WHITEPAPER

Verifiable AI: The Foundation for Trust in Cybersecurity

A White Paper for Security Leaders

uptycs

Table of Contents

Executive Summary

Security operations face a paradox: more AI-powered tools than ever, but less trust in their decisions.

Teams process 960+ alerts daily, spending 25% of their time chasing false positives. AI was supposed to solve this. Instead, it created a new problem: **black-box intelligence that can't be verified or audited.**

When AI flags a threat, analysts face an impossible question: How do we know this is right?

Current AI systems tell analysts what they concluded. Verifiable AI shows them why—with evidence chains traced to specific log entries, timestamps, and event IDs.

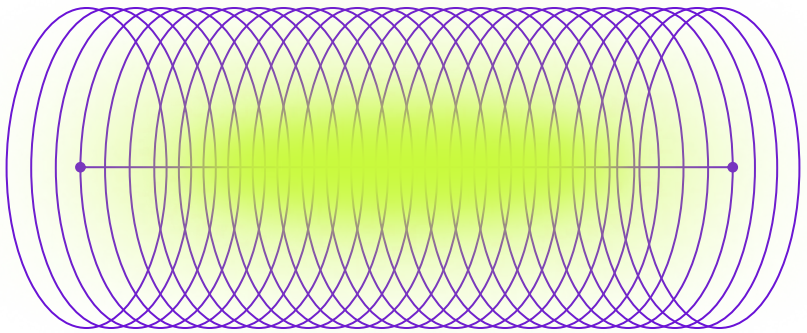
This paper introduces Verifiable AI as the foundation for trustworthy security operations, demonstrated through Juno – the first AI Security Analyst, built on Uptycs' unified security platform and designed to show its work.

What makes Verifiable AI different:

- Every finding traces to specific log entries with timestamps and event IDs
- Analysts verify claims independently rather than trusting blindly
- Audit trails are byproducts of investigation, not separate work
- Junior analysts learn from verifiable reasoning, not black boxes

The impact: Investigation time cut by more than half. Complete audit compliance. Knowledge that transfers across teams. AI that earns trust by enabling verification.

In security, proof has always mattered more than persuasion. Verifiable AI delivers both.



I. The AI Trust Crisis in Security Operations

We Solved Collection, Not Comprehension

The modern Security Operations Center represents a fundamental contradiction. Organizations have invested heavily in detection tools, threat intelligence platforms, SIEM systems, and AI-powered analytics. Data collection has never been more comprehensive. Alert generation has never been more sophisticated.

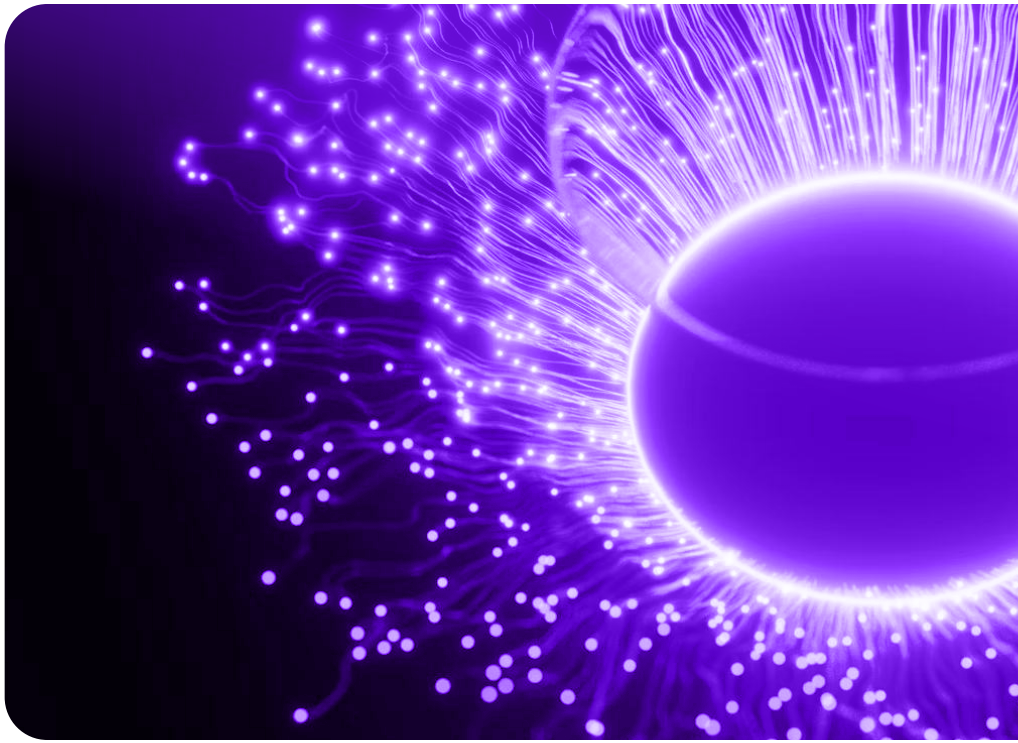
Yet security teams are overwhelmed, not empowered.

Recent research reveals the scale of the crisis: organizations process an average of 960 alerts per day, with large enterprises handling over 3,000 daily alerts from 30+ different security tools.[1] Close to half of analyst teams battle false positive rates exceeding 50%,[2] spending approximately 25% of their time—15 minutes of every hour—chasing false positives.[3][4] Meanwhile, 40% of alerts go uninvestigated entirely, and full investigation of a single alert averages 70 minutes.[1]

The signal-to-noise ratio hasn't improved. It's gotten worse.

AI was supposed to fix this—filtering noise, prioritizing threats, accelerating response. Modern security tools use AI to detect network anomalies, identify malware through behavioral analysis, correlate events across disparate sources, and predict attack patterns. These capabilities deliver real value.

But something fundamental is missing: trust.



The Black Box Problem

When an AI system surfaces a critical alert, analysts face an impossible choice:[6]

- **Trust it blindly** and potentially waste hours on a false positive—or worse, take disruptive action based on faulty reasoning.
- **Ignore it** and risk missing a real threat that leads to a breach.
- **Try to verify it** but lack the reasoning trail to understand why the AI reached its conclusion.

This isn't theoretical. Consider what happens in practice:

An AI-powered SIEM flags unusual database access as critical. The analyst investigates—it's a database administrator performing scheduled maintenance. The AI saw unusual volume and timing but had no context about maintenance windows. Result: Three hours wasted, growing skepticism about future alerts, and no way to verify the AI's reasoning to prevent recurrence.

An endpoint detection system identifies "lateral movement" with 94% confidence.

The security team isolates affected systems, disrupting a critical business process. Post-incident analysis reveals the "threat" was a legitimate system administrator using standard admin tools. The 94% confidence was based on behavioral patterns, not actual evidence of compromise.

During a compliance audit, auditors ask: "How did your AI determine this was a security incident?" Answer: "Machine learning detected anomalous behavior." Follow-up: "What specific evidence supported this determination?" No clear answer exists. The finding can't be independently verified.

The pattern: AI systems optimized for speed and detection have created a new form of technical debt—**reasoning debt.**

When Trust Matters Most

Security isn't a domain where "good enough" suffices. Security AI makes decisions with material consequences: incident response actions that can disrupt business operations, threat classifications that determine resource allocation, root cause analysis that shapes future security posture, and compliance findings that must withstand regulatory scrutiny.

In each case, the question isn't just "Is the AI accurate?" It's "Can we prove it?"

CISOs understand this viscerally. A common theme emerges in conversations with security leaders: **"We want AI to make us faster, not just to make us feel faster."** The difference is verifiability.

The Hidden Costs

The trust deficit creates a vicious cycle: analysts spend 30-40% of investigation time validating AI recommendations instead of investigating threats. When explanations don't enable verification, skepticism grows and alert fatigue intensifies. Meanwhile, opaque AI reasoning prevents knowledge transfer—junior analysts can't learn from black boxes, concentrating expertise in senior staff who work around the AI. Sophisticated attackers probe these systems to map their blind spots, while compliance auditors flag documentation gaps that black-box decisions create.

These aren't future concerns. They're present realities shaping how security teams adopt—or resist—AI today.

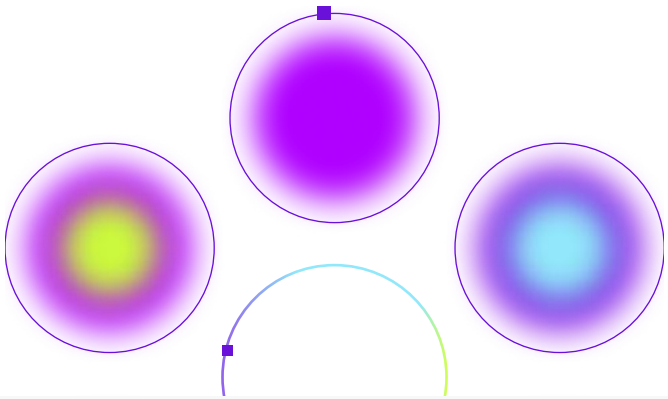
II. Introducing Verifiable AI

A New Standard for AI-Assisted Security

The gap between what current AI systems provide and what security operations actually need points toward a fundamentally different approach. Rather than asking AI to explain its thinking, we need AI that constructs reasoning chains that security teams can verify independently.

This is **Verifiable AI:** AI systems that provide evidence-based reasoning with complete audit trails, enabling human analysts to independently confirm every step from observation to conclusion.

Verifiable AI fundamentally changes the AI adoption question. The question changes from "Do we trust this AI?" to "Can we verify this reasoning?"



Core Principles

Three foundational principles distinguish Verifiable AI from previous approaches:

1. **Human-Verifiable Inference** Every conclusion traces back to specific evidence that a human analyst can independently confirm. No "trust the model" steps. No statistical abstractions that obscure the underlying facts.

2. **Complete Reasoning Chains** The path from observation to conclusion is explicit and documented. Each logical step is stated clearly enough that a different analyst—or an auditor—could follow the same evidence to the same conclusion.

3. **Source-Level Provenance** Every claim references specific data sources: log entries, API responses, database records. Citations include timestamps, event IDs, and exact locations—making verification straightforward and reproducible.

How Verifiable AI Differs

The distinction between Verifiable AI and previous approaches becomes clear when examining what each system delivers:

Aspect	Traditional AI	Verifiable AI
Primary Output	Prediction/Classification	Evidence Chain + Reasoning
Decision Basis	Model weights	Specific data points
Validation Method	Accuracy metrics over time	Independent evidence verification
Audit Trail	Model version + inputs	Complete source citations
Analyst Role	Accept or reject output	Verify evidence and logic
When It Fails	Unclear why	Can identify exact failure point

The key difference: **Verifiable AI doesn't ask analysts to trust it—it gives them the tools to verify it.**

Technical Foundation

Verifiable AI systems combine advanced language models with structured reasoning frameworks. Unlike traditional AI that outputs conclusions, Verifiable AI constructs explicit reasoning processes:

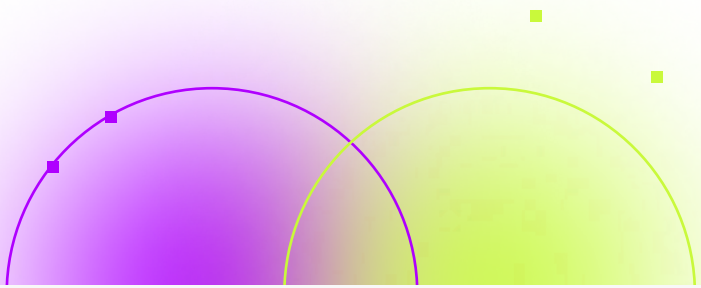
- **Evidence Collection** Queries across multiple data sources (SIEM, EDR, network logs, cloud audit trails) to gather relevant information. Rather than just flagging patterns, the system identifies specific events, logs, and data points.

- **Claim Generation** Makes specific, verifiable statements about what evidence shows. Instead of "unusual behavior detected," it states "User X accessed System Y at time Z from IP address A—the first time this user accessed this system from this location."

- **Source Citation** Links every claim to exact log entries, timestamps, and event IDs. Citations are precise enough that another analyst can pull the same log entry and verify the claim independently.

- **Reasoning Assembly** Connects evidence points into logical chains that follow investigative reasoning: "Because A is true (here's the evidence), and B is true (here's that evidence), and A+B together indicate C, we conclude C."

- **Verification Packaging** Presents findings with verification steps built in. Rather than forcing analysts to figure out how to validate claims, the system provides the roadmap: "To verify this, check log X at timestamp Y."



Why This Matters for Security

Verifiable AI addresses the unique challenges of security operations:

- **Against Adversarial Attacks** Attackers can't game the system by understanding feature weights—they would need to forge actual evidence across multiple independent log sources.

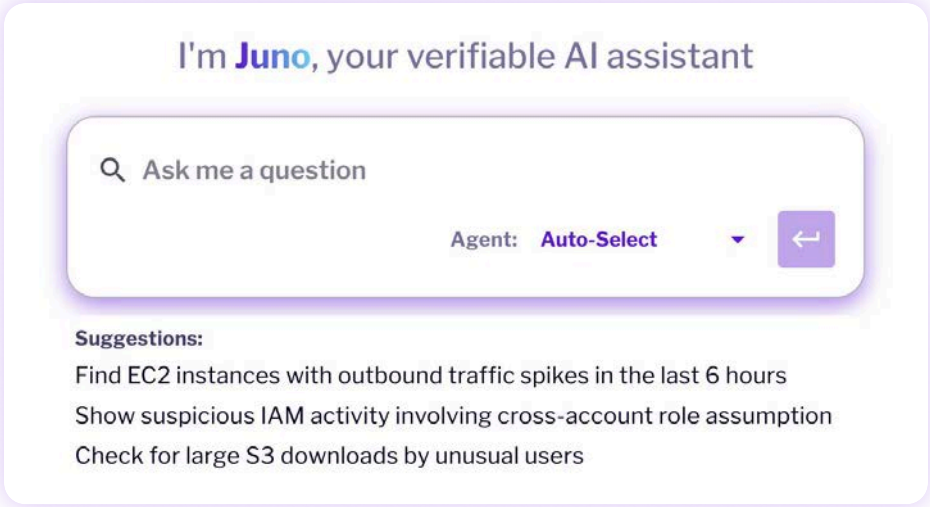
- **For Compliance and Audit** Every AI-driven decision has a complete evidence trail that meets regulatory requirements for automated decision-making.

- **For Team Development** Junior analysts learn investigative reasoning by following verifiable chains. AI becomes a teaching tool, not a black box.

- **For Incident Response** Evidence chains constructed during detection become the foundation of incident reports—no separate documentation effort required.

- **For Continuous Improvement** When the AI makes a mistake, the verifiable reasoning chain shows exactly where it went wrong, enabling precise corrections.

III. Juno: Verifiable AI in Action



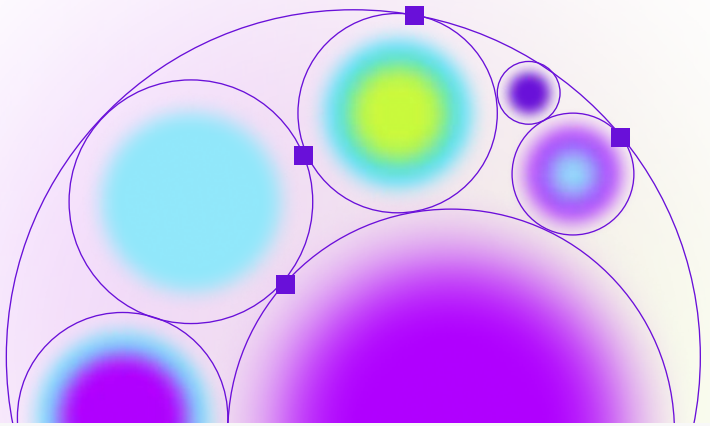
Juno is the first AI Security Analyst built on Verifiable AI principles, proving these concepts work in operational security environments. Running on Uptycs' unified telemetry platform, Juno has access to endpoint, cloud, container, and identity evidence in a single data model—allowing it to build complete, cross-domain reasoning chains.

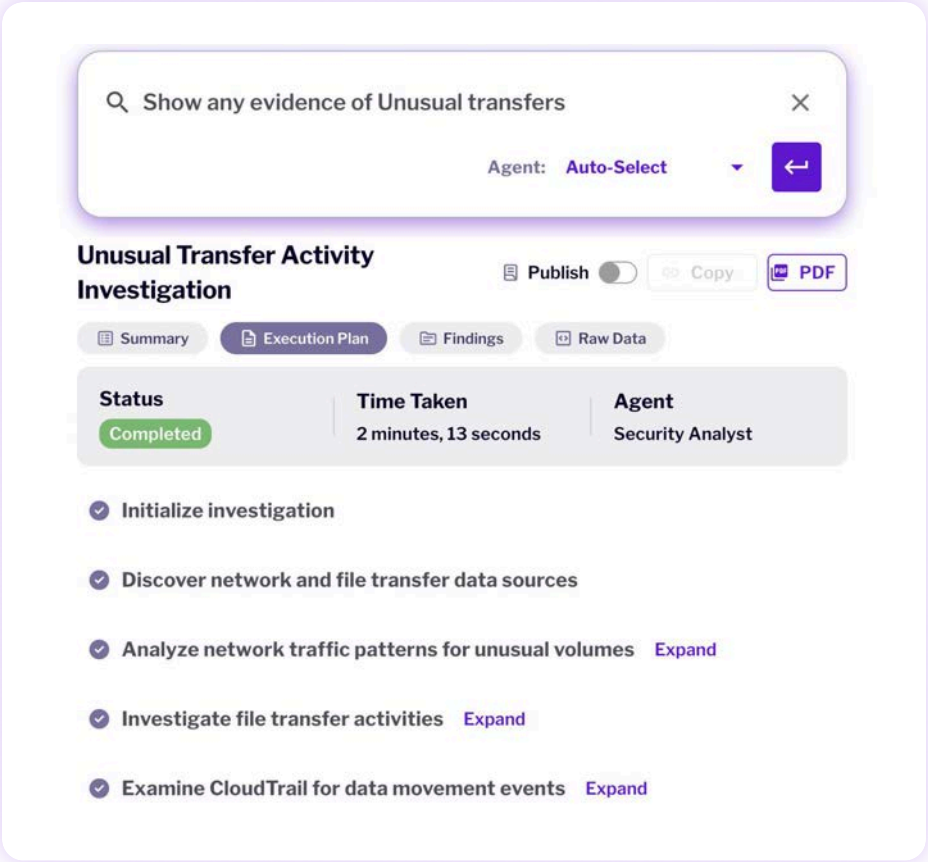
How it works: Juno implements the five-stage process—Evidence Collection, Claim Generation, Source Citation, Reasoning Assembly, and Verification Packaging—across threat investigation, root cause analysis, compliance reporting, and incident response.

A practical example:

When an alert fires for potential data exfiltration, traditional AI provides a risk score (8.7/10) and contributing factors. Juno provides verifiable evidence:

- **Unusual transfer:** sarah.chen@company.com, 2.3 GB to storage.cloudprovider.net (Source: Firewall logs fw-prod-02, entries 194722-194856)
- **Volume anomaly:** 26x above 95th percentile baseline (Source: NetFlow records nf-03)
- **Novel destination:** First-time domain, registered 8 days prior (Source: DNS logs)





Result: Analyst verifies in 15 minutes instead of reconstructing analysis for 60+ minutes. The evidence chain becomes the incident report.

What this demonstrates: Complete evidence chains, audit-ready documentation, transparent reasoning, reproducible investigations, and knowledge transfer—all capabilities that emerge naturally from Verifiable AI's architecture.

Traditional AI asks analysts to trust. Juno gives them tools to verify.

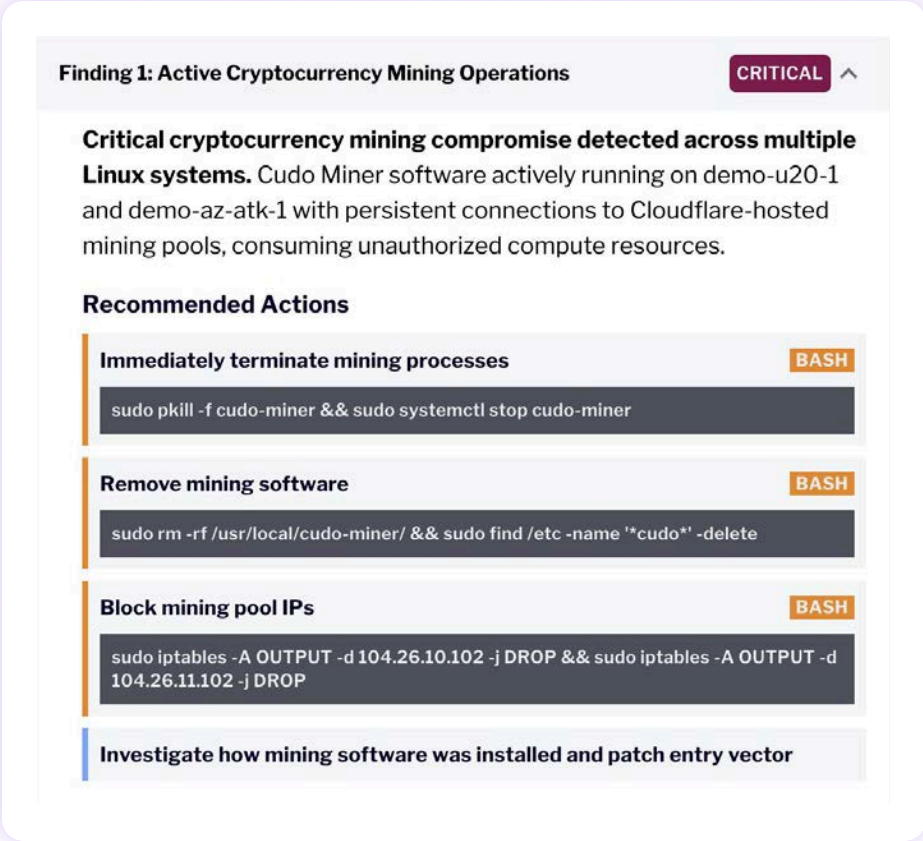
IV. Conclusion: Proof Is the New Intelligence

The evolution of security AI has reached an inflection point.

The first generation of security AI focused on **detection**—finding threats faster than humans could. It delivered value but created new problems: alert fatigue, false positives, and black-box decisions that couldn't be trusted.

The second generation focused on **explanation**—helping humans understand what AI detected. It was an improvement but insufficient. Explanations about how models think don't substitute for evidence about what actually happened.

The third generation is Verifiable AI—systems that provide proof, not just predictions. This is the shift Uptycs has built into its platform and into Juno: AI that shows its work.



This isn't an incremental improvement. It's a fundamental shift in how security teams work with AI:

- From trusting AI to verifying AI
- From accepting conclusions to examining evidence
- From AI as oracle to AI as investigative partner
- From speed versus accuracy to speed with accountability

For CISOs and security leaders, the implications are strategic:

- **Operationally:** Verifiable AI enables faster investigation without sacrificing rigor. Teams move quickly because they can verify quickly, not because they're skipping verification.
- **Organizationally:** Junior analysts learn from verifiable reasoning chains. Institutional knowledge accumulates rather than concentrating in senior team members. Turnover becomes less disruptive.
- **Compliance-wise:** Audit-ready documentation is a byproduct of investigation, not a separate task. Regulatory requirements for AI transparency and accountability become manageable.
- **Competitively:** Organizations that can investigate thoroughly and quickly gain advantage over those forced to choose between speed and accuracy. Verifiable AI resolves that tension—and Uptycs delivers it at platform scale.

The Standard Has Changed

Five years ago, the question was: "Can AI detect threats?" Three years ago, it became: "Can AI explain its detections?" Today, the question is: "Can we verify AI's reasoning?"

Organizations still asking the first two questions risk falling behind. The market is moving toward verifiability—driven by compliance requirements, adversarial sophistication, and security teams who refuse to work blind.

The choice isn't whether to adopt AI in security operations.This is no longer optional.

The choice is whether to adopt AI you can trust because it shows its work—or to continue managing AI you must trust blindly because it doesn't.

Verifiable AI, exemplified by Juno and enabled by Uptycs' unified security platform, represents a new standard: AI that respects analyst expertise, enables independent verification, and produces audit-ready documentation as a natural byproduct of investigation.

In security, proof has always mattered more than persuasion. It's time for AI to meet that standard.

References

1. "The State of AI in the SOC 2025," The Hacker News, September 2025. Survey of 282 security leaders showing average of 960 alerts per day, with large enterprises handling 3,000+ daily alerts. https://thehackernews.com/2025/09/the-state-of-ai-in-soc-2025-insights.html

2. "Osterman Report 2024: SOC Trends, Challenges, and Solutions," Dropzone AI and Osterman Research, 2024. Survey of 125 SOC professionals showing 97.6% of organizations report increasing daily alerts, with rising backlogs of uninvestigated alerts. https://www.dropzone.ai/blog/osterman-report-2024-soc-trends-challenges-and-solutions

3. "Alert Fatigue in Security Operations Centres: Research Challenges and Opportunities," ACM Computing Surveys, 2024. Trend Micro survey finding 51% of SOC teams feel overwhelmed by alert volume, with analysts spending over 25% of their time handling false positives. https://dl.acm.org/doi/10.1145/3723158

4. "Global Security Operations Center Study," IBM and Morning Consult, 2023. Survey of 1,000 SOC members finding that one-third of analysts' workday is spent on incidents that are not real threats, with false positive and low-priority alerts comprising roughly 63% of daily alerts. Referenced in: https://panther.com/blog/identifying-and-mitigating-false-positive-alerts

5. "SANS 2024 SOC Survey: Facing Top Challenges in Security Operations," SANS Institute, July 2024. Comprehensive survey examining SOC architecture, staffing challenges, and operational efficiency improvements through hyperautomation. https://www.sans.org/white-papers/sans-2024-soc-survey-facing-top-challenges-security-operations

6. "Navigating the AI Black Box Problem," Gibraltar Solutions, March 2025. Analysis of trust, accountability, and transparency challenges with AI systems in cybersecurity, including issues with data poisoning and adversarial attacks. https://gibraltarsolutions.com/blog/navigating-the-ai-black-box-problem/

7. "AI's Black Box Problem: When Security Fixes Fall Short," Bank Info Security (Cobalt's State of Pentesting Report 2025), June 2025. Report finding organizations can fix only 21% of generative AI vulnerabilities, highlighting critical transparency and trust gaps in AI security systems. https://www.bankinfosecurity.com/ais-black-box-problem-when-security-fixes-fall-short-a-28707

8. "Artificial Intelligence in fraud detection: Revolutionizing financial security," International Journal of Science and Research Archive, October 2024. Research on explainability challenges in AI fraud detection, including black-box limitations, false positives, and trust issues with model transparency. https://www.researchgate.net/publication/384606692_Artificial_Intelligence_in_fraud_detection_Revolutionizing_financial_security

9. "99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms," USENIX Security Symposium, 2022. Academic research on false positive prevalence in SOC operations. https://www.usenix.org/conference/usenixsecurity22/presentation/alahmadi

uptycs

Uptycs is the leading cloud security platform for large hybrid cloud environments. We extend security visibility from development to runtime, ensuring consistent protection and compliance across the application infrastructure. That's why enterprises like PayPal, Comcast, and Nutanix rely on Uptycs to secure the development ecosystems they use to build their applications and run their workloads.

Learn more at Uptycs.com