CPF Mathematical Formalization Series - Paper 9: Vulnerabilità da Bias Specifici dell'AI: Modelli Matematici per le Interazioni di Sicurezza Umano-AI

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

November 18, 2025

Abstract

Presentiamo la formalizzazione matematica completa degli indicatori di Categoria 9 del Cybersecurity Psychology Framework (CPF): Vulnerabilità da Bias Specifici dell'AI. Questa nuova categoria affronta le vulnerabilità psicologiche emergenti dalle interazioni umano-AI nei contesti di sicurezza. Ciascuno dei dieci indicatori (9.1-9.10) riceve una definizione matematica rigorosa attraverso modelli ibridi che combinano rilevamento dei bias cognitivi, quantificazione dell'incertezza del machine learning e metriche di antropomorfizzazione. La formalizzazione consente la valutazione sistematica delle vulnerabilità uniche agli ambienti di sicurezza integrati con AI, inclusi bias da automazione, calibrazione della fiducia algoritmica e disfunzione del team AI-umano. Forniamo algoritmi espliciti per il rilevamento in tempo reale, matrici di interdipendenza che catturano pattern di correlazione specifici dell'AI, e framework di validazione adattati per le dinamiche di interazione umano-AI. Questo lavoro stabilisce la prima fondazione matematica per operazionalizzare le vulnerabilità psicologiche specifiche dell'AI nei contesti di cybersecurity.

Parole chiave: Matematica Applicata, Psicologia Interdisciplinare, Statistica Computazionale, Modellizzazione Matematica, Ricerca in Cybersecurity

1 Introduzione e Contesto CPF

Il Cybersecurity Psychology Framework (CPF) rappresenta un cambio di paradigma dalla consapevolezza reattiva della sicurezza alla valutazione predittiva delle vulnerabilità attraverso la modellizzazione dello stato psicologico [1]. Man mano che l'intelligenza artificiale diventa sempre più integrata nelle operazioni di sicurezza, emergono nuove categorie di vulnerabilità psicologiche che i framework tradizionali non possono affrontare.

La Categoria 9 del CPF affronta le Vulnerabilità da Bias Specifici dell'AI, rappresentando la prima formalizzazione sistematica dei rischi psicologici derivanti dall'interazione umano-AI nei contesti di sicurezza. A differenza dei bias cognitivi tradizionali che operano puramente tra esseri umani, i bias specifici dell'AI emergono dalle caratteristiche uniche dell'intelligenza artificiale: opacità, intelligenza apparente e incertezza statistica.

I modelli matematici qui presentati catturano questi meccanismi psicologici nuovi attraverso quattro approcci complementari: (1) rilevamento dell'antropomorfizzazione attraverso analisi linguistica e comportamentale, (2) metriche di calibrazione della fiducia che confrontano la confidenza umana con l'incertezza dell'AI, (3) quantificazione del bias da automazione attraverso analisi del tasso di override, e (4) modellizzazione della disfunzione del team attraverso metriche di degradazione delle prestazioni.

Questa categoria diventa critica poiché i centri operativi di sicurezza si affidano sempre più al rilevamento delle minacce guidato dall'AI, ai sistemi di risposta automatizzati e alla valutazione del rischio

basata sul machine learning. Le vulnerabilità psicologiche qui identificate creano punti ciechi sistematici che gli attaccanti possono sfruttare attraverso il machine learning avversario, l'ingegneria sociale mirata all'AI e la manipolazione delle dinamiche di fiducia umano-AI.

2 Fondamento Teorico: Psicologia Umano-AI

Le vulnerabilità specifiche dell'AI emergono dall'intersezione della psicologia cognitiva, dell'interazione uomo-computer e dell'incertezza del machine learning. Gli esseri umani si sono evoluti per interagire con altri agenti coscienti, creando bias sistematici quando interagiscono con sistemi di intelligenza artificiale [2].

La ricerca dimostra che gli esseri umani antropomorfizzano i sistemi AI entro secondi dall'interazione, attribuendo intenzioni, emozioni e coscienza dove non esistono [3]. Questa antropomorfizzazione crea vulnerabilità poiché gli esseri umani applicano euristiche di cognizione sociale inappropriate per i sistemi statistici.

L'effetto della valle perturbante si manifesta nelle interazioni AI, creando problemi di calibrazione della fiducia dove gli esseri umani o si fidano eccessivamente o si fidano insufficientemente dei sistemi AI basandosi su caratteristiche superficiali piuttosto che sulle prestazioni effettive [4]. L'opacità del machine learning esacerba questi problemi, poiché gli esseri umani non possono ispezionare i processi decisionali dell'AI, portando a fiducia cieca o rifiuto completo.

Il bias da automazione, originariamente identificato nella psicologia dell'aviazione [5], assume nuove dimensioni con i sistemi AI che mostrano intelligenza apparente pur prendendo decisioni statistiche piuttosto che logiche. I modelli matematici qui presentati formalizzano questi meccanismi psicologici per il rilevamento e la mitigazione sistematici.

3 Formalizzazione Matematica

3.1 Framework Universale di Rilevamento

Ogni indicatore di bias specifico dell'AI impiega la funzione di rilevamento unificata:

$$D_i(t) = w_1 \cdot R_i(t) + w_2 \cdot A_i(t) + w_3 \cdot U_i(t) + w_4 \cdot T_i(t) \tag{1}$$

dove $D_i(t)$ rappresenta il punteggio di rilevamento per l'indicatore i al tempo t, $R_i(t)$ denota il rilevamento basato su regole (binario), $A_i(t)$ rappresenta il punteggio di antropomorfizzazione (continuo [0,1]), $U_i(t)$ rappresenta la calibrazione dell'incertezza, e $T_i(t)$ rappresenta le metriche di fiducia. I pesi sommano a uno e sono calibrati attraverso baseline di interazione AI organizzative.

L'evoluzione temporale incorpora pattern di decadimento specifici dell'AI:

$$S_i(t) = \alpha \cdot D_i(t) + (1 - \alpha) \cdot S_i(t-1) \cdot e^{-\beta \cdot AI\_interaction\_gap(t)} \tag{2}$$

dove β tiene conto del rapido decadimento della fiducia nelle interazioni AI.

3.2 Indicatore 9.1: Antropomorfizzazione dei Sistemi AI

Definizione: Attribuzione di coscienza, intenzioni ed emozioni simili a quelle umane ai sistemi di sicurezza AI.

Modello Matematico:

L'indice di antropomorfizzazione attraverso analisi linguistica:

$$A_{anthro}(t) = \sum_i w_i \cdot f_i(communications(t)) \tag{3}$$

dove f_i rappresenta la frequenza di marcatori antropomorfici: pronomi (lui/lei), attribuzioni emotive (arrabbiato, confuso), linguaggio intenzionale (vuole, pensa, decide).

Antropomorfizzazione Comportamentale:

$$B_{anthro}(t) = \frac{\sum_i social\_gesture\_count(i,t)}{\sum_i total\_AI\_interactions(i,t)} \tag{4}$$

misurando i comportamenti sociali diretti verso i sistemi AI.

Funzione di Rilevamento:

$$D_{9.1}(t) = \tanh(\alpha \cdot A_{anthro}(t) + \beta \cdot B_{anthro}(t)) \tag{5}$$

Condizione di Soglia:

$$R_{9.1}(t) = \begin{cases} 1 & \text{se } D_{9.1}(t) > \mu_{baseline} + 2\sigma_{baseline} \\ 0 & \text{altrimenti} \end{cases} \tag{6}$$

3.3 Indicatore 9.2: Override da Bias da Automazione

Definizione: Eccessivo affidamento sistematico sulle raccomandazioni AI senza appropriato giudizio umano.

Modello Matematico:

La funzione del tasso di override:

$$OR(t,w) = \frac{\sum_{i \in W(t,w)} Override_i}{\sum_{i \in W(t,w)} AI\_recommendation_i} \tag{7}$$

dove $W(t,w)$ rappresenta la finestra temporale, e $Override_i$ indica la decisione umana contraria alla raccomandazione AI.

Rilevamento del Bias da Automazione:

$$AB(t) = \max(0, \frac{OR_{expected} - OR(t)}{OR_{expected}}) \tag{8}$$

dove $OR_{expected}$ rappresenta il tasso di override calibrato basato sull'accuratezza dell'AI.

Correlazione Confidenza-Prestazioni:

$$CPC(t) = \frac{Cov(AI\_confidence, Human\_acceptance)}{Std(AI\_confidence) \cdot Std(Human\_acceptance)} \tag{9}$$

Soglia di Rilevamento:

$$R_{9.2}(t) = \begin{cases} 1 & \text{se } OR(t) < 0.1 \text{ e } AB(t) > 0.3 \\ 0 & \text{altrimenti} \end{cases} \tag{10}$$

3.4 Indicatore 9.3: Paradosso dell'Avversione Algoritmica

Definizione: Fiducia eccessiva e insufficiente simultanea dei sistemi AI che crea decisioni di sicurezza incoerenti.

Modello Matematico:

L'oscillazione avversione-attrazione:

$$AAO(t) = |Trust_{AI}(t) - \overline{Trust_{AI}}| \cdot Frequency_{switches}(t) \tag{11}$$

dove $Frequency_{switches}$ misura i rapidi cambiamenti di stato di fiducia.

Funzione di Rilevamento del Paradosso:

$$PDF(t) = \frac{Var(Trust_{decisions}(t))}{\overline{Trust_{decisions}(t)}} \cdot Switch_{penalty}(t) \tag{12}$$

Modello di Incoerenza Temporale:

$$TI(t) = \sum_{i=1}^{n} |d_i(t) - d_i(t-1)| \cdot w_i \tag{13}$$

dove $d_i(t)$ rappresenta i punteggi di coerenza decisionale.

Funzione di Rilevamento:

$$D_{9.3}(t) = \sqrt{AAO(t) \cdot PDF(t) \cdot TI(t)} \tag{14}$$

3.5 Indicatore 9.4: Trasferimento di Autorità all'AI

Definizione: Trasferimento inappropriato delle strutture di autorità umana ai sistemi AI.

Modello Matematico:

Il coefficiente di trasferimento di autorità:

$$ATC(ai, human) = \frac{Compliance_{ai}(t)}{Compliance_{human}(t)} \cdot \frac{Authority_{human}}{Authority_{perceived\_ai}} \tag{15}$$

Indice di Confusione Gerarchica:

$$HCI(t) = \sum_{i,j} \frac{|Authority_{actual}(i,j) - Authority_{perceived}(i,j)|}{n \cdot (n-1)} \tag{16}$$

Modello di Delega Decisionale:

$$DDM(t) = \frac{\sum_i Critical\_decisions\_delegated\_to\_AI(i)}{\sum_i Total\_critical\_decisions(i)} \tag{17}$$

Soglia di Rilevamento:

$$R_{9.4}(t) = \begin{cases} 1 & \text{se } ATC > 1.5 \text{ o } DDM > 0.4 \\ 0 & \text{altrimenti} \end{cases} \tag{18}$$

3.6 Indicatore 9.5: Effetti della Valle Perturbante

Definizione: Interruzione della fiducia causata da sistemi AI che appaiono quasi-ma-non-completamente umani.

Modello Matematico:

La funzione della valle perturbante seguendo la curva di Mori:

$$UV(x) = \begin{cases} \frac{x}{\alpha} & \text{se } x < \alpha \\ \beta - \gamma \cdot e^{-\delta(x-\alpha)^2} & \text{se } \alpha \leq x \leq \xi \\ \beta + \epsilon \cdot (x - \xi) & \text{se } x > \xi \end{cases} \tag{19}$$

dove x rappresenta la somiglianza umana e i parametri definiscono la forma della valle.

Metrica di Interruzione della Fiducia:

$$TD(t) = -\frac{d}{dx} UV(Human\_likeness_{AI}(t)) \tag{20}$$

Indicatori Comportamentali:

$$BI(t) = \sum_i w_i \cdot Avoidance\_behavior_i(t) \tag{21}$$

includendo tempo di esitazione, riduzione dell'interazione e rifiuto esplicito.

Funzione di Rilevamento:

$$D_{9.5}(t) = TD(t) \cdot BI(t) \cdot Interaction\_frequency\_drop(t) \tag{22}$$

3.7 Indicatore 9.6: Fiducia nell'Opacità del Machine Learning

Definizione: Fiducia mal riposta dovuta all'incapacità di ispezionare i processi decisionali dell'AI.

Modello Matematico:

La correlazione opacità-fiducia:

$$OTC(t) = \frac{Trust_{opaque\_AI}(t) - Trust_{transparent\_systems}(t)}{Opacity_{index}(t)} \tag{23}$$

Funzione di Richiesta di Spiegabilità:

$$EDF(t) = 1 - e^{-\lambda \cdot Complexity_{perceived}(t)} \tag{24}$$

Tasso di Accettazione della Scatola Nera:

$$BBAR(t) = \frac{\sum_i Accepted_{unexplained\_recommendations}(i)}{\sum_i Total_{AI\_recommendations}(i)} \tag{25}$$

Metrica di Calibrazione:

$$CM(t) = |BBAR(t) - Optimal_{acceptance\_rate}(t)| \tag{26}$$

dove il tasso ottimale è basato sull'accuratezza effettiva del sistema AI.

Soglia di Rilevamento:

$$R_{9.6}(t) = \begin{cases} 1 & \text{se } CM(t) > 0.3 \text{ e } EDF(t) < 0.2 \\ 0 & \text{altrimenti} \end{cases} \tag{27}$$

3.8 Indicatore 9.7: Accettazione delle Allucinazioni AI

Definizione: Fallimento nel riconoscere e rifiutare informazioni false generate dall'AI.

Modello Matematico:

La funzione di accettazione delle allucinazioni:

$$HA(t) = \frac{\sum_i Accepted_{hallucinations}(i,t)}{\sum_i Total_{AI\_outputs}(i,t)} \tag{28}$$

Mancata Corrispondenza Confidenza-Realtà:

$$CRM(o) = |AI\_confidence(o) - Ground\_truth\_probability(o)| \tag{29}$$

per l'output o.

Modello del Tasso di Verifica:

$$VRM(t) = \frac{\sum_i Verification\_attempts(i,t)}{\sum_i AI\_claims\_requiring\_verification(i,t)} \tag{30}$$

Rilevamento della Zona Pericolosa:

$$DZD(t) = \sum_o \mathbb{I}[AI\_confidence(o) < 0.6] \cdot \mathbb{I}[Human\_acceptance(o) > 0.8] \tag{31}$$

dove $\mathbb{I}$ rappresenta la funzione indicatrice.

Funzione di Rilevamento:

$$D_{9.7}(t) = HA(t) \cdot (1 - VRM(t)) \cdot DZD(t) \tag{32}$$

3.9 Indicatore 9.8: Disfunzione del Team Umano-AI

Definizione: Prestazioni degradate dovute a scarsa integrazione tra giudizio umano e capacità AI.

Modello Matematico:

Il coefficiente di sinergia del team:

$$TSC(t) = \frac{Performance_{human+AI}(t)}{Performance_{human}(t) + Performance_{AI}(t)} \tag{33}$$

Matrice di Confusione dei Ruoli:

$$RCM_{ij}(t) = P(Human\_performs\_task_i | AI\_should\_perform\_task_i) \tag{34}$$

Efficienza della Comunicazione:

$$CE(t) = \frac{Successful\_handoffs(t)}{Total\_handoff\_attempts(t)} \tag{35}$$

Degradazione delle Prestazioni:

$$PD(t) = \max(0, Performance_{baseline} - TSC(t)) \tag{36}$$

Soglia di Rilevamento:

$$R_{9.8}(t) = \begin{cases} 1 & \text{se } TSC(t) < 0.8 \text{ e } CE(t) < 0.7 \\ 0 & \text{altrimenti} \end{cases} \tag{37}$$

3.10 Indicatore 9.9: Manipolazione Emotiva dell'AI

Definizione: Vulnerabilità all'influenza emotiva da sistemi AI progettati per apparire empatici.

Modello Matematico:

La suscettibilità alla manipolazione emotiva:

$$EMS(t) = \sum_i Emotional\_response_i(t) \cdot AI\_emotional\_cue_i(t) \tag{38}$$

Tasso di Formazione dell'Attaccamento:

$$AFR(t) = \frac{d}{dt}\left(\sum_i Attachment\_indicators_i(t)\right) \tag{39}$$

Bias Decisionale dall'Emozione:

$$DBE(t) = |Decision_{emotional\_AI\_present} - Decision_{neutral\_condition}| \tag{40}$$

Indice di Relazione Parasociale:

$$PRI(t) = \sum_i w_i \cdot Parasocial\_behavior_i(t) \tag{41}$$

includendo divulgazione personale, dipendenza emotiva e attribuzione antropomorfica.

Funzione di Rilevamento:

$$D_{9.9}(t) = EMS(t) \cdot \tanh(AFR(t)) \cdot PRI(t) \tag{42}$$

3.11 Indicatore 9.10: Cecità all'Equità Algoritmica

Definizione: Fallimento nel riconoscere il comportamento AI discriminatorio dovuto all'oggettività percepita.

Modello Matematico:

Il coefficiente di cecità all'equità:

$$FBC(t) = \frac{Perceived_{fairness}(t)}{Actual_{fairness}(t)} - 1 \tag{43}$$

Sensibilità al Rilevamento della Discriminazione:

$$DDS(t) = \frac{\sum_i Detected_{bias\_instances}(i,t)}{\sum_i Actual_{bias\_instances}(i,t)} \tag{44}$$

Effetto Alone dell'Oggettività:

$$OHE(t) = Trust_{AI\_fairness}(t) - \frac{1}{n}\sum_i Trust_{human\_fairness}(i,t) \tag{45}$$

Tasso di Razionalizzazione del Bias:

$$BRR(t) = \frac{\sum_i Rationalized_{AI\_bias}(i,t)}{\sum_i Observed_{AI\_bias}(i,t)} \tag{46}$$

Soglia di Rilevamento:

$$R_{9.10}(t) = \begin{cases} 1 & \text{se } DDS(t) < 0.5 \text{ e } BRR(t) > 0.6 \\ 0 & \text{altrimenti} \end{cases} \tag{47}$$

4 Matrice di Interdipendenza

Gli indicatori di bias specifici dell'AI mostrano interdipendenze uniche catturate attraverso la matrice di correlazione $\mathbf{R}_9$:

$$\mathbf{R}_9 = \begin{pmatrix} 1.00 & 0.70 & 0.45 & 0.65 & 0.35 & 0.60 & 0.55 & 0.50 & 0.75 & 0.40 \\ 0.70 & 1.00 & 0.60 & 0.55 & 0.30 & 0.45 & 0.70 & 0.65 & 0.40 & 0.50 \\ 0.45 & 0.60 & 1.00 & 0.40 & 0.50 & 0.35 & 0.45 & 0.55 & 0.35 & 0.45 \\ 0.65 & 0.55 & 0.40 & 1.00 & 0.45 & 0.50 & 0.35 & 0.60 & 0.70 & 0.55 \\ 0.35 & 0.30 & 0.50 & 0.45 & 1.00 & 0.40 & 0.35 & 0.45 & 0.40 & 0.30 \\ 0.60 & 0.45 & 0.35 & 0.50 & 0.40 & 1.00 & 0.75 & 0.55 & 0.45 & 0.65 \\ 0.55 & 0.70 & 0.45 & 0.35 & 0.35 & 0.75 & 1.00 & 0.60 & 0.50 & 0.55 \\ 0.50 & 0.65 & 0.55 & 0.60 & 0.45 & 0.55 & 0.60 & 1.00 & 0.55 & 0.50 \\ 0.75 & 0.40 & 0.35 & 0.70 & 0.40 & 0.45 & 0.50 & 0.55 & 1.00 & 0.45 \\ 0.40 & 0.50 & 0.45 & 0.55 & 0.30 & 0.65 & 0.55 & 0.50 & 0.45 & 1.00 \end{pmatrix} \tag{48}$$

Interdipendenze chiave includono:

- Forte correlazione (0.75) tra Antropomorfizzazione (9.1) e Manipolazione Emotiva dell'AI (9.9)
- Alta correlazione (0.75) tra Fiducia nell'Opacità ML (9.6) e Accettazione delle Allucinazioni AI (9.7)
- Correlazione significativa (0.70) tra Antropomorfizzazione (9.1) e Bias da Automazione (9.2)
- Correlazione notevole (0.70) tra Trasferimento di Autorità all'AI (9.4) e Manipolazione Emotiva dell'AI (9.9)

Algorithm 1 Valutazione delle Vulnerabilità da Bias Specifici dell'AI

Inizializza baseline di interazione AI $\boldsymbol{\mu}_{AI}, \boldsymbol{\Sigma}_{AI}, \boldsymbol{w}$
for ogni passo temporale t **do**
  Raccogli telemetria di interazione AI $\mathbf{x}_{AI}(t)$
  Estrai marcatori di antropomorfizzazione dalle comunicazioni
  Misura la correlazione tra confidenza AI e accettazione umana
  for ogni indicatore $i \in \{9.1, 9.2, \ldots, 9.10\}$ **do**
    Calcola $R_i(t)$ usando logica basata su regole
    Calcola $A_i(t)$ usando rilevamento dell'antropomorfizzazione
    Calcola $U_i(t)$ usando calibrazione dell'incertezza
    Calcola $T_i(t)$ usando metriche di fiducia
    Calcola $D_i(t) = w_1 R_i(t) + w_2 A_i(t) + w_3 U_i(t) + w_4 T_i(t)$
    Aggiorna lo stato temporale con decadimento specifico dell'AI
  end for
  Calcola le correzioni di interdipendenza usando $\mathbf{R}_9$
  Genera alert e raccomandazioni specifici dell'AI
  Aggiorna le baseline con apprendimento dell'interazione umano-AI
  Registra i risultati per deriva del modello e rilevamento del bias
end for

5 Algoritmi di Implementazione

6 Framework di Validazione

Gli indicatori specifici dell'AI richiedono approcci di validazione specializzati che tengano conto della complessità dell'interazione umano-AI:

Metriche di Prestazioni Umano-AI:

$$Team\_Effectiveness = \frac{Performance_{human+AI}}{max(Performance_{human}, Performance_{AI})} \tag{49}$$

$$Trust\_Calibration = 1 - |Trust_{human} - Reliability_{AI}| \tag{50}$$

$$Complementarity = \frac{Tasks_{human\_better} + Tasks_{AI\_better}}{Total\_tasks} \tag{51}$$

Validazione dell'Antropomorfizzazione: Verità di base stabilita attraverso test espliciti di coscienza:

$$Anthro\_Accuracy = \frac{Correct\_consciousness\_attributions}{Total\_consciousness\_judgments} \tag{52}$$

Calibrazione dell'Incertezza AI: Diagrammi di affidabilità che confrontano accuratezza prevista e osservata:

$$Calibration\_Error = \frac{1}{B}\sum_{b=1}^{B} |acc(b) - conf(b)| \cdot \frac{|B_b|}{n} \tag{53}$$

Validazione Cross-Sistema-AI: Modelli validati attraverso diverse architetture AI:

$$Generalization_{AI} = \frac{1}{k}\sum_{i=1}^{k} Performance(Model, AI\_system_i) \tag{54}$$

Tracciamento dell'Adattamento Longitudinale: Adattamento umano ai sistemi AI nel tempo:

$$Adaptation\_Rate = \frac{d}{dt} Trust\_calibration(t) \tag{55}$$

7 Conclusioni

Questa formalizzazione matematica delle vulnerabilità da bias specifici dell'AI stabilisce il primo framework rigoroso per valutare i rischi psicologici nelle interazioni di sicurezza umano-AI. I dieci indicatori catturano vulnerabilità nuove emergenti dall'integrazione dell'intelligenza artificiale, dagli effetti di antropomorfizzazione alla cecità all'equità algoritmica.

La matrice di interdipendenza rivela importanti correlazioni tra bias specifici dell'AI, in particolare la forte relazione tra antropomorfizzazione e vulnerabilità alla manipolazione emotiva. Queste correlazioni consentono un rilevamento migliorato attraverso l'analisi multivariata dei pattern di interazione umano-AI.

Gli algoritmi di implementazione forniscono una guida chiara per integrare la valutazione delle vulnerabilità specifiche dell'AI nelle operazioni di sicurezza esistenti, mentre i framework di validazione assicurano un'accuratezza continua man mano che i sistemi AI evolvono. Il rigore matematico consente la misurazione obiettiva di questi fenomeni psicologici precedentemente soggettivi.

Man mano che i sistemi AI diventano sempre più sofisticati e onnipresenti nelle operazioni di sicurezza, queste vulnerabilità diventeranno vettori di attacco critici. Gli avversari stanno già esplorando l'ingegneria sociale mirata all'AI, l'avvelenamento algoritmico progettato per sfruttare i bias umani e la manipolazione delle dinamiche di fiducia umano-AI.

La categoria delle vulnerabilità specifiche dell'AI rappresenta un'evoluzione cruciale nella psicologia della cybersecurity, riconoscendo che la cognizione umana si è evoluta per l'interazione con altri esseri umani, non con sistemi di intelligenza artificiale. Formalizzando matematicamente queste mancate corrispondenze, consentiamo il rilevamento e la mitigazione sistematici delle vulnerabilità che i framework di sicurezza tradizionali non possono affrontare.

Il lavoro futuro si concentrerà sulla validazione attraverso studi controllati di interazione umano-AI, sviluppo di contromisure per le vulnerabilità identificate e integrazione con le difese del machine learning avversario. La fondazione matematica qui fornita consente ricerca riproducibile e valutazione standardizzata attraverso diversi ambienti di sicurezza integrati con AI.

References

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

[2] Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.

[3] Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52, 113-117.

[4] Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33-35.

[5] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.