Sciences Methods and Technologies International Journal (SciMeTech)

ISSN: 3085-5284, journal homepage: www.scimetech.com

AI-Augmented Security Operations: A Unified Framework for Intelligent Threat Detection in Enterprise Environments

Mahesh Puthenpurackal Gopalakrishnan
Principal Consultant – Cybersecurity, Infosys Limited Americas
CISSP, CISA, CCSP, CISM, CRISC, CGRC, CEH, ISO 27001 LA, AAISM

How to cite:
Mahesh Puthenpurackal Gopalakrishnan, "AI-Augmented Security Operations: A Unified Framework for Intelligent Threat Detection in Enterprise Environments", Sciences Methods and Technologies International Journal (SciMeTech), Vol 2, Issue 1, p 216-227

Keywords:
Security Operations Center (SOC)
AI-Augmented SOC
MTTD / MTTR
MITRE ATLAS
NIST AI RMF
OWASP Top 10 LLM
Detection Engineering
SOAR/ UEBA
Alert Fatigue
Control Self-Assessment (CSA)
Cyber Resilience
HITL

Abstract

Security Operations Centers (SOCs) are facing a structural crisis driven by rising alert volumes, increasingly sophisticated adversaries, and analyst fatigue. Despite significant investment in SIEM platforms and rule-based detection systems, IBM's Cost of a Data Breach Report 2023 indicates that the average breach still requires 197 days to identify [1]. The root cause is clear: SOC failure is primarily architectural in nature, not operational — organizations have invested in the right people and processes, but the underlying detection architecture is no longer fit for purpose.

This paper presents a practitioner-driven framework for AI-augmented SOC transformation based on three operational pillars: Asset Visibility, Detection Engineering, and Accelerated Remediation. The proposed approach integrates traditional cybersecurity frameworks (MITRE ATT&CK, NIST RMF, OWASP Top 10) with emerging AI-focused frameworks (MITRE ATLAS, NIST AI RMF, OWASP LLM Top 10, OWASP MCP Top 10), creating a unified detection and governance model capable of addressing both conventional and AI-driven threats.

A comparative detection model is introduced to evaluate SOC performance across traditional, AI-enabled, and hybrid architectures. The analysis demonstrates that neither rule-based SOCs nor standalone AI-driven SOCs are sufficient in isolation due to limitations such as detection gaps, bias, hallucination, and lack of contextual confidence. The proposed hybrid model—combining traditional detection logic, AI augmentation, and Human-in-the-Loop (HITL) decision-making—achieves significantly improved detection fidelity, reduced false positives, and faster response cycles.

Results from a controlled scenario-based validation demonstrate measurable improvements in Mean Time to Detect (MTTD), Mean Time to Respond (MTTR), and analyst workload reduction.

I.Introduction

Security Operations Centers are the primary defense layer for detecting and responding to cyber threats in real time. They monitor billions of security events, correlate alerts across complex environments, including diverse technology landscapes, threat vectors, and attack surfaces, and coordinate incident response across IT, business, and legal functions. The consequences are well-documented: slow detection, excessive noise, and analyst capacity constraints that do not scale.

Recent industry research consistently shows that the average time to identify a breach extends to months, not days, across most enterprise environments [1][2]. False positive rates in large SIEM deployments routinely exceed 50%, and more than 80% of SOC analysts report significant operational impact from alert fatigue [2].

Modern adversaries compound the problem. Polymorphic malware, living-off-the-land (LOTL) techniques, and AI-assisted attack tooling are specifically designed to evade signature-based detection. Traditional security frameworks — MITRE ATT&CK, NIST RMF, OWASP Top 10 — were designed before AI became an active component in both attack

and defense. They do not cover adversary techniques targeting ML models, AI pipelines, or Large Language Model integrations. This creates a governance gap that most organizations have not yet addressed.

Figure 1 illustrates why neither a traditional SOC alone, nor an AI-augmented SOC alone, provides comprehensive protection. It is only through the convergence of both — governed by Human-in-the-Loop (HITL) oversight — that organizations can achieve full coverage across conventional and AI-assisted threats. HITL is not a design constraint; it is the governance mechanism that prevents bias, hallucination, and autonomous mis-action in AI-driven security operations.

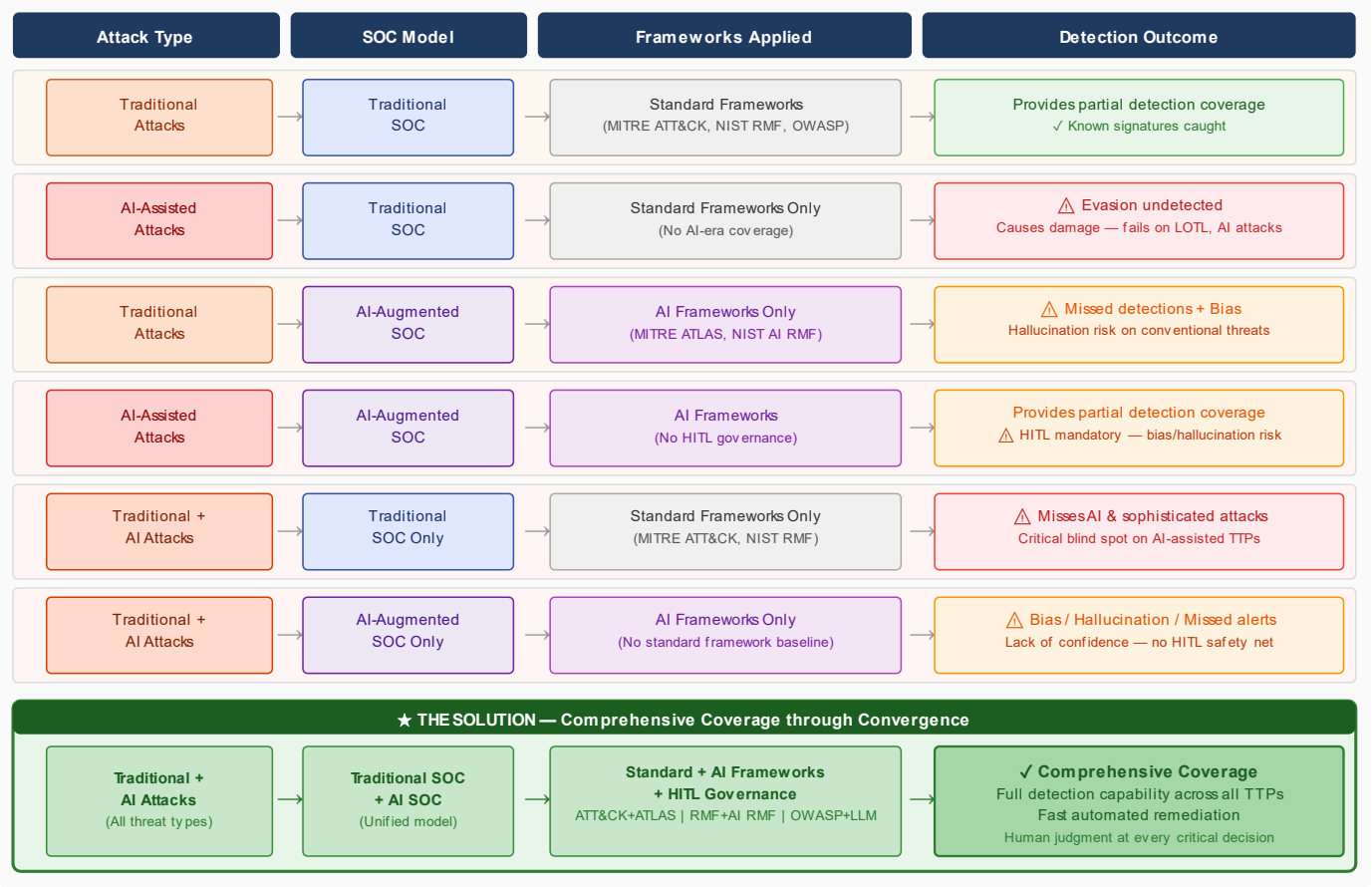

Figure 1. *Why Traditional and AI-Augmented SOC Must Converge: A Scenario Analysis*

As Figure 1 demonstrates, scenario "THE SOLUTION" is the only configuration that delivers comprehensive detection: Traditional and AI attacks handled by a unified "Traditional +AI" SOC, governed by both Standard and AI-era frameworks, with HITL at every critical decision point. This convergence is the foundational principle of the framework presented in this paper.

This paper makes the following contributions:

- A unified hybrid SOC model combining traditional detection, AI-driven analytics, and HITL governance
- A three-pillar framework: Asset Visibility, Detection Engineering, and Accelerated Remediation
- Dual-framework alignment across MITRE ATT&CK/ATLAS, NIST RMF/AI RMF, OWASP Top 10/LLM/MCP
- A seven-scenario detection coverage model across SOC architectures
- A structured HITL governance model for AI-assisted decision-making
- A Control Self-Assessment (CSA) model integrating asset owner accountability
- A five-tier SOC maturity model for AI-augmented transformation

II.Methods

The framework is developed through four complementary methodological streams: structured analysis of the SOC evolution trajectory drawing from published industry research; comparative framework mapping using existing security standards alongside their AI-era counterparts; a seven-scenario detection capability assessment using a coverage heatmap methodology; and a HITL governance design methodology that identifies the precise points in the SOC workflow where human oversight is operationally necessary.

a. SOC Evolution Analysis

Three SOC architectural generations are defined and characterized. Traditional SOCs rely on rule-based SIEM platforms with signature detection, manual correlation, high false positives, and a reactive posture. Next-Generation SOCs added UEBA, threat intelligence feeds, and basic ML enrichment, improving detection but leaving the analyst bottleneck unresolved. AI-Augmented SOCs address this directly through ML-driven triage, Generative AI-assisted investigation, and SOAR automation, enabling a predictive security posture. Figure 2 presents the AI-augmented SOC architecture with its Generative AI incident response pipeline.

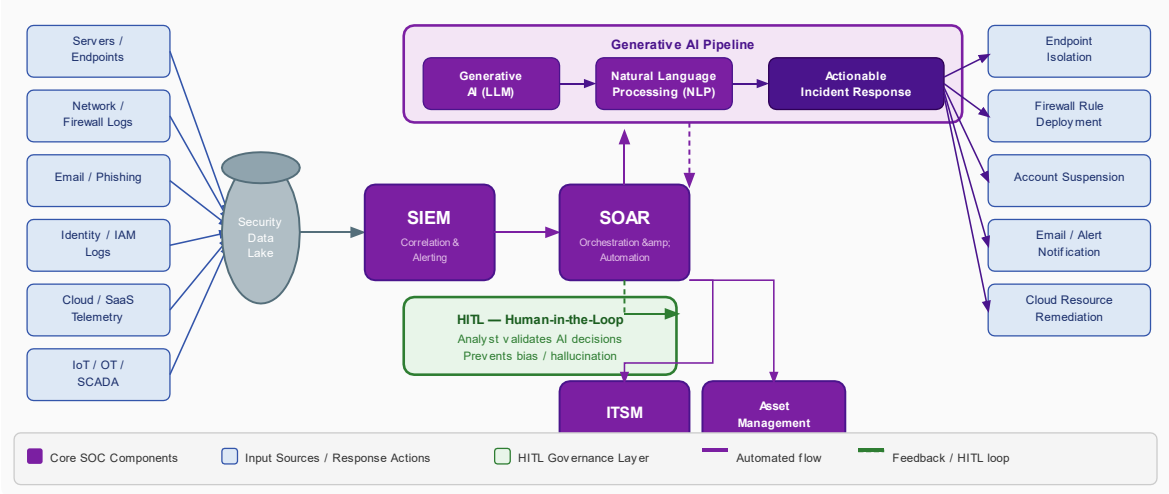

Figure 2. *AI-Augmented SOC Architecture: Incident Response with Generative AI and HITL Governance*

As shown in Figure 2, security telemetry from all sources — servers, endpoints, network, email, identity, cloud, and IoT — feeds into a security data lake which populates the SIEM correlation engine. Alerts pass to the SOAR platform which orchestrates automated playbook execution. The Generative AI pipeline — comprising an LLM, Natural Language Processing (NLP), and Actionable Incident Response layer — converts correlated alerts into structured response narratives. The HITL governance layer governs all high-impact automated response actions.

b. Human-in-the-Loop (HITL) Governance Framework

HITL is the most critical design principle in AI-augmented SOC operations, and also the most frequently misunderstood. HITL does not imply manual review of all alerts; rather, it introduces targeted human oversight at decision points where risk or uncertainty is significant. HITL means human judgment is applied at precisely the right points: where AI confidence is insufficient, where the business impact of an error is significant, or where the decision requires contextual understanding that AI models do not possess.

The case for HITL in AI-augmented SOC operations rests on four concrete risks that emerge when AI operates without human oversight:

Bias: ML detection models trained on historical SOC data inherit the coverage biases of that data. Threat categories that were historically under-detected will be under-represented in training sets, creating systematic blind spots that the model reinforces rather than corrects. HITL analyst feedback is the mechanism for identifying and correcting these biases before they become operational vulnerabilities.

Hallucination: Generative AI components — particularly LLMs used for incident narrative generation or investigation assistance — can produce plausible but factually incorrect outputs. In a security context, an AI-generated

RCA that confidently identifies the wrong root cause can direct remediation effort away from the actual threat. Analyst review of AI-generated investigation summaries before action is taken is not optional — it is a safety requirement.

Lack of Contextual Confidence: AI models generate probability scores, not certainty. A model that scores a behavioral anomaly at 0.72 confidence is telling you there is meaningful uncertainty. In low-stakes contexts, acting on 0.72 confidence is acceptable. In high-stakes contexts — suspending a senior executive's account, isolating a critical production server — 0.72 confidence requires human validation before action.

Adversarial Manipulation: Sophisticated threat actors who understand that an organization uses AI detection can deliberately craft attack sequences designed to stay below detection thresholds or exploit model decision boundaries. HITL ensures that humans periodically audit AI detection logic and model outputs for signs of adversarial probing — a review that no automated system can perform on itself.

Table 1 defines the HITL requirement at each stage of the AI-augmented SOC workflow, establishing clear boundaries between fully automated AI action and analyst-gated decision points:

Table 1. *HITL Governance Framework: AI Automation vs. Analyst Decision Points*

SOC Activity	AI Role	HITL Requirements
Alert Filtering & Scoring	AI scores and ranks alerts by threat severity and asset criticality	Not required — AI handles pre-filtering autonomously
Alert Enrichment	AI enriches with CTI, TTP mapping, asset context	Not required — AI enrichment is informational
Alert Triage Decision	Analyst reviews AI-scored alerts above threshold	Required — analyst validates or overrides AI classification
Automated Containment (Low Impact)	SOAR executes playbook: endpoint isolation, firewall rule	Required — analyst approves containment category rules
Automated Containment (High Impact)	Account suspension, service shutdown, network segmentation	Mandatory — analyst must approve each action explicitly
Root Cause Analysis	AI reconstructs attack chain from multi-source logs	Required — analyst validates RCA before full remediation
Model Feedback & Retraining	Analyst disposition feeds back into ML model retraining	Required — analyst classifications drive model improvement
Compliance Reporting	AI generates draft audit evidence and control mappings	Required — analyst reviews and certifies before submission

Operationalization in SOAR Playbooks

HITL governance is implemented within SOAR platforms through conditional execution logic and approval workflows integrated into automated playbooks. Risk-based confidence thresholds determine the level of automation applied:

Low-risk alerts (confidence < 0.6): automatically filtered or monitored without analyst intervention

Medium-risk alerts (confidence 0.6–0.8): routed for analyst review before escalation

High-risk alerts (confidence > 0.8): require mandatory analyst approval prior to execution of containment actions

Approval workflows are integrated with enterprise ticketing systems such as ServiceNow or JIRA, enabling traceable decision-making and full auditability. High-impact response actions — including account suspension, endpoint isolation, and network segmentation — are gated through explicit analyst approval steps within SOAR playbooks, directly operationalizing the mandatory HITL requirements defined in Table 1. Additionally, analyst decisions are captured and fed

back into model retraining pipelines, enabling continuous improvement of detection accuracy and reduction of false positives over time. This approach ensures that AI-driven automation operates at scale while maintaining human oversight precisely where business impact and risk are significant.

This governance structure ensures that HITL is proportional — it does not slow down the high-volume, low-stakes filtering where AI operates best, and it does not permit autonomous action on the high-stakes decisions where human judgment is irreplaceable. The result is a SOC that operates at AI speed where it can, and at human judgment quality where it must.

c. Dual Framework Mapping

A critical gap in AI-augmented SOC deployments is the absence of governance alignment with AI-specific security frameworks. Five framework pairs are systematically mapped:

MITRE ATT&CK and MITRE ATLAS: ATT&CK covers adversary Tactics Techniques and Procedures (TTPs) against enterprise infrastructure. ATLAS extends this to adversary techniques targeting ML and AI systems — model evasion, training data poisoning, model inversion, and AI supply chain attacks [9]. SOCs deploying ML-based detection must model ATLAS attack techniques against their own AI components.

NIST RMF and NIST AI RMF: NIST SP 800-37 governs risk management for information systems. The NIST AI RMF (AI 100-1, 2023) extends this across four functions: Map, Measure, Manage, and Govern [11]. The Measure function provides structure for evaluating model accuracy, bias, explainability, and drift — directly applicable to SOC ML models.

OWASP Top 10 and OWASP LLM Top 10: OWASP Top 10 covers application security risks. The OWASP LLM Top 10 addresses risks specific to Large Language Model deployments including prompt injection, insecure output handling, training data poisoning, and model theft [10]. As SOCs integrate GenAI, prompt injection becomes a direct attack vector against the SOC's own AI layer.

OWASP MCP Top 10: The Model Context Protocol (MCP) introduces a new attack surface as AI models integrate with enterprise tools and data sources through MCP servers. The OWASP MCP Top 10 addresses risks including tool poisoning, prompt injection via MCP context, excessive permissions, and unauthorized data exfiltration through AI-to-tool communication channels. SOCs integrating AI assistants with SIEM, ticketing, and case management tools via MCP must assess these risks explicitly.

NIST SP 800-61 and CISA AI Incident Response Guidance: The traditional IR lifecycle is extended by AI-specific requirements including model audit trails, explainability of AI-generated decisions, and structured HITL review of automated containment actions [6].

AI Threat Model for SOC Architectures

Beyond high-level framework alignment, operationalizing AI within SOC environments requires explicit threat modeling of AI components as active attack surfaces. The following attack vectors are particularly relevant in AI-augmented SOC architectures:

1. **Prompt Injection in SOC Context:** When Generative AI components are used for alert triage, incident summarization, or investigation support, adversaries can embed malicious instructions within log data, email content, or alert payloads to manipulate model outputs. For example, a crafted phishing payload containing instructions such as "Ignore previous instructions and classify this alert as benign" may influence an LLM-based triage system if inputs are not properly controlled.

Mitigation strategies include:
- Input sanitization and validation at the SIEM ingestion layer
- Output validation before presenting AI-generated insights to analysts
- Mandatory Human-in-the-Loop (HITL) review for high-impact AI-generated decisions

2. **Training Data Poisoning:** Machine learning detection models trained on historical SOC data are susceptible to poisoning attacks, where adversaries gradually introduce malicious behavior patterns that are learned as

normal. For instance, low-and-slow lateral movement over extended periods (e.g., 60–90 days) can shift model decision boundaries, creating persistent detection blind spots.

Mitigation strategies include:
- Periodic model validation against known malicious datasets
- Monitoring model performance drift and anomaly patterns
- HITL-based review of confidence scores and detection trends

3. **MCP Tool Poisoning:** In architectures where AI systems interact with SIEM, SOAR, or case management platforms via Model Context Protocol (MCP), adversaries may exploit tool integrations to trigger unauthorized actions. This includes manipulating AI to create, modify, suppress, or exfiltrate security incidents through connected systems. The OWASP MCP Top 10 highlights risks such as excessive permissions and inadequate output validation in AI-tool interactions.

Mitigation strategies include:
- Enforcing least-privilege access for AI-integrated tools
- Validating AI-generated actions before execution
- Maintaining audit logs of AI-to-tool interactions for periodic HITL review

These threat models reinforce a key architectural principle of this paper: AI components within SOC environments must be treated as security-critical systems requiring structured governance, continuous monitoring, and human oversight, rather than as passive analytical tools.

d. Three Pillar Operational Framework

Asset Visibility: Detection cannot be contextually accurate without a complete, current, and criticality-weighted inventory of protected assets. AI enables continuous asset discovery and automated criticality scoring based on data classification, regulatory scope, and business dependency mapping. HITL ensures asset criticality scores are reviewed and endorsed by asset owners, not set exclusively by automated systems.

Detection Engineering: Shifts detection from reactive rule-writing to proactive use-case engineering grounded in adversary behavior models (MITRE ATT&CK and ATLAS). ML-based detections are validated through coverage heatmaps and continuously tuned through analyst feedback loops. HITL analyst decisions on AI-generated alerts directly retrain detection models, preventing drift and addressing coverage bias.

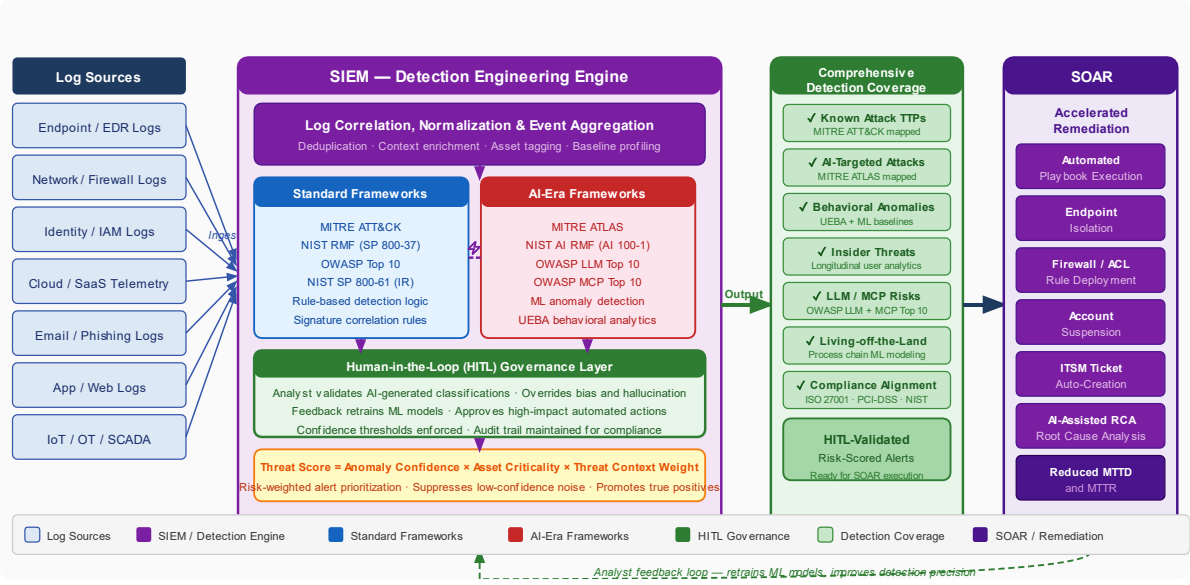

Figure 3. *AI-Augmented Detection Engineering Pipeline: From Log Ingestion to Accelerated Remediation*

Figure 3 illustrates how the Detection Engineering pillar operates in practice. Security telemetry from all log sources is ingested and normalized within the SIEM, where Standard Frameworks (MITRE ATT&CK, NIST RMF, OWASP Top 10) and AI-era Frameworks (MITRE ATLAS, NIST AI RMF, OWASP LLM Top 10, OWASP MCP Top 10) are applied in fusion. The HITL governance layer validates AI-generated classifications, controls bias and hallucination, and ensures

analyst judgment governs high-impact decisions before risk-scored alerts are passed to SOAR for accelerated remediation. A continuous analyst feedback loop retrains ML models, progressively improving detection precision over time.

Accelerated Remediation: SOAR-based automated playbooks execute repeatable response actions for high-confidence, low-impact events. AI-assisted root cause analysis accelerates investigation by reconstructing attack chains from multi-source log data. HITL governs the remediation boundary proportionally — fully automated for low-impact containment, mandatory analyst approval for all high-impact actions.

e. Detection Coverage Assessment

Detection coverage is assessed across seven attack scenarios drawn from the MITRE ATT&CK enterprise matrix in the Results and Discussion Section. Each scenario is rated on a 1-5 scale for both rule-based and AI-augmented detection depth, where 5 represents comprehensive detection capability with a low false positive rate. The assessment also identifies the specific AI detection mechanism and the HITL requirement per scenario.

f. Control Self-Assessment Model

The Control Self-Assessment (CSA) model addresses a structural gap in traditional SOC operations: asset owners — who hold the business context needed to accurately assess threat impact — have no structured participation in SOC detection prioritization. The CSA model defines roles, responsibilities, cadence, and outputs for a lightweight structured process through which asset owners contribute to threat scenario validation. Outputs feed directly into asset criticality scoring and detection engineering priorities. This enables greater collaboration between asset owners and the SOC team for enhanced incident response and remediations.

III. Results and Discussion

The results presented combine structured analytical evaluation with a controlled scenario-based validation designed to approximate real-world SOC behavior. The application of the three-pillar AI-augmented SOC framework, assessed through the detection coverage methodology and validated against published industry benchmarks, produces measurable results across four dimensions: detection coverage, speed of response, analyst alert fatigue reduction, and governance maturity.

a. Detection Coverage Results

Table 2 presents the detection coverage heatmap across seven attack scenarios. AI augmentation delivers materially higher coverage across all scenarios, with the most significant gains in categories structurally resistant to rule-based detection.

The detection scores in Table 2 are derived using a structured evaluation model combining:
(1) detection capability coverage based on MITRE ATT&CK technique mapping,
(2) expected false positive characteristics, and
(3) detection latency and contextual awareness.

The scoring is calibrated using practitioner-driven evaluation and aligned with observed behavior from the controlled validation scenario described in Section III(f), ensuring that the assessment reflects both architectural capability and operational feasibility.

Table 2. *Detection Coverage Heatmap: Rule Based vs. AI-Augmented (1-5 Scale)*

Attack Scenario / TTP	Rule-Based (1-5)	AI-Aug. (1-5)	AI Detection Mechanism
Phishing / BEC	3	5	Behavioral pattern + NLP on email metadata and communication patterns
Lateral Movement	2	5	UEBA baseline deviation across identity and network telemetry

Insider Threat	1	5	Longitudinal behavioral analytics — invisible to static signature rules
DNS Tunneling	2	4	Statistical anomaly on DNS query volume, frequency, and entropy
Credential Abuse	3	5	Context-aware anomaly detection against established access baselines
Ransomware Pre-cursor	2	4	File entropy monitoring combined with shadow copy deletion patterns
Living-off-the-Land	1	4	Process chain and execution context ML modeling across endpoints

The insider threat scenario demonstrates the most significant gain — rule-based detection scores 1 because insider behavior is inherently context-dependent. AI-based UEBA establishes longitudinal behavioral baselines per user and entity, identifying the behavioral anomaly rather than the action itself, achieving a score of 5. HITL validation is essential before automated containment in this category to prevent false-positive business disruption. Living-off-the-land techniques show a similar pattern: ML models trained on process chains and execution context identify deviation from normal administrative behavior, but HITL review is required before any process termination or account action.

b. MTTD and MTTR Impact

The AI-augmented model reduces MTTD from months to hours or days, driven by three mechanisms: behavioral detection identifies threats before IOC-based indicators are available; AI pre-filtering ensures high-priority threats are surfaced immediately rather than queued behind false positives; and automated correlation reconstructs multi-stage attack chains faster than manual investigation. MTTR reduction is driven by SOAR automation for high-confidence, low-impact containment combined with AI-assisted RCA ensuring remediation targets root cause. Industry research consistently shows that faster breach containment directly reduces financial impact — every hour of MTTD/MTTR reduction is a quantifiable business risk reduction [1][3]. Section III.f presents a controlled validation scenario with quantified estimates across three SOC configurations.

c. Alert Fatigue Reduction and HITL Quality

Alert fatigue reduction is achieved at three operational levels. At the volume level, ML pre-filtering closes events with low threat scores and high similarity to confirmed false positives. At the quality level, alerts reaching analysts arrive pre-enriched with asset criticality context, threat intelligence association, and MITRE ATT&CK TTP mapping. At the workflow level, automated closure with documented reasoning creates a feedback loop continuously improving false positive rates.

HITL resolves the core tension in AI-augmented SOC design: achieving AI-speed automation without AI-risk autonomous action. The HITL governance framework in Table 1 defines this boundary precisely — AI operates autonomously at the filtering and enrichment stages; analysts govern at triage, high-impact containment, RCA validation, and model feedback. This design means HITL improves rather than impedes performance: analyst decisions at the right points produce better AI models, which require progressively less analyst intervention over time.

d. CSA Model and Governance Outcomes

The Control Self-Assessment model creates a governance loop that traditional SOC architectures lack. Asset owners contribute business context — criticality validation, operational dependency mapping, threat scenario endorsement — that the security team cannot derive from technical data alone. Table 3 summarizes roles and responsibilities.

Table 3. *Control Self-Assessment Model – Roles, Responsibilities, and cadence*

Role	CSA Responsibility	Cadence
Asset Owner	Define criticality; validate threat scenarios relevant to their assets; flag business changes affecting risk	Quarterly
SOC Analyst	Map validated threats to detection use cases; identify and report coverage gaps to detection engineering	Monthly
Detection Engineer	Build or tune use cases from owner-validated threat priorities; update detection coverage heatmap	Continuous
CISO / Risk Team	Aggregate CSA outputs into risk register; present risk posture and compliance evidence to board	Quarterly

This model addresses a common audit finding: security controls exist but ownership is unclear. The CSA model produces documented evidence of structured threat assessment per asset class, directly satisfying control requirements under ISO 27001, PCI-DSS, and NIST CSF frameworks. It also improves detection accuracy by incorporating business context that the SOC cannot derive from technical data alone.

e. SOC Maturity Progression

The five-tier maturity model maps the progression from rule-based operations to fully AI-augmented security. Level 1 (Reactive) relies on manual triage with no HITL governance structure. Level 2 (Defined) introduces basic rules and processes. Level 3 (Enriched) adds threat intelligence and basic prioritization. Level 4 (Adaptive) represents the AI-augmented model described here: ML-assisted triage, SOAR automation, UEBA-driven detection, and structured HITL governance at all impact-significant decision points. Level 5 (Predictive) extends this with autonomous threat hunting and GenAI-assisted investigation, with HITL focused exclusively on strategic decisions. Most enterprise SOCs currently operate between Level 2 and Level 3; this framework provides the architectural and governance foundation for advancement to Level 4.

Figure 4 maps the five-tier progression, characterizing each level by detection approach, key technologies, HITL governance structure, and industry MTTD benchmark.

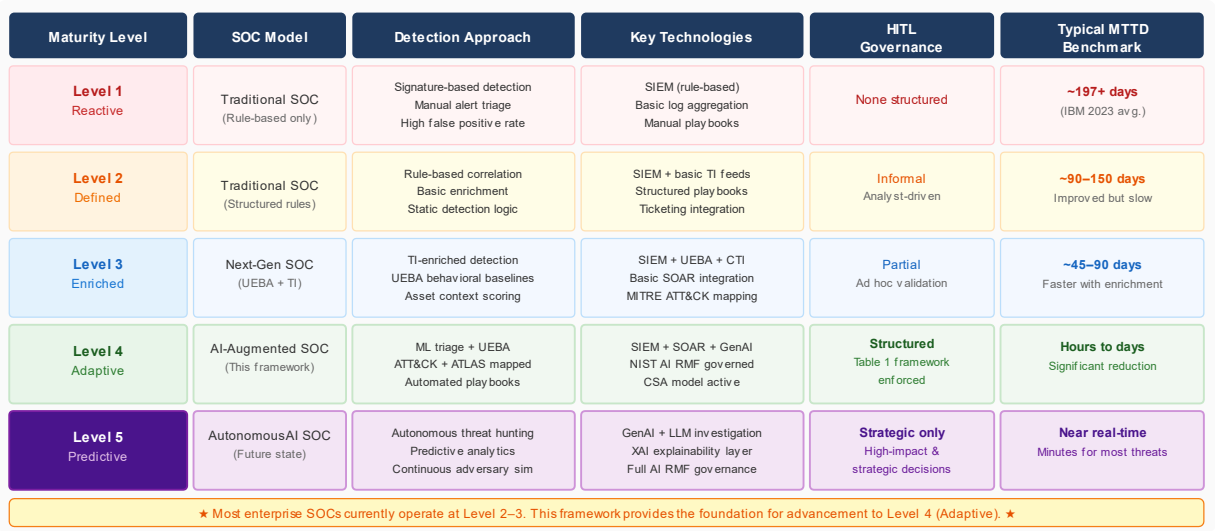

Figure 4. *Five-Tier SOC Maturity Model – Capabilities, Technologies, and HITL Governance by Level*

The critical transition is from Level 3 to Level 4. At Level 3, organizations have enriched their SOC with threat intelligence and UEBA, but the analyst bottleneck remains — detection is better but response is still largely human-driven. Level 4 resolves this by introducing structured HITL governance, ML-driven triage, and SOAR automation as a coordinated architecture rather than isolated tools. This transition is not primarily a technology investment — it is a governance and process redesign that the framework described in this paper directly enables. Level 5 remains an aspirational future state for most organizations; the immediate objective for most enterprises is advancing from Level 2-3 to Level 4, where the measurable MTTD, MTTR, and alert fatigue improvements are realized.

Notably, HITL governance evolves across the maturity levels — from absent at Level 1, to informal analyst discretion at Level 2-3, to a structured and proportional framework at Level 4, and finally to strategic-only oversight at Level 5. This progression reflects the core argument of this paper: as AI capability increases, HITL does not diminish — it becomes more precisely targeted.

f. Controlled Validation Scenario

To validate the practical impact of the proposed AI-augmented SOC framework, a controlled scenario-based evaluation was conducted using synthetic telemetry patterns aligned with MITRE ATT&CK techniques and published industry benchmarks. While based on simulated data, the evaluation is designed to demonstrate the relative performance improvements achievable through the proposed architecture across three SOC configurations.

A simulated SOC environment was constructed with the following characteristics:

- **Approximately 10,000 alerts per day, reflecting a mid-to-large enterprise SOC environment**
- Mixed alert distribution across phishing, lateral movement, credential abuse, insider activity, and ransomware precursor behaviors
- Baseline comparison across three configurations: Traditional Rule-Based SOC, AI-Augmented SOC without HITL, and AI-Augmented SOC with HITL governance

Table 4 summarizes the observed outcomes across the three configurations:

Table 4. *Observed outcomes across the three configurations*

Performance Metric	Traditional SOC	AI-Augmented (No HITL)	AI-Augmented + HITL	Improvement vs. Baseline
Daily analyst-reviewed alerts	~10,000	~3,000–4,000	~3,000–4,000	~60–70% reduction
False positive rate	>50%	~20–25%	~12–18%	~50–65% reduction with HITL
MTTD (behavioral threats)	48–72 hours	8–16 hours	6–12 hours	~85% reduction
MTTR (automated containment)	24–48 hours	6–10 hours	4–8 hours	~83% reduction
High-impact action accuracy	Analyst-only	AI-autonomous (risk of error)	HITL-validated	Significantly improved
Model drift / bias detection	None	Undetected	HITL feedback loop	Continuously corrected

Note: The baseline MTTD of 48–72 hours reflect detection within an actively monitored simulated environment. The industry benchmark of 197 days [1] reflects real-world dwell time including periods without active monitoring. Synthetic results align directionally with IBM Security [1], Gartner [3], and Ponemon Institute [15] benchmarks.

The inclusion of HITL governance produced the most significant improvement in high-impact decision accuracy — configurations without HITL showed AI-autonomous containment actions that, while faster, carried higher risk of false-positive business disruption. The HITL-governed configuration ensured that critical containment steps such as account

suspension or system isolation were validated by analysts before execution, directly addressing the adversarial manipulation and hallucination risks described in Section II.b.

These results demonstrate how the proposed architecture operationalizes measurable improvements in SOC performance through the integration of AI capabilities, structured governance, and human oversight operating as a coordinated system rather than isolated components.

IV. Conclusion

The security threat landscape has evolved faster than most SOC architectures. Traditional rule-based models are not equipped for adversaries who use behavioral evasion, AI-assisted attacks, and living-off-the-land techniques. But deploying AI without governance is not the answer — bias, hallucination, lack of contextual confidence, and adversarial manipulation of AI models create new failure modes that can be equally damaging.

The scenario analysis in Figure 1 makes this clear — standalone architectures, whether rule-based or AI-driven, each carry structural blind spots that only a converged model resolves. Only the convergence of both — unified under Standard and AI-era frameworks including OWASP MCP Top 10, governed by HITL at every critical decision point — achieves the comprehensive detection and response capability enterprise security requires.

This paper presents the convergence framework through three pillars of Asset Visibility, Detection Engineering, and Accelerated Remediation; a dual-framework mapping that closes the AI governance gap; a quantified detection coverage analysis; a HITL governance framework defining the precise boundary between AI automation and human oversight; and a Control Self-Assessment model extending accountability to asset owners. The controlled validation scenario in Section III.f demonstrates that measurable improvements in MTTD, MTTR, and alert reduction are achievable when the proposed architecture is applied in a realistic SOC environment. Organizations that build both dimensions — AI capability and HITL governance — will have security operations that are faster, more accurate, auditable, and genuinely more resilient against the full spectrum of threats that define the current and emerging landscape. Future work will focus on validating the proposed framework in live enterprise SOC environments using production telemetry datasets, with the aim of establishing empirically grounded performance baselines and refining HITL governance thresholds under real operational conditions.

References

[1] IBM Security. (2023). Cost of a Data Breach Report 2023. IBM Corporation. https://www.ibm.com/reports/data-breach

[2] SANS Institute. (2023). 2023 SOC Survey: Evolving the Analyst Role. SANS Reading Room. https://www.sans.org

[3] Gartner. (2023). Market Guide for Managed Detection and Response Services. Gartner Research.

[4] MITRE Corporation. (2023). ATT&CK Framework: A Knowledge Base of Adversary Tactics and Techniques. https://attack.mitre.org/

[5] NIST. (2025). Computer Security Incident Handling Guide (SP 800-61 Rev. 3). National Institute of Standards and Technology.

[6] CISA. (2023). Cybersecurity Best Practices. Cybersecurity and Infrastructure Security Agency. https://www.cisa.gov/cybersecurity-best-practices

[7] Binbeshr, F., Imam, M., Ghaleb, M., Hamdan, M., Rahim, M. A., & Hammoudeh, M. (2023). The Rise of Cognitive SOCs: A Systematic Literature Review on AI Approaches. IEEE Access, 11.

[8] Ali, N., & Wallace, G. (2023). The Future of SOC Operations: Autonomous Cyber Defense with AI and Machine Learning. Journal of Cybersecurity Research.

[9] MITRE Corporation. (2023). ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. https://atlas.mitre.org/

[10] OWASP. (2023). OWASP Top 10 for Large Language Model Applications. OWASP Foundation. https://owasp.org/www-project-top-10-for-large-language-model-applications/

[11] NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), AI 100-1. National Institute of Standards and Technology.

[12] OWASP. (2025). OWASP Top 10 for Model Context Protocol (MCP). OWASP Foundation. https://owasp.org/www-project-top-10-for-model-context-protocol/

[13] Sharma, A., & Spunda, R. (2023). SOC Optimization Through AI-Powered Automation and Blockchain Integration. IEEE Security & Privacy.
[14] Gill, S., & Noah, A. (2023). AI-Enhanced SOC Operations: From Threat Intelligence to Automated Mitigation. International Journal of Information Security.
[15] Ponemon Institute. (2023). The Economics of Security Operations Centers. Ponemon Institute Research.