‘Guardrails First’: Designing Safe, Scalable, and Accountable AI Agents for Security Operations

By : Neal Iyer

Conf42 DevOps 2026

Why Talk Security Operations with a DevOps audience?

Your SecOps colleagues will seek your help to build, deploy AI Agents THIS year!

What qualifies me as a presenter?

- **5+ years building market leading Security Operations Products**
  - Proofpoint
  - Zscaler
  - Splunk
  - Cisco
- **Talk to 150+ SOC teams every year across all shapes and sizes**
  - Fortune 500 organizations with large SOC teams
  - Organizations starting a SOC from scratch for the first time

The SOC Crisis: Overwhelmed and Under Pressure

Alert Fatigue

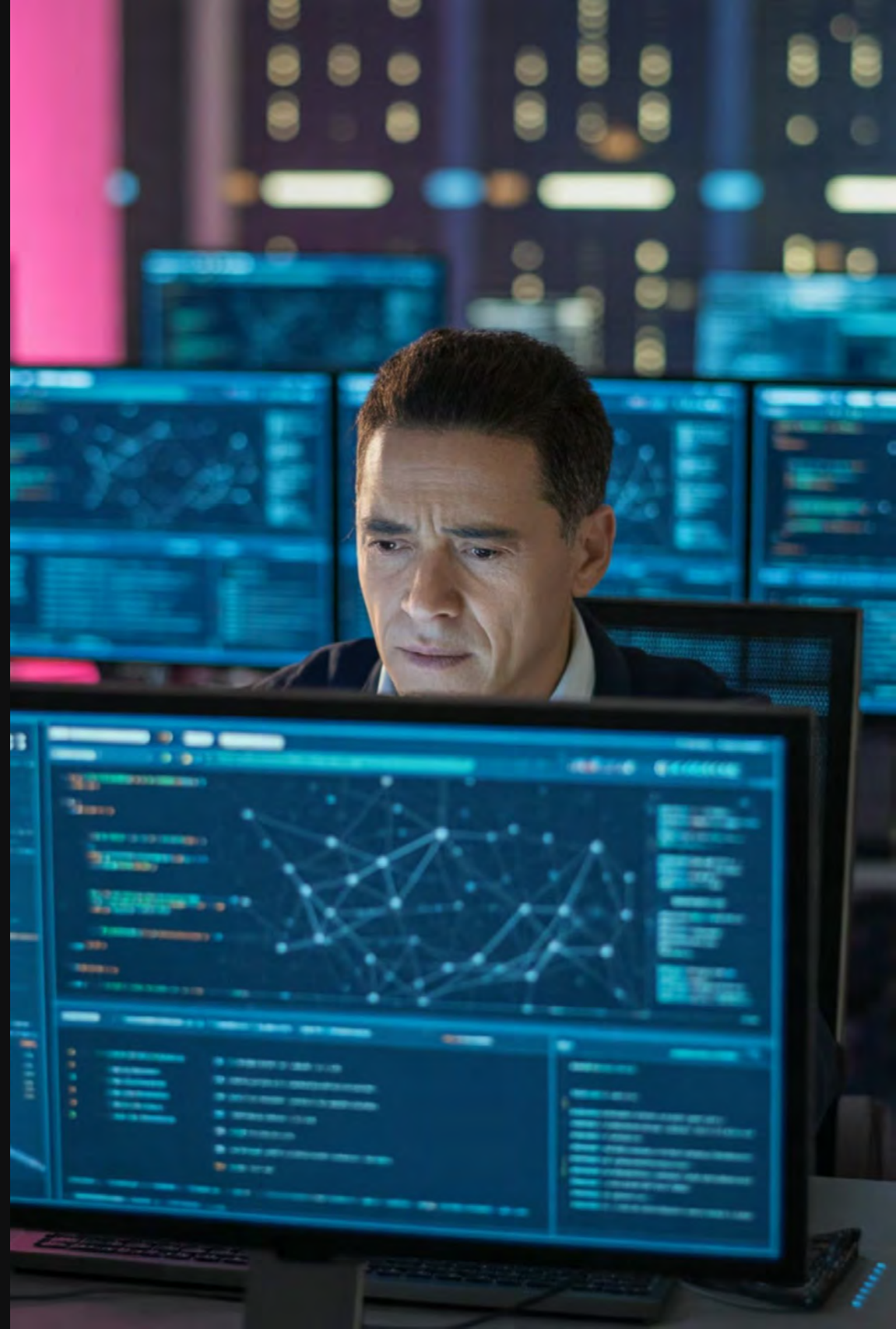
Security teams are drowning in thousands of daily alerts, with analysts spending 80% of their time on false positives and routine triage tasks.

Tool Sprawl

Modern SOCs juggle 15-20 different security tools, creating fragmented workflows and dangerous visibility gaps across the attack surface.

Skills Shortage

The cybersecurity workforce gap continues to widen, with experienced analysts burning out under relentless operational demands and escalating threats.



AI makes it even easier for attackers!

Speed

Increased speed of attacks with AI

Scale

Easier to scale attacks with AI

Skill

AI lowers the skill required to launch attacks

AI Agents to augment SOC analysts are an absolute necessity but very hard!

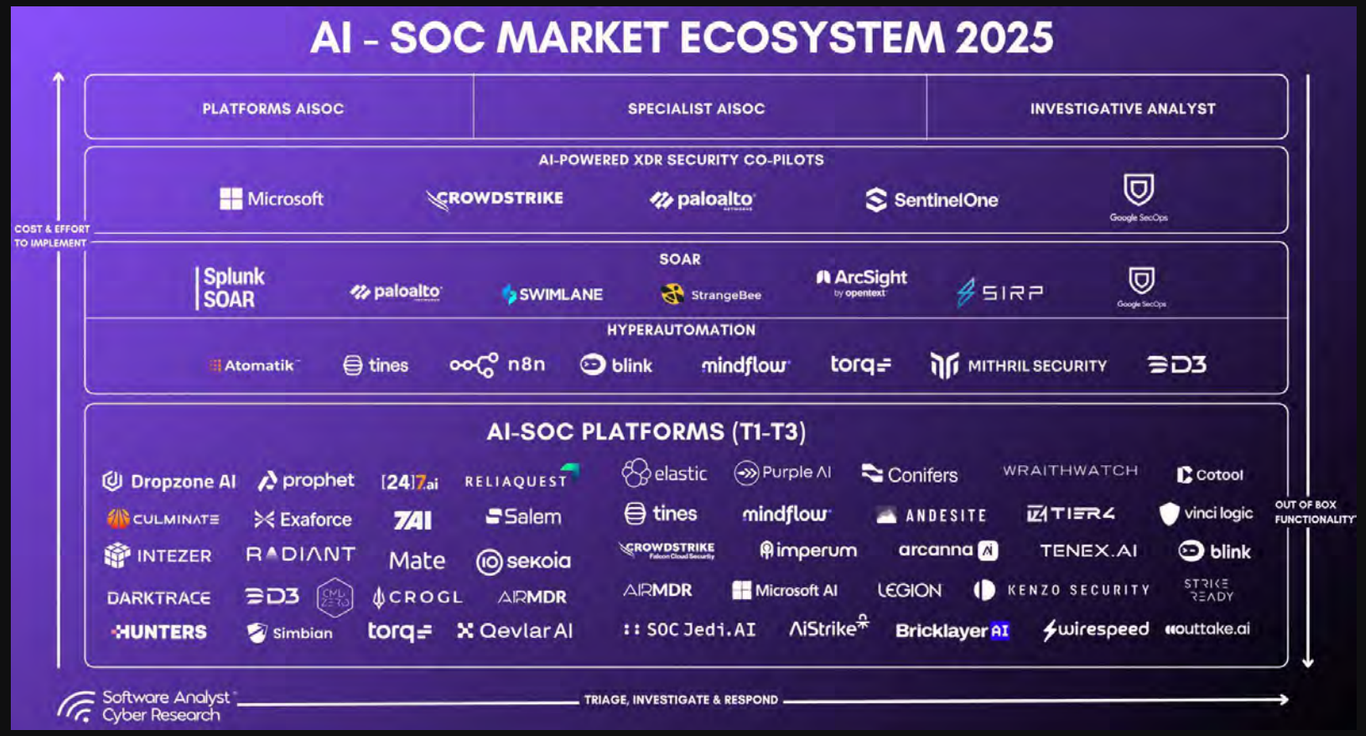
The Opportunity

AI agents offer unprecedented potential: faster triage, consistent decision-making, reduced analyst burnout, and 24/7 coverage that scales beyond human limitations.

The Risk

Yet the high stakes of cybersecurity demand caution. Unguarded AI can miss critical threats, create false positives at scale, or make decisions without adequate oversight.

Many choices available but many may not be suitable



Many vendors with similar marketing

Image credit: Francis Odum on LinkedIn

IC-led - Internal ‘build our own’ projects

Demos are great but how do we build enough trust to productionize?

Introducing the "Guardrails First" Framework

A practical design pattern for building trustworthy AI agents within SOC workflows. This framework transforms human-centric security processes into an AI learning curriculum grounded in real procedures, analyst heuristics, and verified historical casework.

- Organizational expertise at the center
- Appropriately constrained in-line with policy
- Designed with trust and auditability in mind
- Human in the Loop for critical actions

Building the AI Learning Curriculum

- **Leverage Historical Results, Notes**

  Similar alerts investigated by your SOC team are a golden corpus for leveraging for RAG, model Training

- **Utilize Analyst SOPs as context**

  Leverage existing organizational SOPs with rich feature extraction when prompting the AI Agents. Prepare for a significant rewrite/overhaul to account for gaps

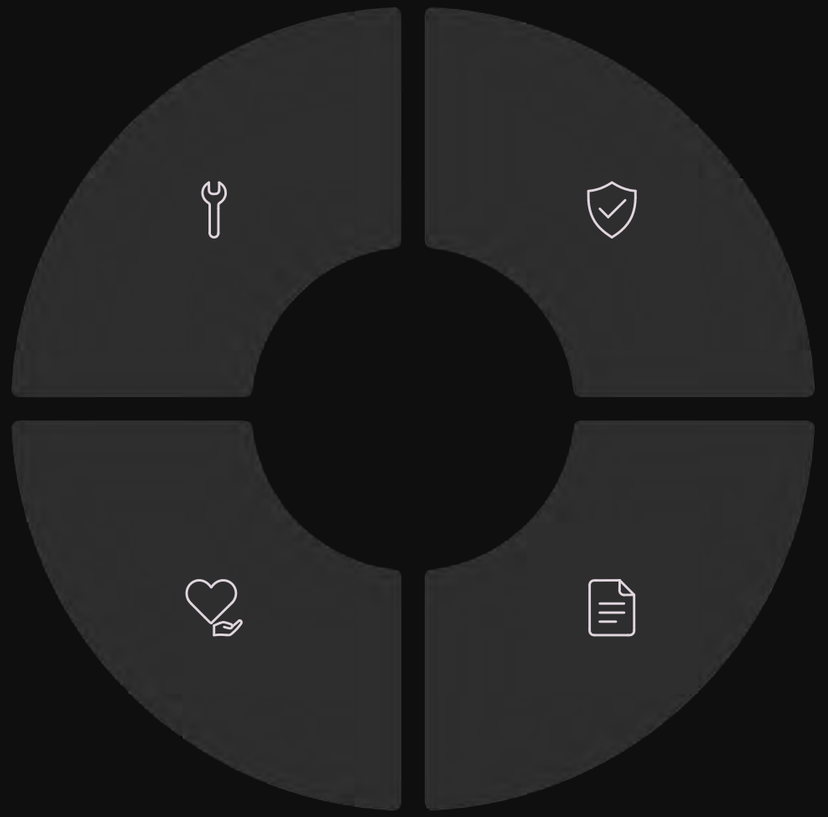
- **Build a back-testing pipeline**

  Test all versions against a historical set of alerts to ensure high efficacy outcomes across multiple runs

Core Pillars of Safe AI Agent Design

Capability-Scoped Tools

Policy-Driven Gating



Human Oversight

Decision Provenance

Capability-Scoped Tools: Constraining AI Actions

The Principle

Each AI tool exposed to your agents should perform specific well-defined functions within a specific security domain. No tool should have access beyond its intended scope.

Why It Matters

Limiting tool scope prevents cascade failures, contains potential errors, and makes auditing straightforward. You can disable a single capability without dismantling your entire AI system.

Alert Enrichment

Query threat intelligence feeds, gather context from SIEM data. Read-only access to specific databases.

Ticket Management

Create, update, and categorise incident tickets. Cannot close high-severity incidents without human approval.

User Lookup

Retrieve user account details from identity systems. Cannot modify permissions or disable accounts.

Notification Dispatch

Send alerts via email or Slack. Pre-approved recipient lists only, cannot escalate beyond defined contacts.

Policy-Driven Gating

- **AI Proposal**

  Agent analyses alert and recommends action based on training

- **Policy Check**

  Proposal evaluated against security policies and compliance rules

- **Gating Decision**

  Approved actions proceed; rejected proposals escalate to human analyst

- **Execution**

  Only validated actions execute, with full audit trail captured

Decision Provenance: Complete Traceability

Why Full Audit Trails Matter

In security operations, every decision must be defensible. Decision provenance captures the complete chain of reasoning behind each AI action.

- **What data did the AI observe?**

  Log all inputs: alerts, threat intel, historical patterns, and contextual signals

- **How did it reason?**

  Document the decision logic, confidence scores, and rule matches

- **What action did it take?**

  Record outputs, side effects, and any escalations to human analysts

- **Was the outcome correct?**

  Track validation, post-incident review, and continuous feedback loops

Human in the Loop Adoption Roadmap: From Observation to Autonomy

Shadow Mode

AI makes recommendations silently. No actions executed. Build confidence and calibrate accuracy.

Supervised Operation

AI proposes actions; analysts approve before execution. Gradually expand scope as trust builds and accuracy improves.

Bounded Autonomy

AI handles routine low-risk tasks independently. Human oversight reserved for high-severity or ambiguous cases.

Continuous Learning

AI refines behaviour based on analyst feedback. Regular audits ensure performance remains within acceptable parameters.

Measuring Success: Key Performance Indicators

- **Reduction in MTTA**

  Mean time to acknowledgement drops as AI handles initial triage and enrichment

- **False Positive Reduction**

  Better context and historical learning lead to more accurate alert classification

- **Analyst Productivity Gain**

  Analysts spend more time on complex investigations, less on routine tasks

- **Recommendation Accuracy**

  AI suggestions align with analyst decisions, demonstrating effective learning

Implementation Blueprint: Your Action Plan

01

Assess Current State

Document existing SOC workflows, pain points, and analyst decision patterns

02

Define Scope

Select initial use cases with clear success criteria and low risk profiles

03

Build Guardrails

Implement capability scoping, policy gates, and audit logging infrastructure

04

Train & Validate

Develop AI models using historical data, test in shadow mode, validate accuracy

05

Deploy Supervised

Launch with human oversight, gather feedback, refine based on real-world usage

06

Expand Gradually

Increase autonomy incrementally, measure continuously, maintain rollback capability

The Future is Safe, Scalable, and Human-Centred

> "AI in security operations isn't about replacing analysts—it's about amplifying their expertise whilst maintaining accountability and trust. Guardrails first ensures we build systems worthy of the critical mission they serve."

Start Small, Think Big

Begin with low-risk, high-value use cases. Prove value before expanding scope.

Measure Everything

Establish baselines, track KPIs, and validate improvements with data, not assumptions.

Keep Humans in the Loop

Augment analyst capabilities; never eliminate human judgment from critical security decisions.

Build Trust Through Transparency

Decision provenance and clear audit trails are non-negotiable for responsible AI deployment.

Neal Iyer | Conf42 DevOps 2026

Thank You!
Questions?
Welcome.