priority assessment and remediation targets.

---

4. Methodological Implications

4.1 Role of AI in Implementation

The framework requires analytical capabilities exceeding traditional rule-based systems:

Natural language understanding for psychological inference from communication samples, public profiles, and organizational documents.

Relationship inference from communication patterns, organizational metadata, and behavioral signals.

Contextual reasoning for pretext generation and attack path viability assessment.

Large language models (LLMs) possess these capabilities. Their application to organizational security modeling—rather than attack generation or defense detection—represents a novel and underexplored domain.

4.2 Validation Testing Protocol

PHLRA does not eliminate simulation-based testing; it transforms testing from discovery mechanism to validation mechanism. Post-modeling, testing serves to:

1. **Validate predictions.** Test highest-predicted-probability paths to confirm model accuracy.
2. **Calibrate weights.** Adjust model parameters based on prediction-outcome divergence.
3. **Address uncertainty.** Test cases where model confidence is low.

This approach dramatically increases testing efficiency by focusing resources on consequential validations rather than broad-spectrum simulation.

4.3 Organizational Resilience Metrics

The framework enables quantified organizational metrics previously unavailable:

Organizational Resilience Index (ORI): Composite score integrating individual vulnerability distribution, trust network characteristics, and attack path exposure.

Maturity classification: Based on ORI, organizations can be classified into maturity levels enabling benchmarking and progress tracking.

What-if analysis: Perturbation of model inputs (organizational structure changes, training interventions, access modifications) yields predicted ORI changes, enabling evidence-based defensive investment.

---

5. Limitations and Future Work

5.1 Data Requirements

ITM requires access to communication samples and behavioral data, raising privacy and consent considerations. Organizations must balance assessment fidelity against employee privacy expectations. Federated or privacy-preserving approaches merit investigation.

5.2 Model Validation

Validating predictive models of human behavior presents methodological challenges. Ground truth requires either longitudinal studies correlating predictions with actual incident data or controlled testing introducing ethical considerations.

5.3 Adversarial Considerations

Sophisticated adversaries may probe for organizational use of predictive assessment and attempt to manipulate model inputs. Robustness against adversarial manipulation requires investigation.

5.4 Cultural Factors

Trust dynamics, authority response, and communication patterns vary across cultures. Model calibration for multi-national organizations requires attention to cultural variables.

---

6. Conclusion

The security industry's approach to human-layer assessment has lagged decades behind technical security methodology. While technical vulnerability management progressed from ad-hoc

penetration testing to risk-based vulnerability management with mature scoring and prioritization frameworks, human vulnerability assessment remains in the "test and see what happens" era.

This paper proposes a path forward: predictive modeling of human vulnerability through AI-driven analysis of individual characteristics, organizational trust dynamics, and attack path viability. The framework enables organizations to quantify human-layer risk, predict attack success, prioritize defensive investment, and validate predictions through targeted testing.

The tools now exist to treat human vulnerability with analytical rigor. What remains is methodological adoption.

References

Ariely, D., & Zakay, D. (2001). A timely account of the role of duration in decision making. *Acta Psychologica*, 108(2), 187-207.

Aizen, K. (2025). Social Engineering Testing (SET) Framework: Tactical Assessment & Resilience Blueprint.

Bada, M., Sasse, A. M., & Nurse, J. R. (2019). Cyber security awareness campaigns: Why do they fail to change behaviour? *International Conference on Cyber Security for Sustainable Society*.

Caputo, D. D., Pfleeger, S. L., Freeman, J. D., & Johnson, M. E. (2014). Going spear phishing: Exploring embedded training and awareness. *IEEE Security & Privacy*, 12(1), 28-38.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1-58.

Cialdini, R. B. (2006). *Influence: The Psychology of Persuasion*. Harper Business.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. *Proceedings of the 16th International Conference on World Wide Web*, 649-656.

Hadnagy, C. (2018). *Social Engineering: The Science of Human Hacking*. Wiley.

Halevi, T., Lewis, J., & Memon, N. (2015). A pilot study of cyber security and privacy related behavior and personality traits. *Proceedings of the 24th International Conference on World Wide Web*, 737-744.

Lain, D., Kostiainen, K., & Capkun, S. (2022). Phishing in organizations: Findings from a large-scale and long-term study. *IEEE Symposium on Security and Privacy*, 842-859.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4), 371-378.

Mitnick, K. D., & Simon, W. L. (2002). *The Art of Deception*. Wiley.

Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2005). The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies*, 62(3), 381-422.

Shropshire, J., Warkentin, M., & Sharma, S. (2015). Personality, attitudes, and intentions: Predicting initial adoption of information security behavior. *Computers & Security*, 49, 177-191.

Wash, R., & Cooper, M. M. (2018). Who provides phishing training? Facts, stories, and people like me. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-12.

Corresponding author: Kai Aizen