

# Categoria 9: Vulnerabilità di Distorsione Specifiche dell'IA

## Contents

<b>Panoramica</b>	<b>2</b>
<b>Indicatori</b>	<b>2</b>
<b>Schema di Implementazione</b>	<b>2</b>
<b>Metriche Chiave</b>	<b>2</b>
Tasso di Override dell'IA . . . . .	2
Punteggio di Distorsione dell'Automazione . . . . .	2
Deriva della Performance del Modello . . . . .	2
<b>Fonti Dati Chiave</b>	<b>2</b>
<b>Approccio di Rilevamento</b>	<b>3</b>
Rilevamento della Distorsione dell'Automazione . . . . .	3
Rilevamento della Deriva del Modello . . . . .	3
Rilevamento di Allucinazioni dell'IA . . . . .	3
<b>Stabilimento della Baseline</b>	<b>4</b>
<b>Tipi di Evento Comuni</b>	<b>4</b>
<b>Livelli di Rischio</b>	<b>4</b>
<b>Strategie di Mitigazione</b>	<b>4</b>
Tecnica . . . . .	4
Organizzativa . . . . .	4
Processo . . . . .	4
<b>Considerazioni Speciali</b>	<b>5</b>
Integrazione LLM . . . . .	5
Robustezza Adversariale . . . . .	5
<b>Risorse Correlate</b>	<b>5</b>

Questa cartella contiene schemi di implementazione dettagliati per tutti i 10 indicatori nella categoria di vulnerabilità specifica dell'IA.

## Panoramica

Le vulnerabilità specifiche dell'IA sfruttano l'eccessiva dipendenza dai sistemi IA/ML, la distorsione dell'automazione, l'opacità algoritmica e i modelli unici di interazione umano-IA.

## Indicatori

1. [9.1] **Antropomorfizzazione dei Sistemi IA** - Attribuzione di qualità umane ai sistemi IA
2. [9.2] **Distorsione dell'Automazione** - Preferenza per le decisioni automatizzate rispetto al giudizio umano
3. [9.3] **Paradosso dell'Avversione agli Algoritmi** - Sfiducia verso sistemi IA affidabili
4. [9.4] **Trasferimento di Autorità dell'IA** - Cessione della responsabilità decisionale all'IA
5. [9.5] **Effetti della Valle Inquietante** - Disagio causato da sistemi IA quasi-umani
6. [9.6] **Fiducia nell'Opacità dell'Apprendimento Automatico** - Accettazione di modelli "black box" senza comprensione
7. [9.7] **Accettazione di Allucinazioni dell'IA** - Trattenere come veritieri gli output dell'IA errati ma confidenti
8. [9.8] **Disfunzione del Team Umano-IA** - Fallimenti nella collaborazione nei team ibridi
9. [9.9] **Manipolazione Emotiva dell'IA** - L'IA che sfrutta le emozioni umane
10. [9.10] **Cecità di Equità Algoritmica** - Mancata riconoscimento dei bias nei sistemi IA

## Schema di Implementazione

Ogni indicatore segue il framework **OFTLISRV** con monitoraggio del sistema IA.

## Metriche Chiave

### Tasso di Override dell'IA

AOR = Sovrascritture\_umane / Raccomandazioni\_IA

Valori molto bassi (<5%) o molto alti (>50%) indicano disfunzione.

### Punteggio di Distorsione dell'Automazione

ABS = Errori\_IA\_accettati / Errori\_IA\_totali

### Deriva della Performance del Modello

MPD = (Accuratezza\_attuale - Accuratezza\_baseline) / Accuratezza\_baseline

## Fonti Dati Chiave

- **Sistemi IA/ML:** Log di previsioni, punteggi di confidenza, importanza delle feature
- **SIEM:** Avvisi generati dall'IA, output del modello ML
- **Decisioni Utente:** Override, accettazioni, modifiche delle raccomandazioni IA
- **Metriche del Modello:** Accuratezza, precisione, recall nel tempo

- Dati di Incidente: Falsi positivi/negativi dai sistemi IA

## Approccio di Rilevamento

### Rilevamento della Distorsione dell'Automazione

```
# Tracciare accettazione vs verifica
ai_recommendations = get_ai_outputs(window=7_days)

for recommendation in ai_recommendations:
    if recommendation.accepted and not recommendation.verified:
        if recommendation.confidence < 0.8: # Bassa confidenza
            flag_automation_bias(user_id)

    # Verificare se gli errori sono catturati
    if recommendation.actual_result == 'false_positive':
        if recommendation.accepted_without_override:
            automation_bias_errors += 1
```

### Rilevamento della Deriva del Modello

```
# Monitorare la performance del modello nel tempo
current_metrics = model.evaluate(recent_data)
baseline_metrics = load_baseline_metrics()

drift = {
    'accuracy': current_metrics.accuracy - baseline_metrics.accuracy,
    'precision': current_metrics.precision - baseline_metrics.precision,
    'recall': current_metrics.recall - baseline_metrics.recall
}

if any(abs(d) > 0.1 for d in drift.values()): # >10% degrado
    alert_model_drift(model_id)
```

### Rilevamento di Allucinazioni dell'IA

```
# Rilevare output confidenti ma errati
predictions = get_ai_predictions()

hallucinations = [
    p for p in predictions
    if p.confidence > 0.9 and p.actual_result == 'error'
]

if len(hallucinations) / len(predictions) > 0.05: # >5% tasso
    flag_hallucination_risk(model_id)
```

## Stabilimento della Baseline

Gli indicatori specifici dell'IA richiedono: - Metriche iniziali di performance del modello - Modelli di collaborazione umano-IA - Baseline del tasso di override - Pianificazione del riaddestramento del modello

## Tipi di Evento Comuni

- `ai_recommendation_followed` → 9.1, 9.2, 9.3
- `model_prediction_error` → 9.4, 9.5, 9.7, 9.10
- `ai_explanation_missing` → 9.6
- `human_ai_disagreement` → 9.8
- `ai_generated_content` → 9.9

## Livelli di Rischio

- **Basso** (0-0.33): Scetticismo salutare, uso appropriato dell'IA
- **Medio** (0.34-0.66): Qualche eccessiva dipendenza, verifica ancora presente
- **Alto** (0.67-1.00): Fiducia cieca nell'IA, pensiero critico sospeso

## Strategie di Mitigazione

### Tecnica

- Soglie di confidenza per l'accettazione automatica
- Revisione umana obbligatoria per decisioni ad alto impatto
- Dashboard di monitoraggio della performance del modello
- Implementazioni di IA Spiegabile (XAI)
- Programmi di test adversariali

### Organizzativa

- Formazione sulla literacy dell'IA
- Comprensione dei limiti del ML
- Requisiti di human-in-the-loop
- Audit regolari del modello
- Team di sviluppo IA diversificati

### Processo

- Pianificazione del riaddestramento del modello
- Avvisi di degrado della performance
- Requisiti di documentazione degli override
- Test di ipotesi alternative
- Audit di bias dei dati di training

## Considerazioni Speciali

### Integrazione LLM

Per sistemi CPF basati su RAG che utilizzano LLM: - Validare gli output dell'LLM rispetto alla verità fondamentale - Monitorare le allucinazioni nelle valutazioni psicologiche - Mantenere la supervisione di esperti umani - Controllo di versione per l'ingegneria dei prompt

### Robustezza Adversariale

- Testare i modelli contro esempi adversariali
- Monitorare i tentativi di evasione
- Implementare metodi ensemble
- Esercizi regolari di red team

## Risorse Correlate

- **Fondamento Denso:** [/foundation\\_docs/core/it-IT/](/foundation_docs/core/it-IT/) - Formalizzazione della vulnerabilità IA
- **Piano Grafico CPF LLM:** Documento CPF principale - Metodologia di integrazione RAG
- **Dashboard:** </dashboard/soc/> - Metriche di performance dell'IA
- **Ricerca:** Team umano-IA nella cybersicurezza