Operacionalizar el Cybersecurity Psychology Framework: Una Metodología Sistemática de Implementación

Companion Técnico de Implementación al CPF v1.0

December 20, 2025

Giuseppe Canale, CISSP

Independent Researcher

kaolay@gmail.com
g.canale@cpf3.org

URL: cpf3.org

ORCID: 0009-0007-3263-6897

Abstract

Este documento proporciona una metodología sistemática para operacionalizar todos los 100 indicadores del Cybersecurity Psychology Framework en capacidades SOC funcionantes. Presentamos el esquema de implementación OFTLISRV, formulaciones matemáticas de detección y modelado de red bayesiana para las interdependencias de los indicadores. Cada indicador se mapea a fuentes de datos específicas, algoritmos y protocolos de respuesta habilitando un deployment inmediato.

1 Arquitectura de Implementación

La operacionalización del CPF sigue un esquema OFTLISRV sistemático aplicado uniformemente a través de todos los 100 indicadores: Observables (O), Fuentes de Datos (F), Temporalidad (T), Lógica de Detección (L), Interdependencias (I), Umbrales (S), Respuestas (R) y Validación (V). Este esquema garantiza coherencia mientras acomoda las características únicas de cada vulnerabilidad psicológica.

La dimensión temporal se revela crítica para los indicadores psicológicos, ya que estos fenómenos exhiben patrones de persistencia y decaimiento distintos de las métricas de security tradicionales. Definimos los parámetros temporales a través de tres componentes: frecuencia de muestreo f_s, ventana de observación W y umbral de persistencia τ. Para el indicador i al tiempo t, el estado temporal $T_i(t)$ se calcula como:

$$T_i(t) = \alpha \cdot X_i(t) + (1 - \alpha) \cdot T_i(t - 1)$$

donde $\alpha = e^{-\Delta t/\tau}$ proporciona decaimiento exponencial, y $X_i(t)$ representa la observación instantánea.

2 Framework Universal de Detección

La lógica de detección de cada indicador combina reglas determinísticas con detección de anomalías estadísticas. La función de detección base D_i para el indicador i evalúa:

$$D_i = w_1 \cdot R_i + w_2 \cdot A_i + w_3 \cdot C_i$$

donde R_i representa la detección basada en reglas (binario), A_i representa el puntaje de anomalía (continuo) y C_i representa la correlación contextual (normalizada). Los pesos w_1, w_2, w_3 se calibran por organización a través de períodos de baseline iniciales.

La detección de anomalías emplea la distancia de Mahalanobis para tener en cuenta la correlación entre observables:

$$A_i = \sqrt{(x_i - \mu_i)^T \Sigma_i^{-1} (x_i - \mu_i)}$$

donde x_i es el vector de observación, μ_i es la media baseline y Σ_i es la matriz de covarianza actualizada a través de media móvil ponderada exponencial.

3 Implementaciones por Categoría

3.1 Categoría 1: Vulnerabilidades Basadas en la Autoridad

Los indicadores basados en la autoridad (1.1-1.10) monitorean los patrones de compliance con la autoridad percibida a través del análisis de los logs de autenticación, encabezados de email y cadenas de aprobación. La implementación aprovecha los sistemas existentes de Active Directory, gateway de email y gestión de accesos privilegiados.

El indicador 1.1 (Compliance No Cuestionante) se operacionaliza a través del monitoreo continuo de la función tasa de compliance $C_r = \frac{N_{executed}}{N_{requested}}$ donde las solicitudes se originan de patrones authority_domain. La detección se activa cuando $C_r > \mu_{baseline} + 2\sigma$ dentro de la ventana $W = 3600s$. Las fuentes de datos incluyen logs de seguimiento de mensajes Exchange filtrados por sender_domain $\in$ {exec_domains} AND action_keywords $\in$ {transfer, send, approve, grant}. La actualización bayesiana para la legitimidad de la autoridad opera como $P(legitimate|factors) = \frac{P(factors|legitimate)\cdot P(legitimate)}{P(factors)}$ con factores que incluyen time_of_day, request_pattern y verification_attempted.

Los indicadores 1.2-1.4 comparten fuentes de telemetría pero aplican lógicas de detección diferentes. La difusión de responsabilidad (1.2) rastrea las transiciones de propiedad de los tickets donde $T_{ownership} > 3$ dentro del ciclo de vida del incidente indica difusión. La susceptibilidad a la impersonificación de la autoridad (1.3) correlaciona controles SPF/DKIM fallidos con interacciones de usuario exitosas, mientras que el bypass por conveniencia (1.4) monitorea exception_grant_rate durante executive_presence_hours versus normal_hours.

Los indicadores de autoridad restantes emplean patrones arquitecturales similares con lógica adaptada. El compliance basado en el miedo (1.5) incorpora análisis lingüístico para urgency_markers en conjunción con compliance_time. Los efectos de gradiente de la autoridad (1.6) utilizan la profundidad jerárquica organizacional como factor de ponderación. Las afirmaciones de autoridad técnica (1.7) detectan jargon_density que excede las baselines específicas del dominio. La normalización de las excepciones ejecutivas (1.8) rastrea bypass_count acumulativo sobre ventanas móviles de 30 días. La prueba social basada en la autoridad (1.9) emplea análisis de grafo sobre las cascadas de compliance, mientras que la escalación de crisis (1.10) activa monitoreo potenciado cuando external_threat_level excede umbrales predeterminados.

3.2 Categoría 2: Vulnerabilidades Temporales

Las vulnerabilidades temporales (2.1-2.10) se manifiestan a través de degradación de security inducida por la presión temporal. La implementación requiere correlación entre indicadores de ritmo empresarial y métricas de comportamiento de security.

El bypass inducido por la urgencia (2.1) cuantifica a través de $U_i = \frac{\Delta t_{normal} - \Delta t_{urgent}}{\Delta t_{normal}}$ donde Δt representa el tiempo de completación del task. Cuando $U_i > 0.5$, indicando una aceleración del 50%, la eficacia de los controles de security se degrada predeciblemente. La detección emplea modelado de regresión de Poisson que modela la tasa de bypass esperada dada la presión temporal: $\lambda = e^{\beta_0 + \beta_1 \cdot pressure + \beta_2 \cdot deadline\_proximity}$.

La aceptación del riesgo guiada por la fecha límite (2.3) se operacionaliza a través de la integración del sistema de gestión de proyectos, extrayendo deadline_distance y correlacionando con security_exception_requests. La función de descuento hiperbólico $V = \frac{A}{1+k\cdot D}$ modela la percepción del valor donde A es el valor efectivo, D es el retraso y k es la tasa de descuento calibrada por organización.

Los patrones de agotamiento temporal (2.6) requieren modelado circadiano con eficacia de security $E(t) = E_0 \cdot (1 + A \cdot \sin(\frac{2\pi(t-\phi)}{24}))$ donde ϕ representa el desplazamiento de fase y A representa la amplitud de la variación. Los indicadores 2.7-2.9 aprovechan modelado temporal similar con parámetros ajustados para ciclos diferentes (diario, semanal, basado en turnos).

3.3 Categoría 3: Vulnerabilidades de Influencia Social

Los indicadores de influencia social (3.1-3.10) detectan la explotación de la programación social humana a través del análisis de los patrones de comunicación y clustering comportamental.

La explotación de la reciprocidad (3.1) rastrea favor_exchange_networks a través de análisis del sentimiento de email y request_grant_patterns. El índice de reciprocidad $R = \sum_{i,j} w_{ij} \cdot favor_{ij}$ donde w_{ij} representa el peso de la relación derivado de la frecuencia de comunicación. La escalación del compromiso (3.2) identifica request_sequences con sensitivity_scores monotónicamente crecientes.

La manipulación de la prueba social (3.3) emplea procesamiento de lenguaje natural para detectar afirmaciones de acción colectiva: los patrones "todos los demás han" activan verificación potenciada. La implementación utiliza embeddings basados en BERT para identificar similaridad semántica con frases de prueba social conocidas, alcanzando 0.92 de precisión en pruebas.

3.4 Categoría 4: Vulnerabilidades Afectivas

Las vulnerabilidades afectivas (4.1-4.10) correlacionan estados emocionales con calidad de las decisiones de security. La implementación aprovecha marcadores lingüísticos e indicadores comportamentales sin monitoreo invasivo.

La parálisis por miedo (4.1) se manifiesta como aumento de decision_time acoplado con resultados no_action_taken. El índice de miedo $F = \alpha \cdot linguistic\_markers + \beta \cdot response\_latency + \gamma \cdot action\_avoidance$ combina señales múltiples. La asunción de riesgo inducida por la rabia (4.2) correlaciona communication_sentiment con tasa subsecuente risky_action_rate.

La transferencia de confianza (4.3) cuantifica a través de trust_scores diferenciales entre interacciones humanas y de sistema. El apego al legacy (4.4) mide resistance_to_change a través de upgrade_deferral_rate y support_ticket_sentiment respecto a los viejos sistemas.

3.5 Categoría 5: Vulnerabilidades de Sobrecarga Cognitiva

Los indicadores de sobrecarga cognitiva (5.1-5.10) detectan cuando los requisitos de security exceden la capacidad de procesamiento humana. La implementación se concentra en métricas de carga de trabajo y análisis de la tasa de error.

La fatiga de alert (5.1) se operacionaliza como $F_a = 1 - \frac{investigated}{presented}$ con modelado de decaimiento temporal que muestra $F_a(t) = F_0 \cdot e^{\lambda \cdot alert\_rate \cdot t}$. La fatiga decisional (5.2) rastrea la degradación de decision_quality a través de correlación error_rate con decision_count dentro de ventanas temporales.

El overflow de la memoria de trabajo (5.7) aplica el límite de Miller 7 ± 2, señalando cuando concurrent_security_requirements excede el umbral. Los errores inducidos por la complejidad (5.9) correlacionan system_complexity_metrics (complejidad ciclomática, conteo de interfaces) con user_error_rates.

3.6 Categoría 6: Vulnerabilidades de Dinámica de Grupo

Los indicadores de dinámica de grupo (6.1-6.10) detectan estados psicológicos colectivos a través del análisis de la red de comunicación y clustering de los patrones decisionales.

La detección de groupthink (6.1) emplea índices de diversidad sobre los patrones decisionales: $D = 1 - \sum p_i^2$ donde p_i representa la fracción que elige la opción i. Baja diversidad acoplada con consenso rápido indica groupthink. El desplazamiento riesgoso (6.2) confronta group_risk_tolerance con average individual_risk_tolerance, señalando cuando el grupo excede al individual en $> 20\%$.

Los asuntos básicos de Bion (6.6-6.8) se operacionalizan a través de marcadores lingüísticos y comportamentales. La dependencia se manifiesta como aumento de referencia a autoridad/proveedores en las comunicaciones. Fight-flight muestra lenguaje polarizado y comportamientos de evitación. El pairing exhibe lenguaje orientado al futuro sin acciones concretas.

3.7 Categoría 7: Vulnerabilidades de Respuesta al Estrés

Los indicadores de estrés (7.1-7.10) correlacionan marcadores de estrés fisiológico y comportamental con degradación de la eficacia de security.

La detección del estrés agudo (7.1) combina señales múltiples: typing_pattern_deviation, email_response_time_variance y error_rate_increase. El índice de estrés $S = \int_0^t stress\_markers(t) \cdot e^{-\lambda(t-\tau)} d\tau$ incorpora decaimiento temporal.

Las respuestas fight/flight/freeze/fawn (7.3-7.6) clasifican a través de pattern matching comportamental usando modelos de Markov ocultos entrenados sobre datos organizacionales etiquetados. Cada patrón de respuesta exhibe firmas características en los logs de comunicación e interacción de sistema.

3.8 Categoría 8: Vulnerabilidades de Procesos Inconscientes

Los indicadores de proceso inconsciente (8.1-8.10) detectan patrones invisibles a la conciencia consciente a través de manifestaciones comportamentales indirectas.

La proyección de la sombra (8.1) identifica patrones de atribución donde las características de la organización aparecen en las descripciones de las amenazas. La compulsión a la repetición (8.3) detecta fallos de security cíclicos a través de análisis de series temporales con descomposición estacional.

La detección del mecanismo de defensa (8.6) emplea análisis psicolingüístico: la negación se muestra en la frecuencia de negación, la racionalización en la densidad de conjunciones causales, la intelectualización en el uso de nombres abstractos que excede el baseline en $> 30\%$.

3.9 Categoría 9: Vulnerabilidades de Bias Específico de AI

Los indicadores específicos de AI (9.1-9.10) abordan vulnerabilidades de interacción humano-AI únicas a la integración de sistemas automatizados.

La antropomorfización (9.1) cuantifica a través del uso de pronombres cuando se hace referencia a sistemas AI y lenguaje emocional en las interacciones AI. El bias de automatización (9.2) rastrea override_rate cuando las recomendaciones AI entran en conflicto con el juicio humano, señalando cuando override_rate < 0.1.

La aceptación de alucinación AI (9.7) correlaciona puntajes de confianza AI con tasas de aceptación humana, identificando zonas peligrosas donde output AI de baja confianza recibe alta confianza humana.

3.10 Categoría 10: Estados Convergentes Críticos

Los indicadores de estado convergente (10.1-10.10) detectan alineamientos peligrosos de vulnerabilidades múltiples a través de análisis multivariado.

La detección de tormenta perfecta (10.1) emplea el índice de convergencia: $CI = \prod_{i=1}^{n}(1 + v_i)$ donde v_i representa el puntaje de vulnerabilidad normalizado. Cuando $CI > threshold_{critical}$, se activa la escalación defensiva automática.

El alineamiento del queso suizo (10.4) modela las capas defensivas como filtros de probabilidad: $P_{breach} = \prod_{i=1}^{n} p_i$ donde p_i representa la probabilidad de fallo de la capa. El cálculo en tiempo real identifica cuando P_{breach} excede el riesgo aceptable.

4 Modelado de las Interdependencias

La red bayesiana captura las dependencias condicionales entre indicadores. Cada nodo indicador mantiene la distribución de probabilidad $P(I_i|parents(I_i))$. La probabilidad conjunta:

$$P(I_1, ..., I_{100}) = \prod_{i=1}^{100} P(I_i|parents(I_i))$$

Las interdependencias clave incluyen el estrés que amplifica el compliance a la autoridad ($P(1.1|7.1) = 0.8$), la presión temporal que aumenta la sobrecarga cognitiva ($P(5.x|2.x) = 0.7$) y las dinámicas de grupo que enmascaran vulnerabilidades individuales ($P(\neg 4.x|6.x) = 0.6$).

La red habilita consultas predictivas: dados los indicadores observados, calcular la probabilidad de vulnerabilidades no observadas usando propagación de belief. Esto identifica riesgos ocultos que requieren investigación.

5 Framework del Protocolo de Respuesta

Los protocolos de respuesta siguen escalación graduada basada en la gravedad del indicador y en el estado de convergencia. Las respuestas de Nivel 1 se ejecutan automáticamente dentro de 100ms (bloqueo, aislamiento). El Nivel 2 requiere aprobación humana dentro de 5 minutos (suspensión de privilegios, congelación de transacciones). El Nivel 3 activa investigación dentro de 1 hora (análisis comportamental, threat hunting).

La función de respuesta $R(s, c, t)$ considera gravedad s, confianza c y criticidad temporal t:

$$R = \begin{cases} automatic & \text{if } s \cdot c > 0.8 \\ semi\_auto & \text{if } 0.5 < s \cdot c \leq 0.8 \\ manual & \text{if } s \cdot c \leq 0.5 \end{cases}$$

Las operaciones en modo degradado se activan cuando los sistemas primarios fallan, utilizando telemetría de fallback con puntajes de confianza ajustados.

6 Metodología de Validación

Cada indicador sufre validación continua a través de pruebas sintéticas y análisis de correlación. Las pruebas sintéticas inyectan condiciones psicológicas conocidas y miden la precisión de detección. El puntaje de validación:

$$V = \frac{TP \cdot TN - FP \cdot FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

proporciona el coeficiente de correlación de Matthews para clasificadores binarios. Los indicadores continuos usan RMSE entre resultados predichos y observados.

La calibración emplea regresión isotónica garantizando que las probabilidades predichas correspondan a las frecuencias observadas. La detección de drift usando pruebas de Kolmogorov-Smirnov activa recalibración cuando $p < 0.05$.

7 Pragmática de Implementación

El deployment sigue un enfoque por fases: establecimiento de baseline (30 días), deployment piloto (10 indicadores, 60 días), rollout graduado (20 indicadores/mes) y capacidad operativa completa (mes 8). Cada fase incluye ciclos de calibración, validación y ajuste.

La integración con herramientas SOC existentes aprovecha protocolos estándar: syslog para ingestión de logs, STIX/TAXII para threat intelligence, playbooks SOAR para automatización de respuesta. El motor CPF opera como middleware, consumiendo telemetría diversa y produciendo indicadores enriquecidos para sistemas downstream.

Los requisitos de recursos escalan linealmente con el tamaño de la organización: aproximadamente 1TB storage por 1000 usuarios/año, 16 cores para procesamiento en tiempo real por 10000 usuarios y 1 analista por 50 indicadores para mantenimiento y tuning.

8 Conclusión

Esta metodología de implementación transforma las intuiciones teóricas del CPF en capacidades operativas. El esquema OFTLISRV sistemático garantiza implementación coherente a través de todos los 100 indicadores mientras acomoda variaciones organizacionales. La red bayesiana captura interdependencias complejas, habilitando evaluación predictiva del riesgo más allá de los indicadores individuales. Los protocolos de respuesta graduada equilibran automatización

con juicio humano, mientras que la validación continua garantiza eficacia sostenida. Las organizaciones pueden iniciar la implementación inmediatamente usando fuentes de datos existentes, alcanzando mejoras de security medibles dentro del primer ciclo de deployment.