

When Marketing Fails

*AI SOC and the curious gap
between vendor claims and user
experience.*



Anton Chuvakin & Oliver Rochford

When Marketing Fails: AI SOC and the curious gap between vendor claims and user experience.

Anton Chuvakin & Oliver Rochford

Over the past 18 months, AI-driven SOC platforms, particularly those built around large language models, have been marketed as a step-change in security operations. Vendors promise autonomous investigations, dramatic analyst productivity gains, and rapid paths to “SOCless” or even humanless operations. Yet across dozens of briefings with both vendors and buyers, a striking and persistent disconnect emerges between how these systems are positioned and how they are actually experienced by practitioners in production.

This paper is based on qualitative analysis drawn from 30+ vendor interactions, OSINT, and direct conversations with CISOs, SOC leaders, and detection engineers actively evaluating or deploying AI SOC capabilities.

We document a recurring pattern: rather than surfacing known limitations in external messaging, vendors reframe adoption friction and accuracy shortfalls as buyer “AI readiness” problems, change resistance, or trust deficits.

This dynamic closely mirrors the social psychology described in [When Prophecy Fails](#), seminal study on cults and failed predictions, where deeply held beliefs persist even when confronted with contradictory evidence. Rather than revising the underlying belief (“the prophecy was wrong”), adherents reinterpret reality to preserve it (“the prophecy was misunderstood,” conditions weren’t right,” or “outsiders caused the failure”). In the AI SOC context, unmet expectations are similarly rationalized: low adoption becomes a change-management and change capacity problem, inconsistent results become a trust deficit, and technical limitations are reframed as buyer readiness issues. The narrative is preserved, while the evidence is reinterpreted.

We argue that this “prophecy-driven AI marketing” is not merely a marketing excess problem, but a structural failure of feedback integration between engineering reality, customer experience, and go-to-market narratives. When technological immaturity and adoption friction are explained away rather than examined, vendors optimize for narrative velocity instead of operational value.

We also argue that this represents a form of narrative-led gaslighting that damages trust across the ecosystem. It shifts accountability away from technology readiness and toward buyer psychology, discouraging the feedback loops necessary for genuine product maturation. Over time, this risks eroding confidence not only in individual vendors, but in the AI SOC category as a whole.

The paper concludes by outlining a more defensible framework for evaluating AI SOC claims: distinguishing assistive from autonomous capabilities, measuring adoption depth rather than feature usage, and treating trust as an emergent property of reliability, not a buyer mindset issue. As AI becomes embedded into core security operations, the cost of narrative-led misalignment will increasingly be borne not by marketing teams, but by customers operating under real operational constraints and adversarial pressure.

Adoption Trends

If the performance and cost improvements promised by AI SOC platforms are as material and miraculous as is being claimed, we would expect to see this reflected in high conversion rates from proof-of-value to production deployment.

For example, Gartner's [2025 Hype Cycle for Security Operations](#) places AI SOC agents at the Innovation Trigger stage with 1–5% market adoption (Nunez & Livingstone, 2025, ID G00825402).

Transformational technology drives explosive adoption. That signal is not present. AI SOC appears to behave like other recent “normal” cybersecurity markets, like SOAR or UEBA.

Across conversations with vendors and buyers, a more cautious and fragmented picture emerges.

Where organizations are actually landing

Default to incumbent platforms. Many organizations are waiting for AI capabilities to be integrated into their existing SIEM, XDR, and SOAR platforms rather than introducing a new tool and vendor relationship. The perceived marginal gain does not justify the integration cost, procurement friction, or additional attack surface. For these buyers, AI SOC is not a category, but a feature they expect their current vendors to ship. This is particularly pronounced among CrowdStrike, Microsoft, and Palo Alto customers, who see “AI SOC” announcements from their incumbents and reasonably ask: *why would I buy this separately?*

Selective in-house builds. Mature teams are assembling their own solutions using general-purpose AI and tools (Claude, GPT-4, open-source models, n8n, etc) wrapped around internal tooling. This suggests they do not see a sufficient capability gap between vendor offerings and what they can build themselves, and they may want to retain control and oversight over prompt engineering, data handling, and failure modes. Several detection engineering teams report that a well-prompted LLM with access to their internal documentation outperforms vendor products that lack context on their environment. This shifts the build-vs-buy calculus for teams with the skills to build, especially if AI is seen as a workflow enabler and staff augments.

Deliberate waiting. A large cohort is simply observing: running limited pilots, tracking peer experiences, attending demos, but deliberately not committing budget yet. They are in information-gathering mode. For these organizations, the question is not “should we adopt AI SOC?” but “when will the technology mature enough to warrant the integration effort and operational risk?”

Pilot purgatory. A pattern reported by both vendors and buyers is that proofs-of-value that convert to small-scope production deployments, often then stall. The initial use case works well enough to justify continued spend, but expansion into higher-stakes workflows does not follow. The AI handles enrichment or summarization, but humans retain decision authority. This is not failure, per se. But it is also not the “land and expand” trajectory vendors forecast.

The gap between "adoption" metrics and operational impact

Vendor and analyst narratives frequently cite surveys claiming 50%, 70%, or even 90% of organizations are "exploring," "planning," or "considering" AI SOC adoption. Such figures conflate *interest* with *commitment* and obscure the more telling reality: most usage today is light, constrained, and intentionally kept away from mission-critical activities. Several distinctions matter:

- **Feature enablement is not the same as feature adoption.** A capability can be technically enabled, appearing in dashboards, logging activity, generating outputs, all without analysts actually ever fully relying on it for decisions. Vendors often report "adoption" based on feature activation rates. Buyers describe a different reality: features turned on for evaluation, then ignored or actively worked around.
- **Exposure does not mean trust.** Analysts may see AI-generated summaries, suggested actions, or risk scores on every alert. But that does not mean they actually trust them. In conversations, practitioners describe a frequent pattern, that AI output is treated as one input among many, often weighted below their own judgment and often below the raw telemetry. While exposure may seem high, actual authority granted is low.
- **Adoption breadth is not the same as adoption depth.** An organization may use AI capabilities across multiple workflows (alert triage, report generation, IOC enrichment) while keeping each use case shallow. Breadth of deployment can mask the absence of depth, no single workflow has been fully entrusted to AI judgment.

When vendors claim "strong adoption," it is worth asking: *adoption of what, measured how, and with what decision authority?*

Where AI is being used today

In practice, AI is being applied cautiously in lower-risk areas first:

- **Enrichment:** Pulling context from threat intel feeds, internal documentation, or external sources to annotate alerts.
- **Summarization:** Condensing investigation notes, incident timelines, or compliance evidence for reporting.
- **Report drafting:** Generating first-draft incident reports, executive summaries, or compliance documentation.
- **Workflow acceleration:** Automating repetitive steps (query generation, log formatting, ticket creation) that do not require judgment.

These are not negligible productivity gains. But they are also explicitly *not* the autonomous investigation, agentic response, or analyst-replacement capabilities that dominate vendor positioning. Core detection logic, escalation decisions, and response authority remain mainly human. The business value, or rather the budget this frees up is highly variable and unevenly distributed.

What the pattern suggests

Organizations appear to be operating squarely in the Observe–Orient phases of the OODA loop: gathering information, testing assumptions, and waiting for clearer evidence of reliability before acting.

This is not resistance to AI. It is disciplined risk management from buyers who have lived through previous automation hype cycles like UEBA and SOAR, and learned that vendor enthusiasm often outpaces operational readiness. The appropriate response to uncertainty

in adversarial environments is caution, not faith.

Several factors reinforce the wait-and-see posture:

- **Unclear failure modes.** Buyers do not yet have confidence in how AI SOC tools fail. A SIEM that drops logs fails visibly. An AI that misclassifies a sophisticated attack as benign may fail silently.
- **Measurement difficulty.** Proving ROI requires measuring counterfactuals (incidents prevented, time saved on investigations that would have happened anyway). Vendors offer benchmarks; buyers lack confidence in their applicability.
- **Talent constraints.** Adopting AI SOC tools creates new operational dependencies: prompt tuning, output validation, integration maintenance. Teams already stretched thin are cautious about adding complexity.
- **Vendor instability.** The AI SOC startup landscape is volatile. Buyers are wary of adopting tools from vendors who may pivot, get acquired, or fail to reach sustainability.

Strategic implications

The gap between vendor narratives, e.g. "adoption is accelerating", and buyer reality, "we're evaluating cautiously", is not a communication problem. It is actually a market signal.

When transformational technology arrives, adoption is pulled by buyer demand, not pushed by vendor enthusiasm. The current pattern, fragmented uptake, shallow deployment, persistent skepticism, suggests that AI SOC has not yet crossed the threshold from "interesting" to "essential." The value is real but bounded, but the revolution is not yet quite here. So far the adoption cycle for AI SOC technologies looks like for any other new normal security technology.

IS AI SOC marketing prophecy-driven?

When evidence contradicts belief...

“*When deeply held beliefs are challenged by reality, people rarely abandon the belief. They reinterpret the evidence to preserve it.*”
— **When Prophecy Fails**

Most of what AI SOC vendors promise lies in the future. As such, it would be fair to call it prophecy-driven marketing, or future-expectation-driven marketing.

AI SOC marketing is therefore better understood as prophetic rather than technical. Most claims made today describe a future state, e.g. autonomous investigations, analyst replacement, agentic operations, rather than demonstrable, repeatable capabilities in production. Under investor, market, and career pressure, these future expectations harden into present-tense narratives, even when operational evidence contradicts them.

[When Prophecy Fails](#) describes a simple but durable pattern: when strongly held beliefs collide with contradictory evidence, people tend to reinterpret the evidence rather than revise the belief. Success is redefined, timelines are extended, and responsibility is displaced to preserve the original narrative. This pattern maps cleanly to AI SOC marketing, where unmet expectations are reframed as buyer readiness or trust issues instead of signals of product immaturity.

When adoption is low, accuracy inconsistent, or trust fragile, the narrative is rarely revised; instead, the evidence is reframed as a buyer readiness problem, a trust issue, or a temporary phase on the way to inevitability. Like prophecy, the claim survives by shifting timelines, redefining success, and relocating responsibility, until reality intervenes. In cybersecurity, however, belief systems are not validated by conviction or repetition, but by incidents, and incidents are indifferent to prophecy.

When ambitious claims are made publicly, especially under investor, market, or career pressure, evidence that contradicts those claims creates tension. Rather than adjusting the narrative (“the technology isn’t ready yet”, or “here is what the technology is ready for”), organizations often adjust the interpretation (“users don’t trust it,” “they’re not ready,” “adoption takes time”). The belief is preserved; the evidence is reframed.

In the case of marketers, this is understandable. They have no control over the maturity of the technology. Their job is to sell it regardless. But it actually makes their job harder, not easier, because trust is easy to lose but hard to restore.

The AI SOC Adoption Trap Lifecycle

| Claim | Friction | Rationalization | Risk |
|--|--|---|---|
| What is promised <ul style="list-style-type: none"> ● “Autonomous investigations” ● “AI analysts” ● “50% productivity gains” ● “SOCless-ready operations” | What happens in reality <ul style="list-style-type: none"> ● Analysts selectively ignore AI output ● AI used for summaries, not decisions ● Low trust during live incidents ● Features technically enabled, practically dormant | How the gap is explained away <ul style="list-style-type: none"> ● “It’s a trust problem” ● “Users aren’t ready yet” ● “Change takes time” ● “Adoption is actually strong if you count exposure” | Where this would end up <ul style="list-style-type: none"> ● False confidence in coverage ● Slower escalation during incidents ● Missed detections hidden behind “AI-assisted” labels ● Strategic decisions made on narrative, not performance |
| Narrative: <i>The technology is almost mature. The leap is inevitable.</i> | Reality: <i>The system works... until it matters. ->The demos look great. The alert incident queue looks great. The 2am incident is where it breaks.</i> | Psychological move: Reality is reinterpreted to preserve belief <i>-> Exactly the pattern described in When Prophecy Fails</i> | Key asymmetry: Marketing can survive contradiction. Adversarial systems cannot. |

Mature SOCs break the misconception loop early: They revise claims to match reality, not the other way around.

Why Too Much Optimism About AI Is Dangerous in Cybersecurity

Marketing narratives can survive contradiction. Security operations cannot. Reality asserts itself during incidents, not on slides.

Cybersecurity has a uniquely low tolerance for optimism divorced from evidence. Unlike productivity tooling, marketing automation, or customer analytics, security systems are not evaluated by average-case performance or aspirational roadmaps. They are judged by worst-

case outcomes, adversarial pressure, and what fails when assumptions break. In this context, “toxic optimism”, or the insistence that AI capability will rapidly mature into reliability if users simply trust it more, is not only misplaced, it is actively dangerous.

CISOs are not new to AI promises. For more than a decade, the industry has cycled through claims that machine learning would dramatically reduce false positives, eliminate analyst toil, or autonomously detect unknown threats. UEBA, anomaly detection, behavioral analytics, SOAR, XDR, and now LLM-driven “AI SOCs” have all arrived with similar narratives. Each wave delivered incremental value, but none delivered the wholesale autonomy or analyst replacement implied in its marketing. As a result, security leaders have developed a well-earned skepticism, not toward AI itself, but toward AI vendors and their claims, that often try to collapse future potential into today’s inevitability.

The challenge is not that AI fails outright, but that its failure modes are subtle. An AI system that misses a critical alert once in a thousand cases can still appear “highly accurate” in dashboards, demos, and benchmarks. In security operations, however, that single miss can define the outcome. Optimism that ignores asymmetric risk, where for example one false negative outweighs hundreds of correct benign classifications, conflicts directly with how CISOs are accountable. Security leaders are paid to assume things will break, not to assume they will improve on schedule.

As we have [detailed](#) before, this is why cybersecurity has historically been resistant to black-box automation. Trust is not granted by novelty or vendor confidence, but earned through repeated exposure to failure and recovery under real conditions. When AI systems are framed as inevitable autonomous agents rather than probabilistic tools with bounded reliability, skepticism is not resistance to change, it is professional risk management. CISOs have spent years cleaning up the consequences of overpromised automation; they are unlikely to suspend that judgment simply because the latest model sounds more affable and personable.

In this light, toxic optimism around AI SOCs does not fail because CISOs are conservative or emotionally resistant. It fails because security is one of the few enterprise domains where reality consistently wins, loudly and without compromise.

As such, our negativity bias is not a bug, it is by design. Product Manager and Marketers ignore this at their peril, or they show that they not just fail to understand their buyer, they fail to empathise with them as well.

OSINT Synthesis: What Practitioners Actually Say About AI SOCs

Source characteristics

- Public OSINT (Reddit, Discord)
- Age: < 1 year (current sentiment)
- Practitioner-heavy (SOC analysts, MSSP operators, staff engineers)
- Low incentive for narrative alignment or vendor appeasement

These are *not* buyer survey responses or post-sales references.

The Architectural Pattern Behind the Failures

Almost every negative outcome traces back to one of five systemic errors:

1. Predictions misrepresented as Classification
2. Binary decisions in a probabilistic domain
3. Too tight coupling between AI output and irreversible actions
4. Too much authority granted before reliability is proven
5. Lack of context and other data for the AI to be more accurate

OSINT Analysis

| Practitioner Quote / Observation | Primary Failure Mode | What's Broken Architecturally | Why This Fails in Practice |
|---|----------------------|--|--|
| "LLM hit 71% accuracy... missed the actual test incident entirely." | Classification | Over-reliance on probabilistic LLM classification for adversarial events | Accuracy metrics mask asymmetric cost: one missed true positive outweighs dozens of correct benign calls |
| "Calling obvious false positives 'Malicious'." | Classification | No adversary-aware decision boundaries | LLMs optimize for linguistic plausibility, not operational risk minimization |
| "Cry-wolf problem is real." | Confidence Modeling | No calibrated confidence or uncertainty scoring | Without confidence gradients, analysts cannot triage trust — everything feels equally wrong |
| "Binary Malicious/Benign makes things messy." | Confidence Modeling | Forced binary outcomes | SOC work is tri-state at minimum (Benign / Suspicious / Malicious); binaries destroy judgment |
| "We need a 'Suspicious' middle tier." | Escalation Logic | Missing human-in-the-loop routing tiers | Escalation must be conditional, not absolute |

| | | | |
|--|-------------------------------|--|---|
| "Analysts stop trusting it pretty fast." | Trust Feedback Loop (missing) | No negative-feedback learning loop | Systems fail silently and do not adapt to analyst override behavior |
| "Someone created a rule to auto-close every alert." | Workflow Coupling | AI outputs directly bound to closure actions | Tight coupling amplifies failure instead of containing it |
| "Noise reduction... outcome was zero visibility." | Workflow Coupling | No guardrails or blast-radius limits | Automation lacked circuit breakers or kill-switch thresholds |
| "AI is only good enough to support a L1 Analyst." | Role Misalignment | AI positioned above its competence envelope | LLMs are assistive reasoners, not accountability-bearing actors |
| "Agentic AI takes the right info and interprets it incorrectly." | Context Binding | Weak semantic grounding across data sources | Correlation ≠ comprehension; LLMs lack causal validation |
| "Sales said agentic AI could replace L1 SOC." | Escalation Logic | Premature autonomy claims | Escalation authority transferred without reliability proof |
| "Why do they need internal ticketing data?" | Workflow Coupling | Learning procedural artifacts instead of signals | Ticket data encodes human workarounds, not ground truth |
| "Can't imagine missing a critical alert due to misclassification." | Escalation + Classification | No fail-safe bias toward safety | Systems optimize efficiency over resilience |
| "AI works better | Correct | Assistive, not | AI augments |

| | | | |
|---|-----------------------------|---|---|
| for enrichment than final close.” | Pattern | decisive integration | cognition without owning consequences |
| “Here’s why it’s probably fine... here’s why it’s flagged.” | Healthy Confidence Modeling | Explicit uncertainty + rationale | Trust emerges when ambiguity is preserved, not hidden |
| “AI SOC reduced toil; auto-triaging with guardrails.” | Constrained Autonomy | Scoped automation with reversible actions | Autonomy limited to low-risk, high-confidence domains |

Buyer Checklist

This framework shifts the conversation from ‘what the AI could do’ to ‘what the system is trusted to do today. The checklist focuses on control boundaries: where decisions are made, how uncertainty propagates, and who retains accountability

Critical Question

“ Which part of your system is allowed to say ‘I don’t know’ and what happens next? ”

1. Classification: What decisions does the AI actually make?

Before evaluating AI accuracy, buyers must understand which decisions the system is trusted to make on their behalf.

| Ask: | Red flags: | What good looks like: |
|--|---|--|
| <ul style="list-style-type: none"> ● What decisions does the AI make <i>without</i> human approval? ● Which classifications are advisory vs authoritative? ● What happens when the AI is wrong operationally? | <ul style="list-style-type: none"> ● “High accuracy” without asymmetric risk discussion ● Binary Malicious / Benign answers ● No explicit false-negative handling strategy | <ul style="list-style-type: none"> ● Clear separation between <i>analysis</i> and <i>decision</i> ● Explicit bias toward resolving false positives over false negatives ● Documented blast radius for misclassification |

2. Confidence Modeling: *Where does uncertainty live?*

This question examines how uncertainty is represented, surfaced, and acted upon within the system. Confidence modeling determines whether analysts can see *how sure* the AI is, not just *what* it concludes. If the system cannot express uncertainty, it cannot be trusted with decisions.

| Ask | Red flags: | What good looks like: |
|--|---|--|
| <ul style="list-style-type: none"> ● Does the system expose confidence or uncertainty scores? ● Is there a “Suspicious / Needs Review” tier? ● Can analysts see <i>why</i> confidence is low? | <ul style="list-style-type: none"> ● Forced binary outcomes ● “Trust the AI” rhetoric ● Confidence hidden behind UX simplification | <ul style="list-style-type: none"> ● Tri-state or multi-state outcomes ● Rationale + uncertainty presented together ● Confidence drives routing, not just display |

3. Escalation Logic: *Who decides when humans get involved?*

This question examines how and when responsibility is transferred from automated analysis to human decision-making. Escalation logic defines who ultimately decides when an alert is serious enough to require human judgment. If escalation is vague, accountability is too.

| Ask: | Red flags: | What good looks like: |
|--|--|--|
| <ul style="list-style-type: none"> ● What triggers human escalation? ● Can escalation thresholds be tuned per use case? ● Can analysts override escalation logic? | <ul style="list-style-type: none"> ● “Fully autonomous investigation” ● Escalation described as a roadmap feature ● No audit trail for escalation decisions | <ul style="list-style-type: none"> ● Deterministic escalation rules ● Human-in-the-loop by default for high-impact alerts ● Clear separation between detection, investigation, response |

4. Workflow Coupling: *What happens automatically?*

This question set examines how tightly AI outputs are connected to downstream actions within operational workflows. Workflow coupling determines which AI outputs trigger automatic actions and which require explicit human approval. Automation without decoupling turns small errors into cascading systemic failures.

| Ask: | Red flags: | What good looks like: |
|--|---|---|
| <ul style="list-style-type: none"> ● What actions can the AI take without approval? ● Are closures reversible? ● Are there circuit breakers or kill switches? | <ul style="list-style-type: none"> ● Auto-close enabled as a rule, not the exception ● Tight coupling between classification and closure ● “Noise reduction” without visibility guarantees | <ul style="list-style-type: none"> ● Reversible actions ● Guardrails on automation scope ● Automation limited to low-risk, high-confidence paths |

5. Learning Inputs: *What data does the AI learn from and why?*

This section evaluates which data sources influence the system’s behavior and how learning is constrained. Learning inputs determine whether the AI is learning from security signals or inheriting human workarounds. If the AI learns from broken workflows, it will automate their flaws.

| Ask: | Red flags: | What good looks like: |
|--|--|--|
| <ul style="list-style-type: none"> ● Why does the system need ticketing or case-management data? ● Is it learning signals or human workarounds? ● How do you prevent procedural bias from becoming “truth”? | <ul style="list-style-type: none"> ● “We need your tickets to learn your playbooks” ● No distinction between signal data and process data ● Model retraining tied to analyst behavior, not outcomes | <ul style="list-style-type: none"> ● Clear separation between security telemetry and workflow artifacts ● Learning bounded to enrichment, not authority ● No silent retraining on analyst overrides |

6. Trust Feedback: *How does the system learn from rejection?*

Here we assess how analyst overrides and rejections are captured and used by the system. Trust feedback determines whether the system improves when humans disagree with it. If analyst distrust is not integrated as a signal, the system can never learn or progress beyond its own failures.

| Ask: | Red flags: | What good looks like: |
|---|---|---|
| <ul style="list-style-type: none"> ● What happens when analysts ignore or override AI output? ● Is distrust captured as a signal? ● Can I see override rates and patterns? | <ul style="list-style-type: none"> ● Overrides treated as user error ● No negative-feedback loop ● Adoption metrics based on exposure, not usage | <ul style="list-style-type: none"> ● Override telemetry is first-class data ● Feedback visibly improves system behavior ● Trust measured longitudinally, not claimed |

7. Adoption Reality: *What does “used” actually mean?*

This question set focuses on how adoption is defined and measured across different operational contexts. Adoption only matters if it holds under incident pressure. If it’s disabled during real incidents, it was never truly trusted or adopted.

| Ask: | Red flags: | What good looks like: |
|---|---|--|
| <ul style="list-style-type: none"> ● How many customers rely on this during live incidents? ● Which features are <i>consistently</i> used under pressure? ● What gets disabled first when things go wrong? | <ul style="list-style-type: none"> ● “50% adoption” without definition ● Pilot usage counted as production ● Reference customers who won’t discuss incidents | <ul style="list-style-type: none"> ● Clear differentiation between pilot, assistive, and decisive use ● Honest discussion of failure modes ● Customers who describe <i>limits</i>, not miracles |

8. Failure Mode Disclosure: *What breaks first?*

This section examines how failure modes are documented, communicated, and contained. Mature systems acknowledge where they fail before customers discover it themselves. If nothing is documented as unsafe, the system lacks a meaningful way to determine or communicate its own operational boundaries.

| Ask: | Red flags: | What good looks like: |
|--|--|--|
| <ul style="list-style-type: none"> ● Where does the system fail today? ● Which use cases should we <i>not</i> automate? ● What have customers disabled, and <u>why</u>? | <ul style="list-style-type: none"> ● “We haven’t seen that” ● No documented limitations ● Everything framed as roadmap solvable | <ul style="list-style-type: none"> ● Explicit “do not use” guidance ● Known failure modes published ● Architecture designed to degrade safely |

9. Roadmap Independence: *What improves without better models?*

This section examines whether the product roadmap delivers meaningful improvements without relying on future advances in model capability. Mature systems strengthen control, safety, and reliability even when models are imperfect. If progress depends primarily on “better AI later,” the roadmap lacks a credible path to near-term operational maturity.

| Ask: | Red flags: | What good looks like: |
|--|---|--|
| <ul style="list-style-type: none"> ● What roadmap items deliver value if model accuracy does <i>not</i> improve over the next 18–24 months? ● Which upcoming feature, improve auditability, explainability or strengthen human control? ● Which roadmap items explicitly assume better reasoning, fewer hallucinations, or higher trust in the model? | <ul style="list-style-type: none"> ● Most roadmap value tied to “next-gen models” ● Promises of autonomy without corresponding control or decision intelligence improvements ● Roadmap language dominated by “smarter,” “more capable,” or “more autonomous” without explaining how ● Critical safety and explainability features deferred to undefined future releases | <ul style="list-style-type: none"> ● Clear separation between model capability and system safety ● Roadmap investment in confidence-aware routing, multi-state outcomes, escalation and deterministic controls. ● Improved audit trails, override visibility, and post-incident explainability ● Guardrails, circuit breakers, and reversibility delivered <i>before</i> expanded autonomy |

Final Words

Look, let's be real. The problem isn't the technology. LLMs and generative AI are major technological milestones. The underlying capabilities, whether LLM-assisted triage, enrichment, or summarisation, are real and genuinely useful. In constrained contexts.

The real problem is the industry-wide misalignment between how these capabilities are marketed and what is actually possible now, with today's technology and maturity. It's all fun and games theorising about agentic security organisations, but that's simply not what SOCs can buy right now, if ever.

This matters because security operations is not a domain that tolerates optimism disconnected from evidence. Adversaries do not wait for roadmaps to mature. The cost of misplaced confidence is not a failed marketing quarter. It is missed detections and potential breaches!

The practitioners we spoke to are not resistant to AI. They are resistant to being told that a new tool is "transformational" when it is really just another tool, one that requires the same careful scoping, validation, and operational oversight as everything else in the stack, if not more, due to the probabilistic, non-deterministic, and realistically experimental nature of these tools. Practitioners are especially resistant to having their professional scepticism reframed as a psychological barrier. They are doing exactly what security professionals should do: demanding evidence, testing, and cautiously weighing risk versus benefit.

For vendors, the path forward is straightforward, if uncomfortable: stop explaining away friction and start learning from it. Lead with limitations, publish failure modes, measure adoption by what survives incident pressure, not by what gets enabled once but never used in anger. Your job in an early-stage market is to educate users and help them identify early, achievable wins. That business case doesn't write itself (despite what AI influencers may say).

Stop treating buyer caution as a problem to be overcome and start treating it as a signal to work harder on making AI easier to operationalise. Because that's what it is. It's not that enterprises aren't AI-ready. It's that most AI isn't enterprise-ready. That is what the product should be, and what buyers will actually pay for.

And consider your path forward if model capabilities don't experience another major leap. And also if they do. What value can you provide? Most AI vendors need AI models to improve, but not by so much that they won't be needed. That contradiction is easily resolved by providing what frontier models can't: deep integration into security stacks and real-world security domain expertise.

For buyers, we intended the checklist in this paper as a starting point. The most critical question remains: which part of this system is allowed to say "I don't know," and what happens next? Any vendor who cannot answer that question clearly has not yet built a system ready for production.

The AI SOC will arrive, but it will not arrive overnight. When it does, it will be the product of solid engineering, disciplined iteration, and honest feedback, not marketing momentum.

About the Authors

[Oliver Rochford](#) is Lead Analyst at Cyberfuturists and Co-founder of [Aunoo AI](#). A former Research Director at Gartner, Securonix, and Tenable, he co-originated the SOAR (Security Orchestration, Automation and Response) category and conducted over 4,000 engagements with security practitioners, vendors, and enterprise security leaders over his career.

[Dr. Anton Chuvakin](#) is now involved with security solution strategy at Google Cloud, where he arrived via Chronicle Security (an Alphabet company) acquisition in July 2019. Anton was, until recently, a Research Vice President and Distinguished Analyst at Gartner for Technical Professionals (GTP) Security and Risk Management Strategies team.

Anton is a recognized security expert in the field of log management, SIEM and PCI DSS compliance. He is an author of books "Security Warrior", "Logging and Log Management: The Authoritative Guide to Understanding the Concepts Surrounding Logging and Log Management" and "PCI Compliance, Third Edition: Understand and Implement Effective PCI Data Security Standard Compliance" (book website) and a contributor to "Know Your Enemy II", "Information Security Management Handbook" and other books. Anton has published dozens of papers on log management, SIEM, correlation, security data analysis, PCI DSS, security management. His blog "Security Warrior" was one of the most popular in the industry.

Special thanks to

[Johnathan Dempsey](#)

Appendix: Evidence Base

A. Methodology

This paper draws on three evidence streams:

1. **OSINT corpus.** Public practitioner commentary collected from Reddit cybersecurity communities (r/cybersecurity, r/Information_Security). All posts were less than one year old at the time of collection. Contributors self-identified as SOC analysts, MSSP operators, detection engineers, and staff-level security engineers. Posts were coded inductively against the paper's core themes.
2. **Semi-structured interviews.** Two extended interviews with practitioners who have deployed AI SOC platforms in production (not pilots). Informants were selected through theoretical sampling to span different market segments. Both are anonymised below.
3. **Vendor interactions.** 30+ vendor briefings conducted over 18 months, referenced in the body text but not reproduced here.

Gartner's 2025 Hype Cycle for Security Operations places AI SOC agents at the Innovation Trigger stage with 1–5% market adoption (Nunez & Livingstone, 2025, ID G00825402). Finding production-deployed practitioners to interview was difficult in itself, which corroborates the low adoption figure and explains the small interview N. The two informants were selected to maximise contrast across organisation size, team structure, and market segment.

B. Case Study Profiles

INF-01 is a CISO at a large European mobility company with a dedicated SecOps team. His previous human MDR arrangement was distrusted to the point where analysts routinely redid investigations themselves. He has deployed an AI-driven MDR in full production, eventually replacing the human MDR entirely.

INF-02 is a sole security practitioner at a US-based conservation NGO, handling everything from policy to operations across roughly 1,000 users. His previous \$60K MDR did no real investigation, just forwarded alerts. He came in as a design partner with an AI SOC startup and runs it in production at around half the cost of the MDR, with his CrowdStrike detection engineering background shaping how he evaluates what he sees.

C. Interview Evidence Summarised by Theme

C.1 The "80% vs 0%" argument

INF-01 framed AI accuracy against the alternative of not investigating at all:

"80% is better than 0% accuracy, right? So if you don't look at something, I would rather have something look at it with the 80% accuracy and not just ignore it."

He argued that human analysts also make mistakes, are not auditable in the same way, and suffer fatigue and churn. The AI's advantage was consistency at scale, not perfection.

C.2 Both informants draw the same line at no autonomous response

INF-01 restricted AI to detection and investigation only:

"We don't do any automated write access stuff on anything. Not remediation, not patching, not anything. The potential to break stuff is so large that you could end up with more damage – self-inflicted damage."

His team operated in "hybrid mode": AI handles scale (triage, enrichment, investigation steps), humans handle response, containment, and edge cases.

INF-02 drew the same boundary independently, framing it differently:

"You can't hold a computer accountable, right, like IBM says? [...] If it comes down to that point in investigation, it rises to the point where a human needs to be in that loop."

Response actions existed in the platform but required a human to execute them. Both informants arrived at this constraint from different starting points (enterprise risk management vs. accountability principle) but reached the same architecture.

C.3 Explainability as trust infrastructure

INF-01 described explainability as his key requirement from day one:

"I want to see why conclusions were derived. I want to be able to see all the reasoning flow for every incident."

He drew a direct parallel to the old MDR problem: his team used to receive verdicts from external analysts and then redo the investigation because they did not trust the reasoning.

The AI's auditable chain solved this. When the platform recommended aggressive actions (factory reset on a VP's laptop), the team could present the full reasoning as justification. Trust came from transparency, not from vendor assurances.

INF-02 described a related but distinct mechanism: the platform's built-in AI chat let non-technical stakeholders query investigations directly and get explanations at their level. Explainability was not a reporting feature; it was how people outside the SOC could participate in decisions.

C.4 Honest uncertainty means a system that says "I don't know"

INF-02 described the platform returning "inconclusive" verdicts:

"There's scenarios where it comes back with either inconclusive or that it should be something that should be monitored. And I think nine times out of ten, the reason why it comes back that way is because it doesn't have enough data to make a decision."

He treated this as a legitimate output, not a failure. The system acknowledged its own limits rather than forcing a binary call. This maps to the "tri-state or multi-state outcome" pattern identified in the OSINT corpus as a marker of well-designed architecture.

C.5 The agent as attack surface

INF-02, drawing on his detection engineering background, raised a concern absent from vendor marketing: prompt injection through attack data.

"If a malicious actor is prompt injecting through their attack [...] the system is pulling in that data [...] and it has the ability and the capability to do actions on its own with no human in the loop – well, then you basically provided your own insider risk at that point."

The tighter the coupling between AI inference and automated action, the greater the blast radius of adversarial manipulation. This extends the workflow coupling failure mode into an adversarial dimension.

C.6 Scope expansion is where the real value showed up

INF-01 described a case where the AI deployed detection rules his team had previously considered impractical due to false positive rates. The system correlated HR data, LinkedIn profiles, and authentication logs across franchise locations to identify systematic credential sharing between branches, a pattern no human analyst would have investigated at scale.

"My team would never be able to go to that depth of investigation for such a low-risk topic. To me, this is the real value – because we can tremendously expand the scope and the depth of investigations we are currently doing."

C.7 Detection engineering is unlocked

A second-order effect in INF-01's environment: deploying new detection rules used to require "two to three months of testing to see that not too many false positives are generated." With AI-driven triage absorbing the false positive load, his team could deploy experimental detections without that overhead:

"Right now we can just try. We can just throw it to the thing, see what comes back. And if it

doesn't work, it's an agent. It's not like I have to waste two FTEs to take care of some false [positives]."

C.8 The one-person SOC

INF-02 was a sole practitioner responsible for everything from policy to operations across ~1,000 users and 20–30 locations. His previous MDR (\$60K/yr) was pass-through alerting: it ingested data from Defender, AWS, and GCP but performed no correlation. An analyst might spend a few minutes reviewing an alert before forwarding it.

"For me, that's not a very good approach to it, right? Especially as being a one-person team, I need to have a lot of that already done so that when it comes to me [...] I can just make a quick decision based off of all the information, the context, the correlation."

The AI platform cost 50% less than the MDR and freed budget to invest in vulnerability management (Tenable).

C.9 Workforce disruption is the surprise neither informant planned for

INF-01 described organisational impact he did not anticipate:

"One of the things that surprised us [...] it required our SecOps team to go through some kind of evolution. People suddenly had no job."

Roles focused on phishing triage, header analysis, and DMARC verification were automated within weeks. Staff were retrained into cloud threat investigation. Threat hunting plans were reconsidered. The transition was reactive:

"I would have invested much sooner in creating new positions or new titles in the team [...] because what we did is we incorporated AI, got it to a stage where it's really working, and then said, okay, now what do we do with the people."

His hiring had changed: "I think it's going to be a while before I hire somebody new."

D. Cross-Informant Analysis

The two informants occupy very different positions in the market. INF-01 is a CISO at a large European enterprise with a mature security team; INF-02 is a sole practitioner at a mid-sized NGO with no dedicated security staff. Despite this, their accounts converge on several points:

Where they agree:

- Both restrict AI to detection and investigation, withholding autonomous response authority
- Both cite explainability and auditable reasoning as critical to building trust, not a nice-to-have
- Both describe the AI's primary value as expanding investigative scope and depth, not replacing analysts
- Both ran the AI platform in production alongside existing services before committing
- Both found that the AI outperformed their previous MDR arrangements on investigation quality

Where they differ:

- INF-01 experienced workforce disruption (roles automated, staff retrained). INF-02 is a one-person team with no staff to displace. The AI substitutes for a team he never had.
- INF-02 raised adversarial risk (prompt injection through attack data). INF-01 did not mention this, likely reflecting different threat models and technical backgrounds.
- INF-01 operates in a compliance-heavy environment where audit trails matter for regulatory reasons. INF-02's compliance requirements are lighter, but he still values explainability for his own decision-making.

The convergence across such different contexts strengthens the paper's arguments about constrained autonomy and transparency as architectural requirements, not preferences.

E. OSINT Evidence: Practitioner Discourse on AI SOC Platforms

Source Description

Commentary was collected from public cybersecurity communities on Reddit (r/cybersecurity, r/Information_Security). All posts were less than one year old at the time of collection. Posts were coded inductively against the paper's core framework and categorised into thematic clusters.

Table E1: Critical Themes and Points of Friction

| Theme | Representative Quote(s) | Source Thread | Failure Mode |
|---|--|--|---|
| Classification accuracy and false confidence | "When we ran an LLM against 348 known false positives plus one synthetic true positive, the LLM alone hit 71% accuracy – which sounds okay until you realize it was calling obvious false positives 'Malicious' and missed our actual test incident entirely." | r/cybersecurity – "Looking for AI SOC Tools That Integrate with Rapid7 InsightIDR" | Over-reliance on probabilistic LLM classification for adversarial events. Accuracy metrics mask asymmetric cost of missed true positives. |
| Binary decision-making | "The 'auto-close false positives' promise is where things get messy in practice. [...] I'd push hard on any vendor about how they handle the classification confidence – whether they have a 'Suspicious' middle tier." | r/cybersecurity – "Looking for AI SOC Tools That Integrate with Rapid7 InsightIDR" | Forced binary outcomes in a domain requiring tri-state classification. |
| Cry-wolf erosion | "If your AI automation | r/cybersecurity – | No negative-feedback |

| | | | |
|--|---|--|--|
| of trust | keeps escalating benign stuff as urgent, analysts stop trusting it pretty fast. One team I talked to found someone had literally created a rule that just auto-closed every alert without investigating." | "Looking for AI SOC Tools That Integrate with Rapid7 InsightIDR" | learning loop. Tight coupling between AI output and closure actions. |
| LLM hallucination risk | "I have seen first hand how LLMs hallucinate and you don't want to take that risk." | r/cybersecurity – "AI for SOC Use Case" | Probabilistic outputs treated as deterministic classification. |
| Role misalignment | "[MSSP] AI is only good enough to support a L1 Analyst at present. It absolutely can't be trusted to run the show yet." | r/cybersecurity – "Agentic AI SOC question" | AI positioned above its competence envelope. |
| Premature autonomy claims | "Sales said agentic AI could replace L1 SOC." | r/cybersecurity – "Agentic AI SOC question" | Escalation authority transferred without reliability proof. |
| Interpretation failures | "The frequency with which I see agentic AI take the right information and interpret it incorrectly is staggering." | r/cybersecurity – "Agentic AI SOC question" | Weak semantic grounding. Correlation ≠ comprehension. |
| Suspicious data access requests | "Why do they need internal ticketing data?" | r/cybersecurity – "Agentic AI SOC question" | Learning procedural artefacts instead of signals. |
| MSSP incentive misalignment | "If it makes an MSSP's life easier, it's probably worse for security." | r/cybersecurity – "AI for SOC Use Case" | Provider efficiency optimisation vs. customer security outcomes. |

Table E2: Positive Themes – Where Practitioners Report Value

| Theme | Representative Quote(s) | Source Thread | Pattern |
|---|--|---|--|
| Investigation enrichment over final verdict | <p>"Where we've seen AI work better is the investigation enrichment part rather than the final close decision. Having an agent pull context, check baseline behavior, look at threat intel, and give an analyst a summary with 'here's why it's probably fine, here's why I'm still flagging it.'"</p> | <p>r/cybersecurity — "Looking for AI SOC Tools"</p> | <p>Assistive, not decisive integration. Human retains closure authority.</p> |
| Massive alert volume reduction with guardrails | <p>"It turns the toil of tuning and managing massive alert volumes into less than an hour a month. It's currently filtering down to about 0.012% of my total volume [...] over half a million alerts/events a month into 1-3 relevant things a day."</p> | <p>r/cybersecurity — "Are AI SOC Analysts the future or just hype?"</p> | <p>Constrained autonomy with scoped auto-triage.</p> |
| Contextual investigation | <p>"It will literally look up internally and externally for relevant data to make a decision. [...] 'does this event fire every time the user's logged in or is it new and only today?'"</p> | <p>r/cybersecurity — "Are AI SOC Analysts the future or just hype?"</p> | <p>AI augments judgment by gathering context a human would skip under time pressure.</p> |
| Won competitive | <p>"Dropzone won two POCs, first against</p> | <p>r/cybersecurity — "Are AI SOC</p> | <p>Cost displacement</p> |

**evaluation
against SIEM**

other AI vendors,
then again over
SIEMs like Panther
and SumoLogic.
Saving over several
hundreds of
thousands a year in
SaaS tooling."

Analysts the future
or just hype?"

through alert
pipeline
reduction.

**False
positive
reduction**

"Similarly used
Prophet Security and
they also significantly
reduced the number
of false positives for
our team. [...] The
biggest advantage
was extracting all the
evidence from Okta,
SIEM, and S3,
correlating it and
coming up with a
determination."

r/Information_Sec
urity — "Anyone
used AI SOC
Platforms"

Multi-source
evidence
correlation as
primary value.