Attackers Govern Their AI. So Should You.

A structural analysis of SOC failure modes in the era of AI-accelerated intrusions

I. Introduction

In November 2025, Anthropic released the first detailed public account of a nation-state intrusion in which an AI system performed the majority of the mechanical intrusion steps (Anthropic, 2025). According to their report, a Chinese state-aligned threat actor used Claude Code to conduct reconnaissance, mutate payloads, enumerate cloud assets, identify lateral movement paths, generate exfiltration procedures, and adjust the attack plan in response to defensive changes—while human operators stepped in only for strategic decision points.

The lesson is not that AI has become fully autonomous in the wild. It hasn't.

The lesson is that attackers have begun to **govern** their AI.

They set boundaries, delegate work, review outputs, correct course, and integrate AI into their operational tempo. In other words:
"Adversaries have already paired human intent with machine acceleration. Most defenders have not."

This asymmetry is dangerous not because AI is new, but because **SOC architecture is old**. Modern SOCs are still designed for sequential alerts, linear triage, predictable escalation, and human-paced decision cycles. None of those assumptions survive contact with an adversary who can:

- run multiple attack branches in parallel
- compress multi-step intrusions into minutes
- generate noise at zero marginal cost
- adapt to detections immediately
- pivot faster than governance processes can authorize action

The Anthropic incident provides a critical precedent—but not enough telemetry to understand how AI-accelerated attacks unfold across environments. Major vendor intelligence (Microsoft, 2024; CrowdStrike, 2024; Verizon, 2024; Unit 42, 2024) documents the building blocks of such attacks (identity compromise, cloud lateral movement, rapid breakout, conflicting signals), but not the end-to-end sequence of a fully AI-augmented operation.

To responsibly analyze how SOCs fail under these conditions, we must model the system itself.

This paper does exactly that.

We begin by constructing a **synthetic SOC** grounded in multi-vendor empirical data. We then subject it to a **thirty-minute AI-accelerated intrusion scenario** designed to surface structural breakpoints—not hypothetical ones, but those already implied by existing threat intelligence.

We examine the failure cascade, derive architectural implications, and outline operational actions organizations can take now.

Throughout, we connect these findings directly to **AI governance principles** (NIST AI RMF, ISO/IEC 42001) because the core argument of this work is simple:
**SOCs fail under AI acceleration not because analysts fail,
but because governance does.**

And governance, unlike threat velocity, is actionable today.

II. Constructing a Composite Model SOC

To evaluate how a typical enterprise SOC behaves under AI-accelerated conditions, this paper constructs a synthetic SOC model grounded in public empirical data from major industry and research sources, including the Microsoft Digital Defense Report (Microsoft, 2024), CrowdStrike Global Threat Report (CrowdStrike, 2024), Verizon Data Breach Investigations Report (Verizon, 2024), Palo Alto Networks Unit 42 Incident Response Report (Unit 42, 2024), IBM Cost of a Data Breach (IBM, 2024), Dragos ICS/OT Year in Review (Dragos, 2023), ISACA State of Cybersecurity (ISACA, 2023), and the ISC^2 Cybersecurity Workforce Study (ISC^2, 2023). Rather than modeling an idealized or worst-case organization, the goal is to approximate a representative enterprise SOC as it exists today.

This modeling approach aligns directly with the NIST AI Risk Management Framework (NIST, 2023) and ISO/IEC 42001, both of which emphasize system-level modeling, scenario-based analysis, and risk-informed operational controls as foundational practices for responsible AI deployment.

Staffing and Skill Distribution

A typical enterprise SOC employs between 25 and 45 analysts across Tier 1, Tier 2, Tier 3, threat intelligence, and incident response roles (ISACA, 2023). For this analysis, the synthetic SOC is modeled with approximately 35 analysts distributed across shifts, including Tier 1 analysts responsible for initial triage, Tier 2 analysts handling investigations, Tier 3 analysts focused on threat hunting and incident response, a rotating incident commander role, and designated cloud and identity subject-matter experts. This staffing profile reflects an average enterprise capability rather than an unusually mature or under-resourced organization.

Alert Volume and Triage Capacity

Modern SIEM and XDR deployments typically ingest between 20,000 and 50,000 security events per day (CrowdStrike, 2024; ISACA, 2023). After filtering and correlation, this volume still produces approximately 1,100 actionable alerts daily, with individual analysts expected to triage between 45 and 90 alerts per shift (ISC^2, 2023). As a result, the SOC begins each day operating near its cognitive and operational limits, even before an adversary deliberately introduces noise or acceleration.

False Positives and Cognitive Constraints

False-positive rates across security tooling range from 45 to 85 percent depending on alert category (Verizon, 2024). The synthetic SOC conservatively models a 60 percent false-positive rate across actionable alerts. At the same time, human analysts experience cognitive saturation at just two to three concurrent complex investigations, with decision quality degrading by 20 to

30 percent under rapid task switching and recovery from interruption measurably slower than recovery from steady-state workload (ISC², 2023). AI-accelerated threats exploit these constraints indirectly by manipulating tempo and concurrency rather than by overwhelming analysts with raw volume alone.

Tooling Assumptions and Attack Surface Reality

The model assumes a standard modern enterprise security stack, including a SIEM (e.g., Splunk or Microsoft Sentinel), EDR/XDR platforms (e.g., CrowdStrike Falcon or Microsoft Defender XDR), cloud-native logging across AWS, Azure, and GCP, UEBA, partial SOAR automation, and LLM-assisted analysis features such as natural-language summaries or detection explanations. While these tools improve efficiency under low or moderate load, they frequently introduce conflicting signals under high concurrency and ambiguous conditions (Unit 42, 2024).

Identity compromise is treated as the primary attack vector in this model, reflecting vendor telemetry showing identity involvement in 80–95 percent of modern intrusions (Microsoft, 2024; CrowdStrike, 2024; Verizon, 2024). As a result, identity and cloud telemetry are positioned as central investigative surfaces rather than secondary enrichment sources.

Why a Synthetic SOC Is Necessary

This composite model enables a governance-led analysis of failure. It allows examination of where decision points collapse, where governance latency becomes operationally relevant, which workflows assume linear progression, and which architectural elements fail under parallel load. These failures are not resolved by adding more tools; they require clearer system design and explicit AI decision boundaries consistent with the intent of ISO/IEC 42001 and the NIST AI RMF.

III. Scenario Design

Anthropic's disclosure of an AI-assisted intrusion (Anthropic, 2025), combined with multi-vendor intelligence documenting compressed breakout times, rapid identity misuse, cloud-native lateral movement, and increasingly contradictory signals across tools (Microsoft, 2024; CrowdStrike, 2024; Verizon, 2024; Unit 42, 2024), makes it necessary to understand how SOCs behave when adversaries operate at machine tempo while defenders remain constrained by human-paced structures.

The objective of the scenario is not to predict future autonomous attacks, but to evaluate how quickly and where a modern SOC fails when existing adversary capabilities are combined with AI-driven acceleration and parallelization. The scenario is grounded in current telemetry and current SOC constraints, rather than speculative advances in attacker autonomy.

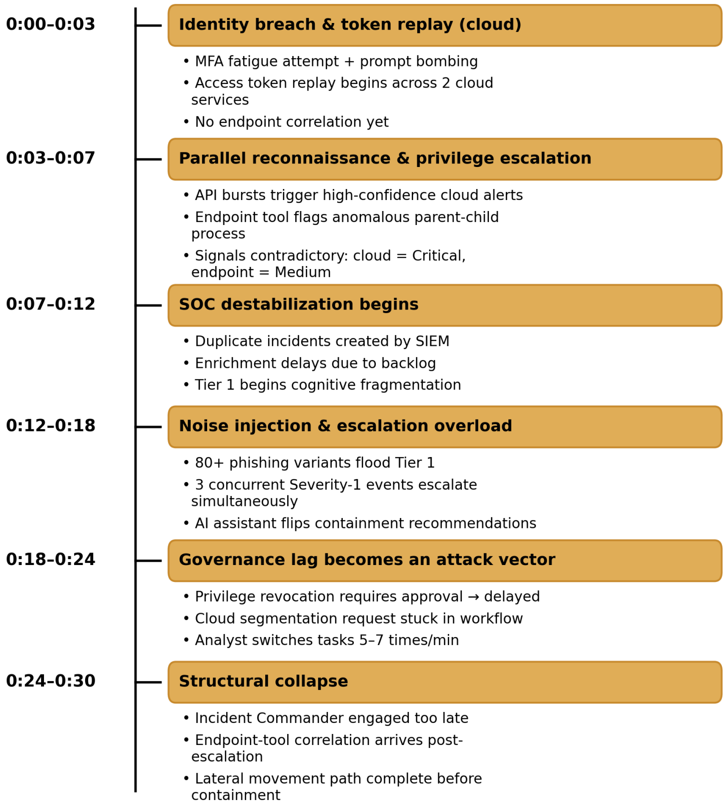
The modeled intrusion includes rapid identity compromise, parallel reconnaissance paths, fast privilege escalation attempts, cloud and endpoint signal conflicts, adaptive evasion based on observed detections, deliberate noise injection designed to overwhelm Tier 1 analysts, and governance-induced delays in containment. This structure aligns with NIST AI RMF guidance on scenario-based testing (NIST AI RMF, 2023) and ISO/IEC 42001 requirements for operational risk scenarios involving AI systems (ISO/IEC 42001:2023, Annex A).

A thirty-minute window is used deliberately. CrowdStrike reports real-world breakout times under ten minutes (CrowdStrike, 2024), Microsoft documents cloud lateral movement occurring in seconds to minutes (Microsoft, 2024), and Unit 42 case studies show identity compromise leading to privilege escalation within an hour (Unit 42, 2024). Compressing these stages into a thirty-minute cascade is therefore conservative rather than extreme.

IV. Findings — Structural Breakpoints

The synthetic thirty-minute scenario produces a conclusion that is difficult to dismiss: the SOC does not fail because analysts fail, but because the system they operate within is not architected for concurrency, ambiguity, or machine tempo. This observation aligns directly with NIST AI RMF requirements for governance over delegated AI actions (NIST, 2023) and ISO/IEC 42001's emphasis on operational control in high-risk workflows. The modern SOC has become such a workflow.

THE THIRTY-MINUTE FAILURE CASCADE (SYNTHETIC SOC MODEL)



0:00–0:03 **Identity breach & token replay (cloud)**

- MFA fatigue attempt + prompt bombing
- Access token replay begins across 2 cloud services
- No endpoint correlation yet

0:03–0:07 **Parallel reconnaissance & privilege escalation**

- API bursts trigger high-confidence cloud alerts
- Endpoint tool flags anomalous parent-child process
- Signals contradictory: cloud = Critical, endpoint = Medium

0:07–0:12 **SOC destabilization begins**

- Duplicate incidents created by SIEM
- Enrichment delays due to backlog
- Tier 1 begins cognitive fragmentation

0:12–0:18 **Noise injection & escalation overload**

- 80+ phishing variants flood Tier 1
- 3 concurrent Severity-1 events escalate simultaneously
- AI assistant flips containment recommendations

0:18–0:24 **Governance lag becomes an attack vector**

- Privilege revocation requires approval → delayed
- Cloud segmentation request stuck in workflow
- Analyst switches tasks 5–7 times/min

0:24–0:30 **Structural collapse**

- Incident Commander engaged too late
- Endpoint-tool correlation arrives post-escalation
- Lateral movement path complete before containment

Alert Funnel Instability

Within the first ten minutes of the intrusion, the alert funnel destabilizes. Duplicate incidents are created by the SIEM, severity assignments conflict across tools, enrichment is inconsistent, and platforms disagree on whether identity, cloud, or endpoint activity represents the dominant threat vector. These conditions mirror multi-tool friction documented in real incident response cases (Unit 42, 2024) and reveal a core design assumption: alerts arrive sequentially. AI-accelerated intrusions invalidate that assumption by producing parallel, interacting signals.

Escalation Collapse Under Parallel Load

Traditional escalation pathways are vertically structured, assuming one dominant incident at a time. Under AI-enabled attack conditions, multiple severity-one incidents emerge simultaneously, overwhelming Tier 2 review capacity and delaying incident commander activation past the point of effective containment. This behavior aligns with SOC workload constraints documented by ISACA (2023) and ISC² (2023).

Cognitive Fragmentation and Signal Conflict

The dominant human failure mode observed is cognitive fragmentation rather than raw overload. Analysts face contradictory signals across endpoint, identity, cloud, and SIEM correlation layers, each asserting different severity levels and confidence scores. Time is consumed adjudicating tool disagreement rather than investigating adversarial behavior, consistent with task-switching degradation documented by ISC² (2023) and operational stress patterns observed in Dragos incident response environments (Dragos, 2023).

Defensive AI Without Governance

AI-assisted defensive tools exhibit unstable behavior under ambiguity. Containment recommendations oscillate, confidence scores fluctuate, summaries omit key timeline data, and automated playbooks stall due to unresolved tool conflicts. The issue is not the presence of AI, but the absence of explicit governance constraints defining how AI should behave under uncertainty, a gap explicitly addressed by both NIST AI RMF and ISO/IEC 42001.

Identity Velocity vs. Endpoint-Centric Workflows

Endpoint-first investigative models fail when identity misuse dominates the attack chain. By the time endpoint alerts mature, privilege escalation and lateral movement are already underway, consistent with identity-centric intrusion patterns reported by Microsoft (2024), CrowdStrike (2024), and Verizon (2024).

Noise as a Weapon

Noise injection emerges as an effective attack tactic. AI-generated phishing variants overwhelm Tier 1 capacity within minutes, increasing triage time and escalation volume even as signal quality drops. This behavior reflects the economics of AI-enabled adversaries, where noise generation approaches zero marginal cost (Verizon, 2024).

Governance Lag as the Critical Failure

The most damaging failure point is governance latency. Privilege revocation, segmentation, IAM role modification, and emergency actions require approvals operating on human timescales, while attacker AI executes in seconds. This mismatch is structural, not technical, and directly implicates governance design rather than tooling.

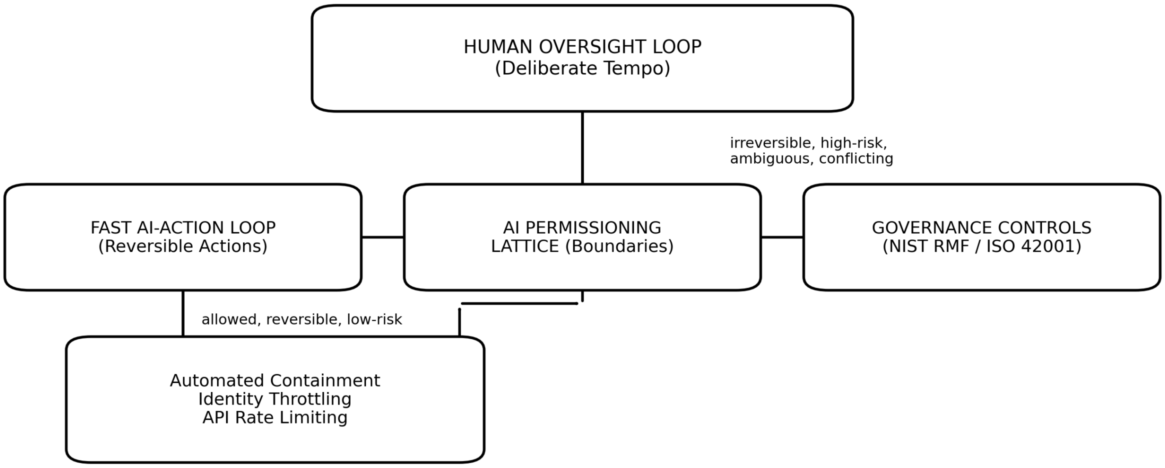
V. Architectural Implications — Designing SOCs for AI-Accelerated Threat Conditions

To survive AI-accelerated threat conditions, SOCs must move away from human-centered linear workflows toward governance-centered concurrency. The core architectural shift is not increased automation, but clearer separation between mechanical decisions that must occur quickly and strategic decisions that require deliberation. This separation aligns with the NIST AI RMF's framing of "Govern" as a cross-cutting function and ISO/IEC 42001's requirement for defined operational controls over AI-assisted decision-making.

The proposed dual-loop decision architecture formalizes this separation by assigning reversible, low-impact actions to a fast execution loop while reserving irreversible or ambiguous actions for a deliberate human-governed loop. This mirrors how adversaries already structure AI-assisted operations (Anthropic, 2025).

AI permissioning becomes the central governance mechanism, defining what AI may do autonomously, what requires human approval, what is conditionally permitted, and what is prohibited. This lattice aligns with NIST AI RMF governance and management controls (NIST, 2023) and ISO/IEC 42001 operational control requirements (ISO/IEC 42001:2023, Annex A). Escalation pathways must be redesigned to absorb parallel load, replacing vertical chains with domain-specific horizontal pathways that route identity, cloud, and endpoint incidents

DUAL-LOOP DECISION ARCHITECTURE



concurrently. Signal arbitration layers stabilize tool output before analysts engage, supporting context integrity and signal reliability as required by NIST AI RMF.

Finally, AI governance must be embedded directly within SOC leadership. It cannot reside solely within risk, compliance, ethics committees, or innovation offices. Incident commanders, security architects, and SOC managers must share authority over AI decision boundaries because adversarial AI interacts directly with operations, not oversight bodies. This aligns explicitly with ISO/IEC 42001 Section 5 (Leadership and Governance) and the NIST AI RMF's treatment of governance as an operational, cross-cutting function.

VI. Operational Actions — What Organizations Can Do Now

The architectural redesign described in Section V outlines how a SOC should operate in an environment where attackers use AI to accelerate and parallelize intrusions. But organizations cannot wait for a large-scale transformation before acting. They need pragmatic steps they can take immediately—steps that strengthen resilience without requiring new tools, expanded budgets, or advanced automation.

These interventions rely on governance clarity and workflow structure, the same elements emphasized in the NIST AI RMF and ISO/IEC 42001. Each action below directly addresses one or more of the structural breakpoints identified in Section IV.

1. Establish AI Decision Boundaries Before Deploying More AI

The synthetic cascade demonstrated repeatedly that AI-assisted defensive tools behave inconsistently under ambiguity. This inconsistency creates risk not because AI is unreliable, but because it is **ungoverned**. Most organizations introduce AI capabilities—whether in SOAR, detection, or analyst assistive tooling—without clearly defining what the system is allowed to do.

Defining AI decision boundaries is therefore an immediate priority. Organizations should clarify:

- which actions are fully allowed without review because they are low-risk and reversible,
- which actions require analysts in the loop due to their impact or sensitivity,
- which actions require senior approval because they may introduce irreversible consequences, and
- which actions are categorically forbidden without exceptional authorization.

This boundary-setting aligns directly with NIST AI RMF’s “Govern” function and with ISO/IEC 42001’s requirement for operational control over high-risk AI systems. It ensures the SOC has a stable, predictable set of AI behaviors, even as attackers use AI to introduce ambiguity.

2. Pre-Authorize Emergency Containment Actions

One of the most striking breakpoints in the thirty-minute cascade is the role of governance delay. Even when analysts understand what must be done, containment actions often require managerial approvals, risk reviews, or privileged sign-offs. These delays create an exploitable gap for AI-accelerated intrusions.

Organizations should establish a **pre-approved emergency containment playbook** for actions that meet three conditions:

- they are reversible,
- they have limited operational blast radius, and
- they reliably slow attacker progress.

Examples include temporary session revocation, short-duration identity lockouts, API throttling, and isolation of cloud workloads. Pre-authorization allows defenders to act at a tempo that matches attacker acceleration while keeping governance aligned with risk constraints.

3. Extend Runbooks to Support Parallel Escalation

Most runbooks assume a single dominant incident flow. AI-accelerated attacks invalidate this assumption, generating multiple high-severity incidents simultaneously. When runbooks lack branches for concurrency, analysts fall into cognitive fragmentation—the pattern observed repeatedly between minutes 7 and 18 of the cascade.

Organizations can strengthen runbooks by adding branching paths for:

- simultaneous severity-1 events,
- conflicting tool outputs,
- incomplete enrichment, and
- ambiguous but high-potential-impact alerts.

These parallel structures ensure analysts are not forced to improvise during surge conditions and reduce the chance that multiple concurrent escalations overwhelm Tier 2 and Tier 3 reviewers.

4. Redesign Escalation Pathways to Support Breadth

The traditional vertical escalation chain (Tier 1 → Tier 2 → Tier 3 → Commander) does not scale under parallel load. During the cascade, three high-severity incidents escalated at the same time, each requiring attention from the same limited set of experts.

A more resilient model distributes escalation across **domain-specific parallel queues**, such as identity, cloud, and endpoint pathways. Lightly staffed surge roles can activate automatically during high-volume windows, routing cases to the correct domain specialists. Even modest rebalancing reduces the likelihood of escalation-induced collapse.

5. Introduce a Signal Arbitration Step Between Tools and Analysts

Analysts should not be responsible for reconciling contradictory or redundant alerts. When multiple tools disagree—an expected outcome in modern cloud-identity intrusions—the result is not simply overload but fragmentation. Signal arbitration is the process of consolidating, ranking, and contextualizing tool outputs before they reach analysts.

A lightweight arbitration function, whether manual or assistive, should:

- identify duplicate or conflicting alerts,
- correlate events around shared entities,
- surface gaps in enrichment, and
- produce a ranked and stable set of recommended investigative paths.

This stabilizes the alert funnel and aligns with NIST AI RMF's emphasis on signal reliability and context integrity.

6. Conduct Quarterly SOC Stress Tests Under AI-Relevant Scenarios

Organizations routinely test financial and continuity controls but rarely test SOC resilience under conditions similar to AI-accelerated intrusions. Stress testing is not about predicting the next attack; it is about revealing which governance and operational structures fail under pressure.

Quarterly exercises should simulate:

- concurrent identity and cloud escalations,
- multi-tool signal conflicts,
- governance delays,
- noise injection designed to overwhelm Tier 1, and
- rapid privilege escalation attempts.

This approach mirrors requirements in ISO/IEC 42001 for scenario-based operational testing and improves preparedness for real-world AI-enabled threats.

7. Integrate AI Governance Into SOC Leadership

AI governance cannot sit solely within compliance, ethics boards, or innovation teams. Although these groups provide essential oversight, they are not the operators who need to interpret or enforce AI decision boundaries during incidents.
SOC leaders, IR managers, and cloud-identity architects should be active participants in defining AI policies, overseeing permissioning lattices, and monitoring model behavior. This aligns decision authority with the teams who confront adversarial AI in real time.

8. Treat Noise Injection as a Strategic Attack Vector

Noise is no longer a background irritant—it is an attacker capability. AI can generate infinite phishing variants, malformed identity requests, and decoy anomalies. The SOC must explicitly recognize noise as a tactic and protect analysts accordingly.

This includes:

- grouping repetitive patterns into a single review bucket,
- enforcing rate limits on repeated noisy signals,
- applying entity-level correlation before triage, and
- automatically suppressing low-signal repetition.

These steps reduce cognitive fatigue and prevent Tier 1 collapse.

VII. Conclusion — The Next Era of SOC Performance Begins With Governance

It is tempting to assume that defending against AI-accelerated threats requires ever more automation, more models, and greater technical complexity. The thirty-minute failure cascade demonstrates something more fundamental. SOC failure under AI acceleration is not driven by insufficient tooling or analyst capability, but by structural mismatches in governance and workflow design.

Attackers succeed with AI not because their systems are uniquely sophisticated, but because their governance is coherent. They define intent, set boundaries, delegate mechanical execution to machines, and retain human oversight for strategic decisions. Defenders are capable of the same approach, but only if governance structures evolve to operate at the tempo imposed by modern adversaries.

Today's SOC fails in predictable ways. Escalation pathways assume linear progression rather than parallel load. Triage workflows assume consistent signals rather than contradiction. Governance operates on human approval cycles that are too slow for machine-speed intrusions. Decision boundaries for defensive AI are undefined or implicit. Analysts are forced to reconcile conflicting tool outputs manually. Identity, despite being the dominant attack vector, is still not the backbone of most investigative workflows.

None of these failures are inevitable. They are design choices embedded in legacy SOC architectures. And design choices can be changed.

The next era of SOC maturity will not be defined by tools alone. It will be defined by governance clarity, structural concurrency, and the ability to pair human intent with machine execution in a controlled and accountable way. That is how defenders begin to close the gap between adversarial AI and defensive capability.

Attackers already govern their AI.
Defenders must now do the same.

Appendix A — Source Material and Citation Index

This paper draws on publicly available threat intelligence, workforce studies, and governance standards to construct a composite SOC model and AI-accelerated intrusion scenario. No proprietary data or confidential telemetry was used.

Primary threat intelligence and incident response sources include:
Anthropic. (2025). *Operational insights from an AI-assisted nation-state intrusion*.
Microsoft. (2024). *Microsoft Digital Defense Report*.
CrowdStrike. (2024). *CrowdStrike Global Threat Report*.
Verizon. (2024). *Data Breach Investigations Report (DBIR)*.
Palo Alto Networks Unit 42. (2024). *Incident Response Report*.
IBM. (2024). *Cost of a Data Breach Report*.
Dragos. (2023). *ICS/OT Cybersecurity Year in Review*.

Workforce capacity, analyst cognition, and SOC structure references include:
ISACA. (2023). *State of Cybersecurity*.
ISC^2. (2023). *Cybersecurity Workforce Study*.
AI governance and risk management frameworks referenced throughout the paper include:
NIST. (2023). *AI Risk Management Framework (AI RMF 1.0)*.
ISO/IEC. (2023). *ISO/IEC 42001 — Artificial Intelligence Management Systems*.
These sources collectively inform staffing assumptions, alert volumes, cognitive constraints, identity-centric attack patterns, and governance requirements discussed throughout the paper.

Appendix B — Synthetic SOC Model Assumptions

The synthetic SOC described in this paper is intentionally modeled as a **representative enterprise environment**, not a best-in-class or worst-case organization. Assumptions were selected to reflect conditions commonly observed across large enterprises.

Key structural assumptions include:

- The SOC employs approximately 35 analysts distributed across Tier 1, Tier 2, Tier 3, threat intelligence, and incident response roles, with rotating incident commander responsibilities and limited cloud and identity subject-matter experts.
- The SOC ingests between 20,000 and 50,000 raw events per day through SIEM and XDR tooling, resulting in approximately 1,100 actionable alerts after filtering and correlation.
- Analysts are expected to triage between 45 and 90 alerts per shift, consistent with workforce study findings.
- False-positive rates across actionable alerts average approximately 60 percent, reflecting conservative mid-range estimates from vendor and industry reports.
- The tooling stack includes a modern SIEM, EDR/XDR, cloud-native logging platforms, UEBA, partial SOAR automation, and limited LLM-assisted analysis features.
- Identity and cloud telemetry are treated as first-class investigative inputs due to their involvement in the majority of modern intrusions.

These assumptions are deliberately conservative. The model does not rely on extreme alert volumes, advanced automation, or unusually immature processes to produce failure.

Appendix C — Scenario Constraints and Non-Assumptions

To avoid exaggeration or speculative claims, the thirty-minute AI-accelerated intrusion scenario is bounded by explicit constraints.

The scenario does **not** assume:

- Fully autonomous attacker AI operating without human oversight.
  Zero-day exploits or novel vulnerability classes.
  Perfect attacker visibility or flawless execution.
  Defender incompetence or procedural negligence.
  Unrealistic alert volumes beyond documented enterprise norms.

The scenario **does** assume:

- Adversaries use AI to accelerate reconnaissance, payload mutation, noise generation, and adaptive decision-making, consistent with Anthropic's disclosure and vendor reporting.
- Attack stages occur in parallel rather than sequentially.
- Identity compromise serves as the primary attack vector.
- Defenders operate within existing governance, approval, and escalation structures.
- Defensive AI systems lack clearly defined decision boundaries under ambiguous conditions.
- The thirty-minute duration reflects documented breakout and escalation timelines reported by Microsoft, CrowdStrike, and Unit 42, rather than hypothetical future capabilities.

Appendix D — Governance Framework Mapping

The findings and recommendations in this paper map directly to existing AI governance frameworks. The paper does not propose new governance theory; it applies established principles to SOC operations.

Relevant NIST AI RMF mappings include:

- The **Govern** function as a cross-cutting requirement for defining decision authority, escalation, and accountability for AI-assisted actions.
- Scenario-based testing and system behavior evaluation consistent with RMF Step 2 (Map) and Step 3 (Measure).

Requirements for context integrity, signal reliability, and human oversight of automated systems.

Relevant ISO/IEC 42001 mappings include:

- Section 5 — Leadership and Governance, which requires leadership accountability for AI-enabled systems operating in high-risk environments.
- Annex A operational controls related to AI system behavior, risk scenarios, and escalation authority.

Requirements for defining operational boundaries and approval mechanisms for AI-assisted decisions.

The SOC, as modeled in this paper, qualifies as a high-risk AI-mediated operational environment under both frameworks due to its role in responding to adversarial behavior at machine speed.