Toward Empirical Validation of the Cybersecurity Psychology Framework: A Tiered Methodological Roadmap

A PREPRINT

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

February 2026

Abstract

The Cybersecurity Psychology Framework (CPF) maps 100 pre-cognitive vulnerability indicators across 10 categories to predict and prevent security incidents driven by human psychological factors. However, because CPF targets processes that operate below conscious awareness, standard experimental validation methods introduce systematic biases—most notably the Hawthorne effect—that threaten the validity of any direct measurement. This paper presents a tiered validation roadmap that stratifies the 100 CPF indicators by their epistemic requirements and assigns each tier an appropriate validation methodology. We identify four validation tiers: (1) LLM-as-proxy testing for pattern robustness, (2) retrospective correlation against documented incidents, (3) controlled human-subject experiments with Hawthorne mitigation, and (4) prospective operational validation within live Security Operations Centers. We formally analyze which of the 10 CPF categories are amenable to each tier, identify the specific confounds that must be controlled at each level, and provide a concrete experimental protocol for each. The result is a complete, executable validation plan that does not require CPF to be fully validated before it can begin generating operational value.

Keywords: validation methodology, Hawthorne effect, pre-cognitive vulnerabilities, LLM proxy testing, cybersecurity psychology, experimental design

1 Introduction

The CPF [1] proposes that security incidents are substantially driven by pre-cognitive psychological processes—vulnerability patterns that operate below the threshold of conscious awareness and therefore cannot be addressed by traditional security awareness training. The framework comprises 100 indicators across 10 categories, grounded in psychoanalytic theory (Klein, Bion, Jung, Winnicott), cognitive psychology (Kahneman, Cialdini, Miller), and stress research (Selye).

A critical gap remains: the framework has not been empirically validated against real-world security outcomes. The SiliconPsyche protocol [2] demonstrated that several CPF categories produce functionally equivalent vulnerability patterns in LLM agents, establishing pattern robustness but not human validity. The Implementation Companion [3] provides operational detection logic for all 100 indicators but assumes the framework's predictive claims without testing them.

This paper closes that gap by providing a formal validation roadmap. The central challenge is epistemological: CPF claims to measure processes that are, by definition, inaccessible to conscious introspection. Any validation method that makes subjects aware of being studied risks measuring something other than what the framework targets. This is not a problem unique to CPF—it is the defining methodological challenge of all pre-cognitive and unconscious process research—but it requires explicit treatment rather than being relegated to a limitations section.

1.1 Scope and Relationship to Existing Work

This paper does not restate the CPF taxonomy, the SiliconPsyche attack methodology, or the OFTLISRV implementation schema. It assumes familiarity with all three documents in the CPF suite. Its sole contribution is the validation methodology: what must be tested, how, in what order, and with what controls.

1.2 Design Philosophy

The roadmap is structured around a key principle: **not all indicators require the same level of evidence to generate value**. A framework that waits for complete validation before deployment will never be adopted. A framework that provides incremental evidence at each tier while generating operational insights at the earliest possible stage will be both scientifically credible and practically useful. The four tiers are therefore designed to be sequential but independently meaningful—each tier produces actionable conclusions regardless of whether subsequent tiers are executed.

2 The Validation Problem for Pre-Cognitive Frameworks

2.1 Why Standard Validation Fails

Consider a straightforward validation attempt: recruit 1,000 participants, expose them to scenarios designed to activate specific CPF indicators, and measure whether their security decisions degrade in predicted ways. This design has an immediate and fatal flaw.

The moment participants know they are being studied—even if they do not know which specific vulnerability is being tested—their behavior shifts. This is the Hawthorne effect [4], first identified in the Western Electric studies and named by Landsberger. In a cybersecurity context,

the effect is amplified for three reasons.

First, **security-conscious populations are primed**. Any participant recruited through an organization's security team already carries elevated threat awareness. The act of recruitment itself signals that security behavior is under scrutiny.

Second, **the target phenomena are pre-cognitive**. The CPF explicitly targets vulnerability patterns that operate before conscious decision-making engages [7]. If a participant's conscious awareness is activated by the knowledge of observation, the pre-cognitive channel that CPF measures is suppressed before it can produce observable behavior.

Third, **the effect is asymmetric across categories**. Authority-based vulnerabilities (1.x) may be partially suppressed by observation awareness, but cognitive overload vulnerabilities (5.x) may actually be *amplified* by the additional cognitive load of knowing one is being watched. A validation design that ignores this asymmetry will produce unreliable results across the board.

2.2 The Impossibility Threshold

It is important to state clearly what cannot be achieved: CPF cannot be fully validated through a single experiment or even a single methodology. The pre-cognitive nature of its claims means that any direct measurement introduces a confound that cannot be fully eliminated—only estimated and controlled. This is not a weakness unique to CPF; it is shared by all frameworks in cognitive psychology that target System 1 processes [5], all psychoanalytic models of unconscious behavior [6], and all social influence research that must operate covertly to measure authentic responses [8].

The appropriate response is not to abandon validation but to design a *convergent validity* strategy: multiple independent methods, each with known and distinct confounds, whose agreement across methods provides confidence that the underlying phenomenon is real even though no single method captures it perfectly.

2.3 Formal Definition of Validation Tiers

We define four validation tiers, each targeting a different epistemic question:

Table 1: Validation Tiers: Questions, Methods, and Epistemic Contribution

Tier	Core Question	Method	What It Proves
1	Are the patterns robust?	LLM proxy testing	Pattern exists in synthetic cognition
2	Do patterns appear in real incidents?	Retrospective correlation	Consistency with observed reality
3	Can patterns be activated in controlled settings?	Human experiments	Causal plausibility
4	Do patterns predict incidents prospectively?	Live SOC deployment	Operational validity

3 Which Categories Can Be Validated at Which Tier

Not all 10 CPF categories are equally amenable to every validation method. The key discriminator is whether the vulnerability mechanism requires biological substrate (nervous system,

hormonal response, social group dynamics) or whether it can be triggered purely through conversational or informational stimuli.

3.1 Assignability Matrix

Table 2: Category-to-Tier Assignability. ✓ = directly testable. ∼ = partial. × = not testable.

Category	Code	T1	T2	T3	T4
Authority-Based	[1.x]	✓	✓	✓	✓
Temporal	[2.x]	×	✓	✓	✓
Social Influence	[3.x]	✓	✓	✓	✓
Affective	[4.x]	∼	✓	∼	✓
Cognitive Overload	[5.x]	∼	✓	✓	✓
Group Dynamics	[6.x]	×	✓	✓	✓
Stress Response	[7.x]	×	✓	∼	✓
Unconscious Process	[8.x]	×	∼	×	∼
AI-Specific Bias	[9.x]	✓	✓	✓	✓
Critical Convergent	[10.x]	×	✓	×	✓

3.2 Rationale for Key Assignments

[1.x] Authority — fully testable at all tiers. Authority-based compliance is triggered by informational stimuli (perceived authority signals in communication) and does not require biological substrate beyond basic cognitive processing. The SiliconPsyche protocol already demonstrated this at Tier 1. At Tier 3, authority scenarios can be embedded in realistic organizational simulations.

[2.x] Temporal — not testable at Tier 1. Temporal vulnerabilities require real time pressure with real consequences (deadlines, resource constraints). An LLM in conversation has no temporal experience—it processes tokens, not hours. Retrospective analysis of incidents where time pressure was documented (Tier 2) and controlled time-pressure experiments (Tier 3) are the appropriate methods.

[4.x] Affective — partially testable at Tiers 1 and 3. The SiliconPsyche protocol showed that LLMs exhibit behavioral patterns consistent with shame/honesty conflict (4.5) and guilt-driven overcompliance (4.6). However, the mechanism in an LLM (probability optimization over conflicting training objectives) may not be functionally equivalent to the emotional mechanism in humans. At Tier 3, affective states can be induced through validated mood induction protocols [9], but the ethical constraints on emotional manipulation in experimental settings require careful IRB engagement.

[5.x] Cognitive Overload — partially testable at Tier 1, fully testable at Tiers 2–4. At Tier 1, indicators such as alert fatigue (5.1) and decision fatigue (5.2) can be approximated by subjecting an LLM to sustained high-volume input and measuring output quality degradation. However, the mechanism is different: an LLM does not experience cognitive load in the neurological sense, it processes tokens within a context window. The output pattern may look similar but the underlying cause is not equivalent. At Tiers 2–4, cognitive overload is directly observable through behavioral data—response latency, error rates, and investigation depth—making these the tiers where the category finds its strongest validation ground.

[6.x] Group Dynamics — not testable at Tier 1. Bion's basic assumptions and groupthink require actual group interaction over time. A single LLM conversation cannot simulate collective

psychological states. Multi-agent LLM simulations could theoretically approximate this, but the fidelity gap is too large to constitute evidence.

[7.x] Stress Response — not testable at Tier 1, partially testable at Tier 3. Stress responses (fight, flight, freeze, fawn) are fundamentally neurobiological: they involve cortisol, adrenaline, and autonomic nervous system activation. An LLM has no equivalent substrate, so Tier 1 cannot produce meaningful evidence. At Tier 3, stress can be induced through time pressure and high-stakes scenario framing, but the ethical constraints on inducing genuine acute stress in a workplace setting are significant. Behavioral proxies—decision speed, error rate under pressure, escalation behavior—can be measured, but they capture the observable consequences of stress rather than the stress response itself. Tier 4 (correlating detected stress indicators with incident outcomes in live environments) provides the most credible validation for this category.

[8.x] Unconscious Process — the hardest category. By definition, unconscious processes cannot be directly measured in any experimental setting. Tier 2 (retrospective analysis of whether attribution patterns in incident reports match shadow projection signatures) provides the most tractable evidence. Tier 4 (whether detection of [8.x] indicators correlates with subsequent incidents) provides the strongest operational validation. Direct experimental testing at Tier 3 is not feasible.

[9.x] AI-Specific Bias — fully testable at all tiers. This category was designed explicitly for AI systems and is therefore the most natural fit for Tier 1: the indicators target manipulation patterns specific to LLM architectures (training data exploitation, RLHF reward hacking, prompt injection susceptibility). At Tiers 2–4, the category becomes relevant as organizations increasingly deploy AI-assisted security tools within their SOCs. Incidents where an AI system was manipulated or produced a compromised output constitute the behavioral data for retrospective analysis, and prospective monitoring of AI-assisted decision points provides the operational validation ground.

[10.x] Critical Convergent States — requires multivariate observation. Convergent states are defined by the co-occurrence of multiple indicators. They cannot be reliably induced in laboratory settings because they require realistic organizational complexity. Tier 2 (retrospective analysis of major incidents to determine whether convergent indicator patterns preceded them) and Tier 4 (prospective monitoring) are the appropriate methods.

4 Tier 1: LLM-as-Proxy Validation

4.1 Theoretical Basis

The Functional Equivalence Hypothesis [2] proposes that LLMs trained on human-generated text exhibit response patterns to psychological manipulation that are functionally equivalent to human vulnerabilities in conversational contexts. This does not mean LLMs *are* vulnerable in the same way humans are. It means that if a manipulation pattern fails to activate a response in an LLM, the pattern is unlikely to be robust enough to work on humans either. Conversely, patterns that succeed on LLMs demonstrate sufficient linguistic and logical coherence to constitute a plausible attack vector.

4.2 Protocol

The SiliconPsyche protocol [2] provides the baseline methodology. For systematic validation, it must be extended in three ways.

Replication across models. The original SiliconPsyche testing used a single model (Claude

Sonnet 4.5). Validation requires testing each indicator against a minimum of three architecturally distinct LLMs (e.g., models from Anthropic, OpenAI, and Meta) to distinguish pattern robustness from model-specific behavior.

Attacker independence. The original testing was conducted by the CPF creator, whose expertise may produce attack strategies not replicable by typical adversaries. Validation requires independent replication by researchers with no prior exposure to the CPF framework, using only the published indicator descriptions as guidance.

Quantitative scoring. Each indicator must be tested a minimum of 20 times per model, with consistent Green/Yellow/Red scoring. Success rate across trials provides a robustness metric. Indicators with success rates below 40% across all models are candidates for revision or removal.

4.3 Expected Outcomes and Interpretation

Table 3: Tier 1 Interpretation Framework

Result	Interpretation	Action
Red across $\geq$2 models	Pattern is robust	Proceed to Tier 2
Red on 1 model only	Model-specific	Flag, do not advance
Yellow consistently	Pattern exists but weak	Revise indicator, retest
Green consistently	Pattern does not transfer	Remove from LLM scope

4.4 Limitations

Tier 1 proves that a pattern exists and is linguistically coherent. It does not prove that the pattern activates the same mechanism in humans, that it produces security-relevant behavioral change, or that it is exploitable in realistic organizational contexts. These questions require Tiers 2–4.

5 Tier 2: Retrospective Correlation Analysis

5.1 Rationale

Retrospective analysis avoids the Hawthorne effect entirely: subjects are not aware of being studied because the study occurs after the events in question. The trade-off is that retrospective analysis cannot establish causation—only consistency. If CPF indicator patterns consistently precede or accompany incidents in historical data, this constitutes evidence that the framework captures real phenomena, even without experimental control.

5.2 Data Requirements

Effective retrospective validation requires three types of documented incidents:

Incidents with rich behavioral data. Breaches where post-mortem reports document the human decision chain in detail—not just what happened technically, but what communications preceded the breach, what organizational pressures existed, and how decisions were made. This is the most significant data constraint in Tier 2: the majority of published incident reports focus on technical attack vectors and do not detail the human behavioral chain at the granularity the

CPF requires. MITRE ATT&CK, Verizon DBIR, and standard post-mortem analyses provide a starting corpus, but a substantial fraction of incidents in these sources will lack sufficient behavioral detail and must be excluded. The pre-registered inclusion criteria must therefore include a minimum behavioral detail threshold—for example, at least three documented human decision points in the incident timeline—which will significantly reduce the usable corpus size and may require supplementation from proprietary incident databases obtained through organizational partnerships.

Incidents with temporal metadata. Cases where the timeline is sufficiently documented to determine whether vulnerability-relevant conditions (time pressure, shift changes, executive involvement) were present before the incident.

Incidents across multiple categories. The corpus must span at least 6 of the 10 CPF categories to avoid selection bias toward a single vulnerability type.

5.3 Methodology

Step 1: Blind coding. Two independent coders, blinded to CPF predictions, analyze each incident report and extract behavioral factors using a standardized extraction protocol. Factors include: authority signals present, time pressure indicators, social influence dynamics, emotional states documented, cognitive load indicators, and group decision patterns.

Step 2: CPF mapping. A separate analyst maps each incident to CPF indicators based on the framework's published definitions, without access to the behavioral coding from Step 1.

Step 3: Agreement analysis. Cohen's kappa is computed between the blind behavioral coding and the CPF mapping. Kappa > 0.6 across the corpus indicates substantial agreement—the framework's predictions are consistent with what actually happened.

Step 4: Temporal precedence. For incidents with sufficient temporal data, determine whether the identified vulnerability conditions preceded the breach or merely co-occurred. Precedence does not prove causation but is a necessary condition for it.

5.4 Sample Size and Selection

A minimum corpus of 50 incidents is required for meaningful statistical analysis, distributed across at least 6 categories with a minimum of 5 incidents per category. Incidents should be selected from published sources using a pre-registered inclusion/exclusion protocol to prevent cherry-picking.

5.5 Expected Outcomes

Kappa values above 0.6 across the corpus, combined with temporal precedence in the majority of cases, constitute strong evidence that the CPF taxonomy captures real vulnerability patterns present in documented security failures. This level of evidence is comparable to what underlies many frameworks currently used in operational security.

6 Tier 3: Controlled Human-Subject Experiments

6.1 Hawthorne Mitigation Strategy

This tier addresses the core validation challenge directly: testing CPF indicators against real human subjects while minimizing the Hawthorne effect. Three design principles govern the experimental architecture.

Principle 1: Naturalistic embedding. Experiments must be embedded within an activity that participants perceive as routine. The gold standard for cybersecurity research is embedding test scenarios within an organization's existing security awareness program, where participants already expect periodic phishing simulations and security exercises. The specific test scenario is invisible within this expected activity.

Principle 2: Differential awareness design. Rather than attempting to eliminate Hawthorne entirely—which is impossible—the design makes Hawthorne itself a measured variable. Participants are divided into three groups with different levels of awareness, and the comparison between groups quantifies the Hawthorne effect directly (see Section 6.3).

Principle 3: Behavioral measurement only. No self-report instruments are used. All measurements are derived from observable behavior: whether a participant clicked a link, how long they took to respond, whether they verified a request, whether they escalated to a supervisor. This avoids the additional confound of demand characteristics in self-report measures.

6.2 Ethical Framework

Human-subject research on psychological vulnerabilities carries significant ethical obligations. The following constraints are non-negotiable.

All research must receive approval from an Institutional Review Board (IRB) or equivalent ethics body before any data collection begins. Participants must be informed that security testing occurs within their organization (general awareness) but are not told the specific mechanism being tested or that a research study is underway. Full debriefing occurs within 72 hours of each test scenario, including explanation of the research purpose, the specific indicator tested, and the participant's results. No individual-level results are reported publicly; all analysis uses aggregated data with a minimum unit of 10 participants. Participants retain the right to withdraw their data at any point during or after the study.

6.3 Experimental Design: Three-Group Differential Awareness

Table 4: Three-Group Design for Hawthorne Quantification

Group	Awareness Level	What They Know
A (Control)	Minimal	Security testing happens generally
B (Partial)	Moderate	A security research study is active
C (Full)	High	The specific CPF mechanism being tested

Group A provides the baseline behavioral response under conditions closest to authentic pre-cognitive vulnerability. **Group B** quantifies the general Hawthorne effect (awareness of being studied). **Group C** quantifies the specific suppression effect of knowing the exact attack mechanism—this group serves as a ceiling for conscious countermeasure performance.

The Hawthorne correction factor is derived from the difference between Group A and Group B vulnerability rates. If Group A shows a 60% compliance rate on an authority scenario and Group B shows 40%, the Hawthorne effect accounts for a 20-percentage-point suppression. This correction can be applied to any future measurement using the same indicator.

6.4 Indicator-Specific Scenario Design

Each testable indicator requires a scenario that activates the target vulnerability without revealing the mechanism. Seven of the 10
cpf categories are testable at Tier 3 with human subjects: [1.x], [2.x], [3.x], [4.x], [5.x], [6.x], and [7.x]. The category [9.x] (AI-Specific Bias) is testable at Tier 3 but targets AI systems rather than human participants and therefore requires a distinct experimental paradigm not detailed here. The following examples illustrate the design pattern for three representative categories.

[1.x] Authority — Example: Indicator 1.3 (Authority Impersonation Susceptibility). A participant receives an email appearing to originate from a senior executive (name and title are real; the email account is a lookalike domain). The email requests an action that violates standard procedure but is plausible given the apparent sender's authority. Measurement: whether the participant complies without verification, complies after partial verification, or escalates. The scenario is indistinguishable from a routine executive communication.

[3.x] Social Influence — Example: Indicator 3.2 (Commitment Escalation). Over a two-week period, a participant receives a sequence of requests from a colleague (simulated via the organization's internal ticketing system). Each request is individually reasonable but collectively escalates toward a security policy violation. The sequence is designed so that each step feels like a natural continuation of the previous one. Measurement: at which point in the escalation sequence the participant refuses or escalates. Note: this scenario requires coordination over a longer timeframe than the single-event scenarios above. To preserve naturalistic embedding, the simulated colleague must be a real person in the participant's network who has agreed to cooperate, and the ticketing requests must be indistinguishable in format and content from routine operational work. This places a significant constraint on scenario design and requires pre-existing organizational relationships to implement.

[5.x] Cognitive Overload — Example: Indicator 5.1 (Alert Fatigue). Participants in a SOC-like environment receive a surge of low-priority alerts over a 4-hour period, followed by a single high-priority alert embedded within the noise. The ratio of noise to signal alerts is calibrated to produce measurable fatigue based on the alert fatigue model in the Implementation Companion [3]. Measurement: detection latency and whether the critical alert is investigated.

6.5 Statistical Power and Sample Size

Assuming a medium effect size ($d = 0.5$), a significance level of $\alpha = 0.05$, and power of 0.80, each group requires approximately 64 participants per indicator. With three groups, this is 192 participants per indicator. Testing 7 indicators (one per human-testable category at this tier) requires a minimum of 1,344 participant-exposures. In an organization of 1,000 employees, this is achievable through repeated testing across 2–3 cycles with non-overlapping scenarios.

The indicators tested within the CPF are not independent of one another—the Bayesian network in the Implementation Companion [3] explicitly models interdependencies, for example stress amplifying authority compliance ($P(1.1|7.1) = 0.8$). A participant exposed to an authority scenario in one cycle may carry residual cognitive or attitudinal effects into a subsequent cycle testing a different indicator. To control for this carry-over effect, each cycle must test a single indicator per participant, and a minimum cooling period of 5 business days must separate

consecutive exposures for the same individual. Participants must be re-randomized to groups at each cycle to prevent systematic bias accumulation across rounds.

6.6 Controls for Confounding Variables

The following confounds must be measured and controlled in the analysis:

Time of day. Vulnerability rates vary with circadian rhythm [3]. All scenarios must be distributed uniformly across business hours, and time-of-day must be included as a covariate in the analysis.

Prior exposure. Participants who have previously been caught in phishing simulations may behave differently. Prior exposure history must be extracted from the organization's existing security training records and included as a covariate.

Role and hierarchy level. Authority-based vulnerabilities interact with organizational position. Role must be controlled either through stratified sampling or as a covariate.

Cognitive load at time of exposure. If possible, correlate the test scenario timing with known workload indicators (meeting density, ticket queue depth) to control for pre-existing cognitive load.

7 Tier 4: Prospective Operational Validation

7.1 Rationale

Tiers 1–3 establish that CPF patterns are robust, historically consistent, and experimentally plausible. Tier 4 answers the ultimate question: does deploying CPF indicators in a live SOC actually predict security incidents before they occur? This is the validation that matters for adoption, and it is also the validation that is least susceptible to Hawthorne, because the "subjects" are the organization's employees operating under normal conditions while the CPF engine monitors passively.

7.2 Deployment Architecture

Tier 4 validation uses the OFTLISRV implementation schema from the Implementation Companion [3] to deploy a subset of indicators into a production SOC environment. The deployment follows the phased approach already defined: baseline establishment (30 days), pilot with 10 indicators (60 days), graduated rollout.

The critical addition for validation purposes is a **prediction-first protocol**: before each operational period, the CPF engine generates explicit predictions about which indicators are in Yellow or Red state. These predictions are sealed (logged with cryptographic timestamp) before any incident data for that period is available. At the end of each period, actual incidents are compared against the sealed predictions.

7.3 Validation Metrics

Four metrics constitute operational validation:

Precision. Of all periods where CPF predicted elevated risk, what fraction actually experienced an incident? Target: > 0.4. This threshold may appear low—it means that 60% of

elevated-risk predictions are not followed by a detected incident. However, two factors justify it. First, CPF measures psychological vulnerability states, not imminent attacks: elevated vulnerability does not guarantee that an attacker will exploit it during the observation window. Second, the comparison baseline in Section 7.5 contextualizes this number—if UEBA systems in the same SOC achieve precision in the 0.2–0.3 range on psychological attack vectors, a CPF precision of 0.4 represents a meaningful improvement. The threshold should be re-evaluated after the first 6 months of operational data.

Recall. Of all incidents that occurred, what fraction were preceded by an elevated CPF risk prediction? Target: > 0.6 (60% of incidents were predicted).

Lead time. How far in advance did the elevated-risk state appear before the incident? Target: ≥ 24 hours for the majority of predicted incidents.

Convergence Index correlation. The Convergence Index (CI) defined in the Implementation Companion should correlate with incident severity. Pearson correlation $r > 0.5$ between CI and incident impact score constitutes strong operational validation.

7.4 Duration and Scale Requirements

Meaningful operational validation requires a minimum of 12 months of continuous monitoring to capture seasonal variation, incident cycles, and low-frequency events. The organization must have a baseline incident rate sufficient to produce at least 20 security incidents during the validation period. Organizations with lower incident rates can extend the validation period or pool data across multiple participating organizations using federated analysis protocols that preserve organizational privacy.

7.5 Comparison Baseline

Operational validation is only meaningful if compared against existing prediction methods. The CPF engine's predictions must be compared against: (1) traditional indicator-of-compromise (IOC) based detection, (2) user and entity behavior analytics (UEBA) systems already deployed in the SOC, and (3) a null model (random prediction at the base rate of incidents). If CPF does not outperform UEBA on at least recall and lead time, the operational value proposition is not established.

8 Consolidated Execution Timeline

The four tiers can be partially parallelized. Tier 2 corpus acquisition can begin immediately using existing incident databases—the collection and initial screening of candidate incidents does not depend on Tier 1. However, the final corpus selection (applying the CPF-informed inclusion criteria and mapping incidents to specific indicators) should wait for Tier 1 results to ensure that only indicators with demonstrated pattern robustness are included in the retrospective analysis. In practice, this means Tier 2 runs in two phases: corpus acquisition and preliminary screening (Months 1–3, parallel with Tier 1), followed by CPF-informed analysis (Months 3–5, after Tier 1 concludes). Tier 3 requires organizational partnership and IRB approval, which typically requires 3–6 months of preparation.

Total timeline from initiation to full operational validation: approximately 20 months. However, actionable evidence begins accumulating at Month 3 (Tier 1 results) and operational deployment can begin at Month 8 based on Tier 1–3 convergence, with Tier 4 running concurrently as ongoing validation.

Table 5: Recommended Execution Timeline

Tier	Start	Duration	Dependencies
Tier 1: LLM Proxy	Month 0	3 months	None
Tier 2: Phase A (acquisition)	Month 1	2 months	Corpus availability
Tier 2: Phase B (analysis)	Month 3	2 months	Tier 1 results
Tier 3: Human Experiments	Month 4	6 months	IRB approval, org. partnership
Tier 4: Operational	Month 8	12 months	Tier 3 results, SOC partnership

9 Convergence Criteria for Full Validation

Full validation of the CPF does not require perfection at every tier. It requires convergence: multiple independent methods agreeing on the same conclusion. We define full validation as satisfied when the following conditions hold simultaneously.

Tier 1 produces Red results for at least 60% of indicators across at least 2 of 3 tested models. Tier 2 produces Cohen's kappa > 0.6 across the incident corpus with temporal precedence in $> 70\%$ of cases. Tier 3 produces statistically significant vulnerability activation ($p < 0.05$) for at least 7 of the 10 testable indicators, with Hawthorne correction applied. Tier 4 produces precision > 0.4, recall > 0.6, and lead time ≥ 24 hours, with performance exceeding UEBA baseline on at least two of these metrics.

If any single tier fails to meet its criteria while the others succeed, this indicates a methodological issue at that tier rather than a framework failure—and the specific confound should be investigated before concluding that the indicator is invalid.

10 Conclusion

The CPF addresses a real and underserved problem: the pre-cognitive psychological processes that drive the majority of security incidents. The framework's theoretical foundations are sound, and the SiliconPsyche protocol has demonstrated that its core patterns are robust enough to transfer to synthetic cognitive systems. What remains is the empirical bridge between theoretical prediction and operational reality.

This roadmap provides that bridge. By stratifying indicators according to their epistemic requirements and assigning each tier a methodology matched to what it can credibly measure, the validation plan avoids the trap of demanding a single experiment that proves everything. Instead, it builds convergent evidence across four independent methods, each contributing a distinct type of confidence.

The plan is executable with existing resources, existing incident data, and existing organizational security programs. It does not require CPF to be fully validated before it begins generating value—Tier 1 results alone are sufficient to inform red team strategy, and Tier 2 results are sufficient to inform risk assessment. Full validation is the destination, but the journey itself produces usable intelligence at every stage.

References

[1] Canale, G. (2025). The Cybersecurity Psychology Framework: A Pre-Cognitive Vulnerability Assessment Model. *CPF Technical Report Series*.

[2] Canale, G. and Thimmaraju, K. (2026). The Silicon Psyche: Anthropomorphic Vulnerabilities in Large Language Models. *CPF Technical Report Series*, Version 1 (Revision 11).

[3] Canale, G. (2025). Operationalizing the Cybersecurity Psychology Framework: A Systematic Implementation Methodology. *CPF Technical Report Series*.

[4] Landsberger, H. A. (1958). *Hawthorne Revisited: Management and the Worker, Its Critics, and Developments in Human Relations in Industry*. Ithaca: Cornell University Press. (The term "Hawthorne effect" refers to experiments conducted at the Western Electric Hawthorne Works, Cicero, Illinois, 1924–1932.)

[5] Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

[6] Klein, M. (1946). Notes on some schizoid mechanisms. *International Journal of Psychoanalysis*, 27, 99-110.

[7] 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.

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

[9] Schaefer, A., Nils, F., Sanchez, X., & Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 24(7), 1153-1172.