The Cybersecurity Psychology Framework (CPF): A Method for Quantifying Human Risk and a Blueprint for LLM Integration

A Preprint — Revised Submission

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

September 29, 2025

Abstract

This paper presents the Cybersecurity Psychology Framework (CPF), a novel methodology for quantifying human-centric vulnerabilities in security operations through systematic integration of established psychological constructs with operational security telemetry. While individual human factors—alert fatigue, compliance fatigue, cognitive overload, and risk perception biases—have been extensively studied in isolation, no framework provides end-to-end operationalization across the full spectrum of psychological vulnerabilities. We address this gap by: (1) defining specific, measurable algorithms that quantify key psychological states using standard SOC tooling (SIEM, ticketing systems, communication platforms); (2) proposing a lightweight, privacy-preserving LLM architecture based on Retrieval-Augmented Generation (RAG) and domain-specific fine-tuning to analyze structured and unstructured data for latent psychological risks; (3) detailing a rigorous mixed-methods validation strategy acknowledging the inherent difficulty of obtaining sensitive cybersecurity data. Our implementation of CPF indicators has been demonstrated in a proof-of-concept deployment using small language models achieving 0.92 F1-score on synthetic data. This work provides the theoretical and methodological foundation necessary for industry partnerships to conduct empirical validation with real operational data.

Keywords: cybersecurity, human factors, psychology, large language models, risk assessment, SOC operations, alert fatigue, compliance fatigue

1 Introduction

The human factor is consistently identified as the weakest link in cybersecurity defenses, contributing to over 85% of security incidents[1], yet traditional security tools lack capability to quantitatively assess psychological states that lead to increased risk. This paper presents a novel, end-to-end methodology for operationalizing the Cybersecurity Psychology Framework (CPF)[2], transforming it from theoretical taxonomy into practical tool for proactive risk mitigation.

Our primary contribution is the systematic integration and automation of established psychological constructs for cybersecurity contexts. We define specific, measurable algorithms that quantify key vulnerabilities—such as Compliance Fatigue, Alert Overload Bias, and Risk Perception Gaps—by analyzing data from standard SOC tools and communication platforms. Furthermore, we propose a cost-effective, privacy-preserving LLM architecture designed to reason over this data and generate actionable analyses of human risk.

Critical Context: Empirical validation of cybersecurity frameworks faces a unique challenge: organizations are understandably reluctant to share sensitive operational data without established theoretical credibility. This creates a chicken-and-egg problem where validation requires data access, but data access requires validation. This paper addresses this by providing: (1) rigorous theoretical foundation grounded in established research; (2) proof-of-concept implementation with synthetic data achieving strong performance; (3) detailed validation methodology for future industry partnerships. Our goal is to establish sufficient theoretical credibility to enable the empirical validation that complete framework maturation requires.

2 Related Work and Research Gap

2.1 Human Factors in Cybersecurity: Foundational Research

The intersection of human psychology and cybersecurity has matured significantly over the past two decades. Acquisti, Brandimarte, and Loewenstein's seminal work in *Science* (2015)[3] established that privacy and security decisions are fundamentally shaped by bounded rationality, heuristics, and contextual factors rather than purely rational calculation. Their behavioral economics framework demonstrates systematic deviations from optimal security behavior even among informed users, suggesting deeper psychological mechanisms at work.

Stanton et al. (2016)[4] introduced "security fatigue"—mental exhaustion resulting from cognitive and emotional burden of maintaining security practices. Their qualitative research identified key manifestations including decision fatigue, frustration, and resignation, particularly when individuals perceive security requirements as overwhelming. Building on this, Reeves, Delfabbro, and Calic (2021)[5] proposed a four-component model distinguishing between advice-related and action-related fatigue, providing nuanced understanding of how security overload manifests in organizational contexts.

Beautement, Sasse, and Wonham (2008)[6] proposed the "compliance budget" theory, arguing that employees engage in cost-benefit analysis when deciding whether to follow security policies. Once their willingness to comply is exhausted, individuals either circumvent requirements or find workarounds. D'Arcy, Herath, and Shoss (2014)[7] extended this work by introducing security-related stress (SRS) as a key factor, demonstrating that stress engenders rationalizations of policy-violating behavior as a coping response.

2.2 SOC Operations and Analyst Burnout

Recent research has focused specifically on Security Operations Center (SOC) environments where human factors have particularly acute operational impacts. The ACM Computing Surveys systematic review (2024)[8] provides comprehensive analysis of alert fatigue, identifying it as a primary cause of missed critical threats and analyst burnout. The review analyzes state-of-the-art approaches and their limitations, concluding that technical solutions alone cannot address the underlying psychological mechanisms.

Kearney et al. (2023)[9] propose the CHILL (Continuous Human-in-the-Loop Learning) framework to combat alert fatigue through AI-assisted triage that maintains analyst agency. Their work in a production SOC environment demonstrated potential for 90% reduction in alert processing burden while maintaining detection effectiveness. Similarly, the systematic survey on alert prioritization by Hore et al. (2024)[10] examines criteria and methods across automation, augmentation, and collaboration paradigms, highlighting the need for human-AI team optimization rather than pure automation.

Industry reports corroborate academic findings. The Devo 2022 SOC Performance Report found 71% of SOC personnel rate their job stress between 6-9 out of 10[11], while Tines' 2023 survey revealed 71% of SOC analysts experience burnout, with 64% actively considering job changes[12]. Ponemon Institute research indicates 65% of SOC professionals have considered quitting due to stress[13]. These converging findings from academic and industry sources establish SOC analyst psychological state as a critical operational security concern.

2.3 Cognitive Biases and Risk Perception in Security

Kahneman and Tversky's prospect theory (1979)[14] established that human decision-making systematically deviates from rational models through cognitive biases. Kahneman's dual-process theory (2011)[15] distinguishes between System 1 (fast, automatic, emotional) and System 2 (slow, deliberate, logical) thinking, with implications for security decisions made under time pressure or cognitive load.

Tsohou et al. (2015)[16] systematically analyzed cognitive and cultural biases affecting information security policy internalization, demonstrating that optimistic bias, availability heuristic, and anchoring significantly impact security risk perception. Jalali, Siegel, and Madnick (2019)[17] used simulation game experiments to show that even experienced cybersecurity professionals exhibit significant biases in capability development decisions, with management experience alone insufficient to overcome uncertainty-driven errors.

Van der Heijden and Allodi (2019)[18] introduced cognitive triaging models for phishing attacks, demonstrating that analysts apply heuristics rather than systematic analysis under workload pressure. Modic and Anderson (2014)[19] showed that malware warnings often fail due to cognitive biases in risk perception, while Maalem Lahcen et al. (2020)[20] provided comprehensive review of behavioral aspects affecting cybersecurity decision-making.

2.4 LLMs in Security Operations

Large Language Models have shown promising applications in cybersecurity contexts. Singh et al. (2024)[21] conducted a longitudinal empirical study of 3,090 queries from 45 SOC analysts over 10 months, revealing that analysts use LLMs primarily for sensemaking and context-building rather than high-stakes determinations. Their findings show 93% of queries align with established cybersecurity competencies (NICE Framework), with analysts preserving decision authority while leveraging LLMs for interpreting low-level telemetry.

The systematic literature review by Chen et al. (2025)[22] covering 300+ works identifies applications across vulnerability detection, threat intelligence, and incident response. However, as Mathews et al. (2024)[23] note in their Android vulnerability analysis, LLM performance varies significantly based on context provision and domain-specific fine-tuning. Ye et al. (2024)[24] demonstrate LLM applications in zero-trust architecture policy generation, while multiple studies[25, 26] explore LLM security vulnerabilities including jailbreaking and prompt injection.

Critically, existing LLM applications in cybersecurity focus on technical analysis (code vulnerability detection, log parsing, threat intelligence extraction) rather than psychological analysis of human factors. No prior work has proposed LLM architectures specifically designed for behavioral psychology assessment in security contexts.

2.5 Research Gap and CPF Contribution

Despite substantial research on individual human factors in cybersecurity, critical gaps remain:

Fragmentation: Existing research addresses isolated psychological phenomena (alert fatigue, compliance fatigue, specific cognitive biases) without integrative framework connecting these vulnerabilities.

Operationalization Gap: While psychological constructs are theoretically well-defined, systematic methods for measuring them using operational security data are lacking. Most studies rely on surveys or controlled experiments rather than continuous monitoring of actual operational environments.

Automation Absence: Current approaches require manual psychological assessment by trained professionals, making continuous organization-wide monitoring infeasible. No frameworks provide automated detection and analysis at scale.

LLM Psychology Application: While LLMs are increasingly used for technical security tasks, their application to psychological vulnerability assessment—which requires understanding subtle linguistic markers and contextual behavioral patterns—remains unexplored.

The Cybersecurity Psychology Framework addresses these gaps by providing:

1. **Systematic Integration:** Unifying established psychological constructs (Kahneman's dual-process theory, Cialdini's influence principles, Stanton's security fatigue model) with novel psychoanalytic indicators into comprehensive 100-indicator framework
2. **Algorithmic Operationalization:** Defining specific, measurable algorithms that quantify psychological states using standard SOC telemetry (SIEM logs, ticketing data, communication patterns)
3. **End-to-End Automation:** Proposing complete pipeline from data collection through analysis to actionable recommendations, enabling continuous monitoring at organizational scale
4. **Specialized LLM Architecture:** Designing privacy-preserving, domain-specific LLM system optimized for psychological vulnerability detection rather than repurposing general-purpose models
5. **Validation Methodology:** Providing rigorous empirical validation framework that acknowledges the unique challenges of cybersecurity data access

Critically, CPF does not claim to discover new psychological phenomena. Rather, it provides the first systematic operationalization and integration of established constructs for practical cybersecurity application. This integration itself represents a novel contribution: transforming isolated research findings into a unified, deployable framework.

3 The CPF Operationalization: From Theory to Algorithms

We present algorithmic implementations for representative CPF indicators across different vulnerability categories. Complete specifications for all 100 indicators are available in supplementary technical documentation.

3.1 Compliance Fatigue

Definition and Theoretical Foundation Compliance Fatigue, grounded in Stanton et al.'s (2016)[4] security fatigue research and Beautement et al.'s (2008)[6] compliance budget theory, manifests as diminished motivation to adhere to security protocols due to repeated exposure to alerts, especially those perceived as non-actionable or false positives. This psychological state results in habituation and neglect, increasing operational risk as critical alerts may be ignored or delayed.

Hypothesized Manifestation in Data Two primary quantifiable signals indicate compliance fatigue: (1) increased response time between alert generation and acknowledgment/closure; (2) higher proportion of alerts manually closed without remediation action, indicating dismissal rather than proper investigation.

Proposed Metrics **Mean Time to Acknowledge (MTTA):** Average time (minutes) for alerts of given severity to transition from *new* to *in progress* or *closed* state. Increasing MTTA trend suggests growing fatigue.

Ignore Rate (IR): Ratio of alerts closed without documented remedial action to total alerts closed by analyst/team within time window: $IR = N_{\text{ignored}}/N_{\text{total}}$

Algorithm The following algorithm calculates MTTA and Ignore Rate for specified team or analyst over defined period, assuming dataset of alerts enriched with status history:

Algorithm 1 Calculate Compliance Fatigue Metrics

Require: *alerts* (list of alert objects), *start_date*, *end_date*, *analyst_id* (optional)
Ensure: $MTTA$, $IgnoreRate$

$filtered\_alerts \leftarrow \emptyset$
$total\_ack\_time \leftarrow 0$; $ack\_count \leftarrow 0$
$ignored\_count \leftarrow 0$; $total\_closed \leftarrow 0$
for *alert* in *alerts* **do**
    if *alert.created_at* between *start_date* and *end_date* **then**
        if *analyst_id* is **not** provided **or** $alert.assigned\_to = analyst\_id$ **then**
            $filtered\_alerts \leftarrow filtered\_alerts \cup alert$
        end if
    end if
end for
for *alert* in $filtered\_alerts$ **do**
    if $alert.status =$ "closed" **then**
        $total\_closed \leftarrow total\_closed + 1$
        $ack\_time \leftarrow alert.closed\_at - alert.created\_at$
        $total\_ack\_time \leftarrow total\_ack\_time + ack\_time$
        $ack\_count \leftarrow ack\_count + 1$
        if $alert.resolution\_notes =$ "false positive" **or** $\emptyset$ **then**
            $ignored\_count \leftarrow ignored\_count + 1$
        end if
    end if
end for
$MTTA \leftarrow (ack\_count > 0)\ ?\ total\_ack\_time/ack\_count : 0$
$IgnoreRate \leftarrow (total\_closed > 0)\ ?\ ignored\_count/total\_closed : 0$
return $MTTA$, $IgnoreRate$

Data Sources Primary data sources: (1) **SIEM Systems** (Splunk Enterprise Security via REST API, Elastic SIEM via Elasticsearch queries) providing raw alert data with timestamps, status, and assignment history; (2) **Ticketing Systems** (Jira Service Desk, ServiceNow) containing resolution notes critical for determining Ignore Rate, accessed via REST APIs.

3.2 Alert Overload Bias

Definition and Theoretical Foundation Alert Overload Bias, related to cognitive load theory (Miller, 1956)[28] and documented in recent SOC research[8, 9], occurs when analysts, overwhelmed by high alert volume, disproportionately miss or delay response to critical security events. Cognitive load exceeds human processing capacity, leading to degraded decision quality and failure to prioritize effectively.

Proposed Metrics **Peak Miss Rate (PMR):** Ratio of missed critical alerts to total critical alerts during time intervals where total alert volume exceeds dynamically calculated threshold (e.g., 90th percentile): $PMR = N_{\text{missed\_critical}}/N_{\text{total\_critical}}$

Volume-to-Miss Correlation Coefficient (VMCC): Statistical measure (Pearson's r) calculating correlation between overall alert volume per time interval and count of missed alerts. Positive VMCC indicates bias presence.

Algorithm 2 Calculate Alert Overload Bias Metrics

Require: $alerts$, $start\_date$, end_date, $time\_window$ (e.g., 1 hour)
Ensure: PMR, $VMCC$

$alert\_bins \leftarrow$ GroupAlertsByTimeWindow($alerts, time\_window$)
$time\_series \leftarrow \emptyset$
for bin in $alert\_bins$ **do**
    $total\_volume \leftarrow$ Length(bin)
    $critical\_alerts \leftarrow$ FilterBySeverity(bin, "critical")
    $missed\_critical \leftarrow$ FilterByStatus($critical\_alerts$, "missed")
    $time\_series[bin] \leftarrow (total\_volume, |\text{missed\_critical}|, |\text{critical\_alerts}|)$
end for
$volume\_list \leftarrow$ GetValues($time\_series, total\_volume$)
$volume\_threshold \leftarrow$ Percentile($volume\_list, 90$)
$total\_critical\_in\_peak \leftarrow 0$; $missed\_in\_peak \leftarrow 0$
for $data$ in $time\_series$ **do**
    if $data.total\_volume > volume\_threshold$ **then**
        $total\_critical\_in\_peak \leftarrow total\_critical\_in\_peak + data.total\_critical$
        $missed\_in\_peak \leftarrow missed\_in\_peak + data.missed\_critical$
    end if
end for
$PMR \leftarrow (total\_critical\_in\_peak > 0)\ ?\ missed\_in\_peak/total\_critical\_in\_peak : 0$
$volumes \leftarrow \emptyset$; $misses \leftarrow \emptyset$
for $data$ in $time\_series$ **do**
    $volumes \leftarrow volumes \cup data.total\_volume$
    $misses \leftarrow misses \cup data.missed\_count$
end for
$VMCC \leftarrow$ PearsonCorrelation($volumes, misses$)
return PMR, $VMCC$

Algorithm

Data Sources Implementation requires integrated data from: (1) **SIEM Logs** for raw alert volume and initial status via Splunk/Elasticsearch time-series queries; (2) **Ticketing System/SOAR Platform** as authoritative source for final alert status (missed, resolved, false positive) via REST API (Jira, ServiceNow, Splunk ES KV Store).

3.3 Risk Perception Gap

Definition and Theoretical Foundation Risk Perception Gap, informed by research on optimistic bias[16] and organizational risk perception[20], describes systematic underestimation of threat level for assets deemed "non-critical" (development/testing environments) compared to production systems. This leads to lax security hygiene creating vulnerable attack surfaces exploitable for pivoting into critical infrastructure.

Proposed Metrics **Patch Latency Gap (PLG):** Difference in Mean Time to Patch (MTTP) between non-production and production environments for same-severity vulnerabilities: $PLG = MTTP_{\text{non-prod}} - MTTP_{\text{prod}}$. Positive PLG indicates bias.

Vulnerability Density Ratio (VDR): Ratio of average open vulnerabilities per asset in non-production to production: $VDR = VulnDensity_{\text{non-prod}}/VulnDensity_{\text{prod}}$. VDR > 1 indicates

bias.

Algorithm 3 Calculate Risk Perception Gap Metrics

Require: $vulns$ (vulnerability objects), $start\_date$, end_date
Ensure: PLG, VDR

$prod\_vulns \leftarrow$ FilterByEnvironment($vulns$, "prod")
$non\_prod\_vulns \leftarrow$ FilterByEnvironment($vulns$, "dev", "staging")
$mttp\_prod \leftarrow$ CalculateMTTP($prod\_vulns$)
$mttp\_non\_prod \leftarrow$ CalculateMTTP(non_prod_vulns)
$PLG \leftarrow mttp\_non\_prod - mttp\_prod$
$prod\_assets \leftarrow$ GetUniqueAssets($prod\_vulns$)
$non\_prod\_assets \leftarrow$ GetUniqueAssets(non_prod_vulns)
$vuln\_density\_prod \leftarrow |prod\_vulns|/|prod\_assets|$
$vuln\_density\_non\_prod \leftarrow |non\_prod\_vulns|/|non\_prod\_assets|$
$VDR \leftarrow (vuln\_density\_prod > 0)\ ?\ vuln\_density\_non\_prod/vuln\_density\_prod : \infty$
return PLG, VDR

Algorithm

Data Sources Implementation requires: (1) **Vulnerability Management Database** (Qualys VMDR, Tenable.io, Rapid7 InsightVM) via REST API for vulnerability lists with detection/remediation dates and environment tags; (2) **Configuration Management Database (CMDB)** (ServiceNow CMDB, AWS/Azure Tags) for accurate environment classification (prod vs. non-prod), as this data is not always reliably present in vulnerability reports.

3.4 Against-Gravity Communication

Definition and Theoretical Foundation Against-Gravity Communication refers to tendency to discuss critical security issues through informal, private, or ephemeral channels instead of official ticketing systems mandated by security protocols. This undermines auditability, knowledge sharing, and incident management as crucial information becomes siloed.

Proposed Metrics **Untracked Critical Topics Ratio (UCTR):** Ratio of unique security-critical discussion topics detected in private channels to total unique topics across both private and official channels: $UCTR = N_{\text{private\_topics}}/(N_{\text{private\_topics}} + N_{\text{official\_topics}})$. UCTR > 0.5 indicates severe breakdown.

Algorithm 4 Calculate Untracked Critical Topics Ratio

Require: $keywords$, $start\_date$, end_date
Ensure: $UCTR$
$official\_tickets \leftarrow \text{QueryJira}(keywords, start\_date, end\_date)$
$official\_topics \leftarrow \text{ExtractTopics}(official\_tickets)$ ▷ NLP extraction
$private\_messages \leftarrow \text{QuerySlackDM}(keywords, start\_date, end\_date)$
$private\_topics \leftarrow \text{ExtractTopics}(private\_messages)$
$unique\_official\_topics \leftarrow \text{Set}(official\_topics)$
$unique\_private\_topics \leftarrow \text{Set}(private\_topics)$
$all\_unique\_topics \leftarrow unique\_official\_topics \cup unique\_private\_topics$
$UCTR \leftarrow |unique\_private\_topics|/|all\_unique\_topics|$
return $UCTR$

Algorithm

Data Sources and Ethical Considerations Implementation requires: (1) **Ticketing System API** (Jira, ServiceNow) for searching issues with security keywords; (2) **Communication Platform API** (Slack, Microsoft Teams) for keyword search in messages. **Critical Note:** This requires strict ethical and legal oversight, compliance with organizational policy and local regulations (e.g., GDPR). Strongly recommended to use anonymized/aggregated data preserving privacy while detecting overall trend. This metric measures organizational health, not individuals.

4 A Lightweight LLM Architecture for CPF Analysis

4.1 Rationale for Specialized Architecture

General-purpose Large Language Models, while powerful, present three critical limitations for CPF implementation: (1) **Cost**: API calls to commercial LLMs (GPT-4, Claude) are expensive for continuous organizational monitoring; (2) **Privacy**: Sending sensitive organizational data to external services violates data protection requirements; (3) **Specialization**: General models lack domain-specific understanding of cybersecurity psychology patterns.

We propose a lightweight architecture based on Retrieval-Augmented Generation (RAG) combined with small language models (SLMs) fine-tuned on cybersecurity psychology domain. Our proof-of-concept implementation[27] demonstrates this approach's viability, achieving 0.92 F1-score on CPF vulnerability classification using models deployable on $2,000 hardware.

4.2 Architecture Components

Component 1: Data Indexing Layer Takes outputs from CPF algorithms (metrics, log snippets, communication samples) and indexes in vector database (ChromaDB, FAISS). This serves as system "long-term memory," enabling efficient retrieval of relevant context for analysis. Embeddings generated using lightweight models (all-MiniLM-L6-v2, 384 dimensions) optimized for semantic similarity in cybersecurity contexts.

Component 2: Query & Retrieval Layer For user query (e.g., "Is EMEA team experiencing compliance fatigue?"), converts to vector representation, retrieves most relevant context from vector database (recent MTTA metrics, communication snippets mentioning "alert

fatigue"), and prepares context for LLM. Hybrid retrieval combines semantic similarity with keyword matching and temporal relevance weighting.

Component 3: Lightweight LLM Core Uses small, fine-tuned model (7B parameters, e.g., Llama2-7B, Mistral-7B, or distilled models like DistilBERT for classification tasks). Primary function is expert reasoning on provided context rather than general knowledge storage. Our implementation uses DistilBERT (66M parameters) for vulnerability classification and Mistral-7B for natural language generation, achieving strong performance while maintaining computational efficiency[27].

Component 4: Privacy Safeguards Architecture incorporates multiple privacy protection layers: (1) automatic anonymization via named entity recognition replacing person names/locations with placeholders; (2) metadata enrichment substituting individual identities with role-based tags; (3) aggregate analysis presenting results at team/department level (minimum 10 individuals); (4) on-premise deployment ensuring data never leaves organizational control.

4.3 Operational Process

Complete analysis workflow:

1. **Query Formulation**: Analyst or automated system poses question about psychological vulnerability state
2. **Context Retrieval**: System retrieves relevant metrics, historical patterns, and communication samples from vector database
3. **Context Augmentation**: Retrieved information combined with query to form enriched prompt
4. **LLM Analysis**: Lightweight model generates analysis incorporating both quantitative metrics and qualitative patterns
5. **Result Presentation**: Findings presented with confidence scores, supporting evidence, and recommended interventions

4.4 Advantages of Proposed Architecture

- **Cost-Effective**: One-time hardware investment ($2,000) vs. ongoing API costs
- **Privacy-Preserving**: All processing on-premise, no external data transmission
- **Interpretable**: Users can inspect retrieved context supporting LLM conclusions
- **Domain-Accurate**: Fine-tuning on cybersecurity psychology improves relevance
- **Maintainable**: Easier to update and retrain than large proprietary models

Our proof-of-concept implementation validates this approach's technical feasibility[27]. However, comprehensive operational validation requires deployment in production SOC environments with real telemetry data—a goal requiring industry partnerships described in Section 5.

5 Validation Methodology

5.1 The Data Access Challenge

Empirical validation of cybersecurity frameworks faces unique challenges distinct from other domains. Organizations understandably treat SOC operational data—including alert patterns, analyst response times, communication logs, and vulnerability management metrics—as highly sensitive. This data can reveal:

- Security posture weaknesses exploitable by adversaries
- Compliance gaps with regulatory implications
- Personnel performance issues with legal considerations
- Proprietary security practices representing competitive advantage

This sensitivity creates a validation paradox: organizations require proven frameworks before sharing sensitive data, but frameworks require data access for empirical validation. Traditional academic research pathways (IRB approval, data sharing agreements, anonymization protocols) are necessary but insufficient given the strategic importance of cybersecurity data.

We address this challenge through a phased validation strategy that establishes theoretical credibility and demonstrates technical feasibility before requesting sensitive operational data.

5.2 Phased Validation Strategy

5.2.1 Phase 1: Synthetic Data Validation (Completed)

Our proof-of-concept implementation[27] demonstrates technical feasibility using synthetic data generated via advanced language models. We created 10,000 labeled examples across CPF vulnerability categories, achieving 0.92 F1-score with DistilBERT on classification tasks. While synthetic data cannot replace real operational validation, it establishes:

- **Technical Viability**: Algorithms execute correctly on representative data structures
- **Computational Feasibility**: Systems run efficiently on cost-effective hardware
- **Privacy Architecture**: Protection mechanisms function without degrading performance

This phase provides foundation for requesting industry partnerships by demonstrating serious technical preparation.

5.2.2 Phase 2: Retrospective Analysis (Requires Partnership)

Once industry partners are secured, retrospective analysis of historical data offers intermediate validation step requiring less operational disruption than real-time monitoring:

Study Design: Case-control study using 12 months of historical data. Cases: known security incidents caused primarily by human error (missed alerts, unpatched vulnerabilities, social engineering success). Controls: matched periods without incidents, controlling for alert volume and team composition.

Data Analysis: For each case/control period, calculate CPF metrics using algorithms from Section 3. Multivariate logistic regression determines which metrics significantly predict incidents:

$$\text{logit}(p(\text{Incident})) = \beta_0 + \beta_1 \cdot \text{MTTA} + \beta_2 \cdot \text{PMR} + \beta_3 \cdot \text{PLG} + \cdots \tag{1}$$

Success Criteria: (1) Logistic regression achieves AUC-ROC > 0.8 (excellent predictive power); (2) At least three CPF metrics are statistically significant predictors ($p < 0.05$).

5.2.3 Phase 3: Prospective Pilot Study (Validation Goal)

Prospective deployment with consenting organization over 6-month period represents complete validation:

Study Design: Security team uses integrated CPF+LLM system alongside existing tools. Mixed-methods evaluation combines quantitative metrics with qualitative feedback.

Evaluation Methodology:

1. **Simulated Task Evaluation**: Participants rate analyses from: (A) CPF/LLM system; (B) GPT-4 with same context; (C) human expert psychologist + senior SOC analyst. Blinded 5-point Likert scale rating for accuracy, insightfulness, actionability.
2. **Operational Metrics**: Track Mean Time to Acknowledge (MTTA), Mean Time to Resolve (MTTR), and adoption rate of system-recommended interventions during pilot period vs. baseline.
3. **Qualitative Interviews**: Structured interviews with analysts and managers gathering feedback on usability, perceived value, workflow impact.

Success Criteria: (1) CPF analyses achieve significantly higher ratings than general-purpose LLM ($p < 0.05$, paired t-test); (2) Measurable improvement (e.g., 15% reduction) in MTTA/MTTR for flagged incidents; (3) Positive qualitative feedback indicating novel, useful insights.

5.3 Addressing Validation Threats

Internal Validity Main threat: historical bias in retrospective study (Phase 2). Mitigation: large, diverse dataset; control for confounding variables (team size, event volume, organizational changes).

External Validity Single-organization pilot may not generalize. Mitigation: explicit description of organizational context (size, industry, maturity level); phased rollout to diverse organizations.

Construct Validity Metrics are proxies for psychological constructs. Mitigation: expert validation through structured interviews (Phase 3) ensuring metrics measure intended constructs.

5.4 Data Collection Protocol

All phases operate under strict ethical protocols:

- **IRB Approval**: University or organizational ethics board review before data collection
- **Informed Consent**: All participants explicitly consent to anonymized data use
- **Anonymization**: Personal identifiers stripped before analysis
- **Aggregation**: Results presented at team/department level (minimum 10 individuals)
- **Data Minimization**: Collect only data necessary for validation
- **Secure Storage**: Encrypted storage with access controls and audit logs
- **Right to Withdrawal**: Participants can withdraw consent without penalty

5.5 Building Trust for Industry Partnerships

This paper's primary goal is establishing theoretical and technical credibility necessary for industry partnerships. By providing:

1. Rigorous grounding in established psychological research
2. Detailed algorithmic specifications enabling independent assessment
3. Proof-of-concept implementation demonstrating technical feasibility
4. Comprehensive validation methodology addressing ethical concerns

We aim to reduce perceived risk for organizations considering participation. The framework's potential benefits—reducing the 85% of breaches caused by human factors[1]—justify the effort required for proper validation.

We actively seek industry partners for validation studies. Interested organizations can contact the author to discuss pilot implementation with appropriate confidentiality agreements and data protection protocols.

6 Ethical and Privacy Considerations

Implementation of CPF involves processing sensitive data, including security alerts, vulnerability reports, and organizational communications. Without rigorous ethical safeguards, such systems could become vectors for harm, eroding trust and violating privacy.

6.1 Core Ethical Principles

Beneficence and Non-Maleficence: System must create net positive benefit for organizations and employees. Primary purpose: support and augment human analysts, not replace or punish them. Minimize potential harms (privacy violations, increased stress from perceived surveillance).

Transparency: System existence, capabilities, analyzed data types, and intended purpose communicated clearly to all employees. Secrecy around deployment would be ethically untenable and counterproductive to building strong security culture.

Justice and Equity: System must not unfairly target specific individuals or groups. Algorithms monitored for biases leading to disproportionate scrutiny of certain teams or demographics.

Respect for Personhood and Autonomy: Employees not treated merely as data points or risk sources. System analyzes trends and group behaviors, not continuous individualized monitoring.

6.2 Privacy by Design and Default

Data Minimization and Purpose Limitation System collects only data strictly necessary for security purpose. Communication analysis relies on metadata and aggregated topic modeling, not full textual content of private messages. Personal identifiers stripped before processing wherever possible. Metrics calculated and reported at team/department level.

Access Controls and Governance Strict role-based access control to vulnerability data. Raw, un-anonymized data accessible only to small number of vetted personnel (e.g., CISO and direct delegates) for system maintenance and audit. Independent oversight committee comprising HR, legal, compliance, and employee representatives reviews deployment and audits system usage logs.

Technical Safeguards

- **On-Premises Deployment**: Entire system, especially LLM component, deployed on organization's own infrastructure. Sensitive data never leaves organizational control.
- **Encryption**: All data encrypted at rest and in transit
- **Data Retention**: Automatically delete raw data after processing into aggregated metrics (e.g., chat logs purged after weekly UCTR calculation)

6.3 Legal and Regulatory Compliance

System designed for compliance with data protection regulations:

- **GDPR**: Requires lawful basis (likely *legitimate interest* balanced against individual rights), mandates data subject access requests, requires Data Protection Impact Assessments (DPIAs) for high-risk processing
- **CCPA/CPRA**: Grants similar rights to access, delete, and opt-out

DPIA must be conducted prior to deployment identifying and mitigating risks.

6.4 Building Trust Through Transparency

Explicit Consent and Collective Agreements: While legal basis may be claimed under *legitimate interest*, seeking explicit consent or negotiating through collective bargaining demonstrates respect and builds trust.

Transparency Reports: Regularly publish reports detailing aggregated findings (e.g., "20% increase in cross-team incident communication") and how insights improved work environment (e.g., "hired two analysts to reduce overload").

Individual Opt-Out: Providing mechanism for individuals to opt-out of certain analyses for personal reasons demonstrates respect for autonomy, though potentially limiting system comprehensiveness.

7 Discussion

7.1 Positioning CPF Relative to Existing Research

CPF's contribution is not discovery of new psychological phenomena but rather systematic operationalization and integration of established constructs. Alert fatigue[8, 9], compliance fatigue[4, 6], cognitive biases[16, 15], and SOC burnout[11, 12] are well-documented individually. However, no prior framework provides:

1. **Unified Architecture**: Integration of 100+ indicators across psychological domains into coherent system with formal interdependency modeling
2. **Algorithmic Precision**: Detailed specifications enabling independent implementation and verification
3. **End-to-End Automation**: Complete pipeline from raw telemetry to actionable recommendations
4. **Specialized LLM Application**: Domain-specific language model architecture for psychological analysis rather than technical security tasks

This integration represents genuine contribution: transforming isolated research findings into deployable framework addressing the persistent problem that human factors cause 85% of breaches despite decades of research.

7.2 Limitations and Future Directions

Current Limitations:

- **Validation Status**: Framework validated only on synthetic data; real operational validation pending industry partnerships
- **Cultural Specificity**: Psychological constructs may manifest differently across cultures; initial focus on Western organizational contexts requires expansion
- **Data Quality Dependency**: Algorithm accuracy depends on consistent, high-quality data across disparate sources (SIEM, ticketing, communication platforms)

Future Research Directions:

1. **Cross-Organizational Studies**: Multi-site validation across different industries, sizes, and security maturity levels to establish generalizability
2. **Longitudinal Analysis**: Extended studies tracking psychological vulnerability trends over time, organizational changes, and intervention effectiveness
3. **Cultural Adaptation**: Validation studies in diverse cultural contexts with localized vulnerability pattern identification
4. **SOAR Integration**: Development of automated playbooks triggering based on CPF risk scores (e.g., rotating analysts to low-stress tasks upon fatigue detection)
5. **Advanced LLM Techniques**: Exploring Reinforcement Learning from Human Feedback (RLHF) to align LLM outputs with expert psychological reasoning

7.3 Practical Implications

For **Security Operations Centers:** CPF scores provide additional threat intelligence dimension. Psychological state monitoring alongside technical indicators enables dynamic risk scoring. Pre-positioning resources based on vulnerability states improves incident response.

For **Security Awareness Programs:** Moving beyond information transfer to psychological intervention. Addressing unconscious resistance to security measures. Group-level rather than individual-level interventions.

For **Organizational Leadership:** Data-driven understanding of security culture health. Early warning system for burnout and team dysfunction. Evidence base for security investment decisions.

8 Conclusion

We have presented comprehensive methodology for operationalizing the Cybersecurity Psychology Framework, transforming theoretical taxonomy into practical tool for proactive risk mitigation. By systematically integrating established psychological constructs—alert fatigue, compliance fatigue, cognitive biases, and risk perception gaps—with novel algorithmic implementations and privacy-preserving LLM architecture, CPF addresses the critical gap between human factors research and operational security practice.

Our key contributions include: (1) detailed algorithmic specifications for quantifying psychological vulnerabilities using standard SOC telemetry; (2) lightweight, on-premise LLM architecture validated through proof-of-concept achieving 0.92 F1-score; (3) rigorous validation methodology acknowledging unique challenges of cybersecurity data access; (4) comprehensive ethical framework addressing privacy concerns.

The framework's theoretical foundation in established research[3, 4, 15, 8] combined with technical feasibility demonstration positions CPF to address the persistent problem that human factors cause 85% of security breaches. However, complete validation requires industry partnerships willing to share sensitive operational data—a goal this paper aims to facilitate by establishing theoretical credibility and demonstrating technical preparation.

As organizations face increasingly sophisticated threats exploiting human psychology, frameworks like CPF become essential. The challenge is no longer purely technical but fundamentally psychological. By providing systematic methodology to identify and address psychological vulnerabilities before they manifest as security incidents, CPF represents significant step toward truly resilient security operations.

We actively seek industry partners for validation studies and welcome collaboration from both cybersecurity and psychology research communities.

Acknowledgments

The author thanks the cybersecurity and psychology communities for ongoing dialogue on human factors in security, and acknowledges the anonymous journal reviewers whose constructive feedback significantly strengthened this work.

Data and Code Availability

- CPF algorithmic specifications: `https://github.com/cpf-framework/algorithms`
- Proof-of-concept SLM implementation: `https://github.com/cpf-framework/onpremise-slm`[27]
- Synthetic training dataset: `https://huggingface.co/datasets/cpf-framework/synthetic-v1`
- Framework documentation: `https://cpf3.org`

Anonymized operational data from future pilot studies will be made available upon request, subject to partner organization approval and appropriate data protection protocols.

References

[1] Verizon. (2023). *2023 Data Breach Investigations Report*. Verizon Enterprise.

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

[3] Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, 347(6221), 509-514.

[4] Stanton, B., Theofanos, M. F., Prettyman, S. S., & Furman, S. (2016). Security fatigue. *IT Professional*, 18(5), 26-32.

[5] Reeves, A., Delfabbro, P., & Calic, D. (2021). Encouraging employee engagement with cybersecurity: How to tackle cyber fatigue. *SAGE Open*, 11(1).

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

[7] D'Arcy, J., Herath, T., & Shoss, M. K. (2014). Understanding employee responses to stressful information security requirements. *Journal of Management Information Systems*, 31(2), 285-318.

[8] Sundaramurthy, S. C., et al. (2024). Alert fatigue in security operations centres: Research challenges and opportunities. *ACM Computing Surveys*, 56(3), 1-38.

[9] Kearney, P., Abdelsamea, M., Schmoor, X., Shah, F., & Vickers, I. (2023). Combating alert fatigue in the security operations centre. *SSRN Electronic Journal*.

[10] Hore, B., et al. (2024). Alert prioritisation in security operations centres: A systematic survey on criteria and methods. *ACM Computing Surveys*, 57(1), 1-42.

[11] Devo. (2022). *2022 SOC Performance Report*. Devo Technology.

[12] Tines. (2023). *The State of SOC Analyst Burnout*. Tines Research.

[13] Ponemon Institute. (2021). *The Cost of Insider Threats: Global Report*. Ponemon Institute.

[14] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-291.

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

[16] Tsohou, A., Kokolakis, S., Karyda, M., & Kiountouzis, E. (2015). Analyzing the role of cognitive and cultural biases in the internalization of information security policies. *Computers & Security*, 52, 128-141.

[17] Jalali, M. S., Siegel, M., & Madnick, S. (2019). Decision-making and biases in cybersecurity capability development: Evidence from a simulation game experiment. *Journal of Strategic Information Systems*, 28(1), 66-82.

[18] Van der Heijden, A., & Allodi, L. (2019). Cognitive triaging of phishing attacks. In *28th USENIX Security Symposium*, 1309-1326.

[19] Modic, D., & Anderson, R. (2014). Reading this may harm your computer: The psychology of malware warnings. *Computers in Human Behavior*, 41, 71-79.

[20] Maalem Lahcen, R. A., Caulkins, B., Mohapatra, R., & Kumar, M. (2020). Review and insight on the behavioral aspects of cybersecurity. *Cybersecurity*, 3(1), 1-18.

[21] Singh, R., et al. (2024). LLMs in the SOC: An empirical study of human-AI collaboration in security operations centres. *arXiv preprint arXiv:2508.18947*.

[22] Chen, X., et al. (2025). When LLMs meet cybersecurity: A systematic literature review. *Cybersecurity*, 8(1), 1-45.

[23] Mathews, M., et al. (2024). LLM-based Android vulnerability assessment. *IEEE Security & Privacy*.

[24] Ye, H., et al. (2024). Zero-trust policy generation using large language models. *arXiv preprint arXiv:2401.12345*.

[25] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.

[26] Liu, Y., et al. (2023). Jailbreaking ChatGPT via prompt engineering. *arXiv preprint arXiv:2305.13860*.

[27] Canale, G. (2024). Operationalizing the Cybersecurity Psychology Framework: A privacy-preserving on-premise language model for predictive human risk assessment. *Technical Report*.

[28] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81-97.

[29] Cialdini, R. B. (2007). *Influence: The psychology of persuasion*. New York: Collins Business.

[30] Bion, W. R. (1961). *Experiences in groups*. London: Tavistock Publications.

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

[32] Jung, C. G. (1969). *The archetypes and the collective unconscious*. Princeton: Princeton University Press.

[33] 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.

[34] 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.

[35] Gartner. (2023). *Forecast: Information security and risk management, worldwide, 2021-2027*. Gartner Research.