2026 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
SCIENCE, TECHNOLOGY & ENGINEERING, ARTS, MATHEMATICS & EDUCATION JUNE 09 - 11, 2026
PRINCE WAIKIKI RESORT, HONOLULU, HAWAII

AINTERNSHIP: THE SUPERVISED INTERN MODEL FOR ENTERPRISE AI AND CYBERSECURITY OPERATIONS

HERTLE, DOUGLAS
SIMVENTIONS
FREDERICKSBURG, VIRGINIA

AInternship: The Supervised Intern Model for Enterprise AI and Cybersecurity Operations

A Phased Implementation Framework Aligned with NIST AI Risk Management

Keywords: AI adoption, human-in-the-loop, enterprise AI, cybersecurity, log analysis, NIST AI RMF, AI governance, workforce transformation, alert fatigue

Abstract

Enterprises increasingly rely on AI, but most fail to capture its value. Human analysts drown in billions of log entries, missing critical threats while attackers lurk undetected for an average of 204 days. Inspired by Ethan Mollick's metaphor of AI as an "intern," this paper introduces the Supervised Intern Model (SIM): a phased, NIST-aligned conceptual framework that turns AI-human collaboration into practical, scalable enterprise processes.

SIM addresses a stubborn reality: 78% of organizations now use AI, yet only 26% generate tangible business value from it. The problem isn't the technology, 70% of implementation failures stem from people and process issues. SIM solves this by prescribing three deployment phases, Observe, Collaborate, and Autonomy with Oversight, each with explicit transition criteria and governance checkpoints. This paper presents the conceptual framework and illustrates its application to cybersecurity, where AI can address the cognitive limitations that create exploitable detection gaps.

1. Introduction

In May 2023, Wharton professor Ethan Mollick published an essay that changed how millions of knowledge workers think about AI. Instead of viewing it as inscrutable software or autonomous agent, he proposed treating it as "a weird, somewhat alien intern that works infinitely fast and sometimes lies to make you happy" (Mollick, 2023).

The metaphor stuck. It gave people permission to experiment, make mistakes, and learn. It acknowledged that AI is powerful but imperfect, just like a new hire who needs guidance.

But here's the problem: what works for one person doesn't automatically work for an entire organization. While individuals figure out their own AI workflows through trial and error, enterprises need standardized processes, governance structures, and change management protocols that scale across thousands of employees.

The numbers tell a stark story. According to the Stanford AI Index (2025), 78% of organizations now use AI in at least one business function. Yet BCG (2024) found that only 26% have developed the capabilities to move beyond pilot programs and generate real value. Even more

telling: 70% of implementation challenges stem from people and process issues, not algorithmic ones.

This paper bridges that gap with the **Supervised Intern Model (SIM)**, a conceptual enterprise framework that operationalizes Mollick's insight. SIM makes three contributions:

First, it provides a phased deployment approach (Observe → Collaborate → Autonomy with Oversight) with concrete milestones and clear transition criteria. Second, it aligns explicitly with the NIST AI Risk Management Framework, connecting the intern metaphor to regulatory compliance. Third, it applies the framework to cybersecurity, a domain where human cognitive limits create exploitable vulnerabilities that AI supervision can address.

2. Theoretical Foundation: From Metaphor to Framework

2.1 Why the Intern Metaphor Works

Mollick's insight grew from a practical observation: AI systems have what he calls a "jagged frontier" of capabilities. They excel unexpectedly at some tasks while failing spectacularly at others that seem simpler. Sound familiar? It's exactly what happens when you supervise a new hire, enthusiasm and raw ability, but unpredictable judgment.

In his 2024 book Co-Intelligence, Mollick distilled this into four principles: always invite AI to the table, be the human in the loop, treat AI like a person, and assume this is the worst AI you'll ever use. For individuals, these principles work beautifully.

But Mollick's advice assumes a single user working with a single AI system. "Figure out your intern's job description," he suggests. "Learn its strengths and weaknesses." That's fine for personal productivity. Enterprise deployment requires something different entirely.

2.2 The Enterprise Gap

Imagine your company of 5,000 employees just received 5,000 interns overnight, interns that each employee struggles to understand and control. That's the enterprise AI challenge in a nutshell. You can't rely on each person individually "figuring out" their AI intern through trial and error. The organization needs standardized processes, defined accountability, and governance that works at scale.

Recent studies show the disconnect clearly. The Stanford AI Index (2025) reports that 71% of organizations now use generative AI regularly, up from just 33% in 2023. But according to ISG (2025), only 31% of AI use cases reach full production, and Deloitte (2026) found that just one in five companies has mature governance for AI agents.

The pattern is unmistakable: enterprises adopt AI eagerly but struggle to operationalize it. They need implementation infrastructure, not just a compelling metaphor. SIM provides that infrastructure while preserving the intuitive power of Mollick's original insight.

3. The Supervised Intern Model

SIM operationalizes the AI-as-intern metaphor through three components: core principles adapted for organizations, a phased deployment model with explicit transition criteria, and governance aligned with NIST's AI Risk Management Framework.

The NIST alignment is deliberate and strategic. Organizations have trusted NIST frameworks for decades to protect critical infrastructure, secure federal systems, and manage cybersecurity risk. If we trust NIST to safeguard human safety, why not leverage that same proven approach to govern AI?

3.1 Core Principles

Oversight: Mollick tells individuals to "be the human in the loop." For enterprises, this means systematic oversight structures, defined roles, review processes, escalation pathways. The 76% of enterprises now implementing human-in-the-loop processes need standardized protocols, not individual discretion.

Understanding: "Learn your intern's capabilities" becomes documented capability assessments that inform decisions organization-wide. With 77% of businesses worried about AI hallucinations, this understanding can't be left to individual discovery.

Gradual Delegation: "Start with low-stakes tasks" becomes a structured progression with explicit criteria. BCG (2024) found that AI leaders scale twice as many products as laggards, not through faster deployment, but through more disciplined progression.

3.2 The Three Phases

SIM prescribes three phases that mirror traditional employee onboarding, but with governance checkpoints built in:

Phase 1: Observe. AI watches and learns. Humans check every output. The AI drafts, flags, and analyzes, but nothing goes live without human validation. This phase maps to NIST's MAP function, identifying risks and documenting capability boundaries. Transition criteria: documented accuracy rates and identified failure modes.

Phase 2: Collaborate. AI proposes actions. Humans guide, correct, and teach. The AI suggests decisions and summarizes findings; humans approve, modify, or reject. This phase implements NIST's MEASURE function through structured feedback loops and performance metrics. Transition criteria: stable performance and defined override protocols.

Phase 3: Autonomy with Oversight. AI acts within safe boundaries. Humans intervene only when needed. The AI executes routine tasks and escalates edge cases; humans monitor exceptions and audit samples. This embodies NIST's MANAGE function, ongoing governance without eliminating human oversight. Transition criteria: continuous monitoring and tested incident response.

Phase	AI Role	Human Role	NIST Alignment	Transition Criteria
Observe	Drafts, flags, analyzes; all outputs reviewed	Validates 100% of outputs; documents boundaries	MAP: Context & risk identification	Accuracy documented; failure modes identified
Collaborate	Suggests actions, summarizes, proposes decisions	Approves, modifies, rejects; provides feedback	MEASURE: Metrics & monitoring	Stable performance; override protocols defined
Autonomy with Oversight	Executes within boundaries; escalates edge cases	Monitors exceptions; audits samples; refines	MANAGE: Ongoing governance	Continuous monitoring; incident response tested

Without defined milestones, organizations either get stuck in perpetual piloting or rush toward autonomy without safeguards. SIM provides the structure to progress deliberately.

With these principles in place, let's see how SIM operates in a real security operations center, where the stakes for human-AI collaboration couldn't be higher.

4. The Cybersecurity Imperative

An analyst scans thousands of log entries each day. Her eyes glaze over as the same patterns repeat: successful logins, routine file access, standard network traffic. Somewhere in the stream, a single duplicated login slips by, a user authenticated from two different network segments within seven minutes. It's a textbook sign of credential compromise.

She never sees it. Neither do her colleagues on the next two shifts.

Three weeks later, the attacker has mapped the entire network.

4.1 The Scale Problem

This isn't a failure of skill or dedication. It's a failure of biology. Modern enterprise environments generate staggering volumes of log data, tens of thousands of events per second, billions of entries per day. Recent research indicates that most organizations receive over 10,000 alerts daily, with more than 50% being false positives (Jalalvand et al., 2024). Industry data shows that 25–30% of alerts go uninvestigated due to overload (Cymulate, 2025).

Meanwhile, according to IBM (2024), threat actors enjoy an average dwell time of 204 days before detection. That's nearly seven months of reconnaissance, lateral movement, and data exfiltration, all while analysts struggle to separate signal from noise.

No human team, regardless of size or expertise, can meaningfully review this volume. Analysts rely on automated filtering just to reduce the firehose to a manageable stream. But even that filtered stream overwhelms human attention.

4.2 Why Humans Miss What Matters

Three cognitive limitations create exploitable gaps:

Alert Fatigue. After reviewing hundreds of similar entries, analysts develop "alert blindness", repetitive stimuli stop registering as meaningful. More than one-third of IT security managers and analysts ignore threat alerts when the queue is full (IDC/FireEye, 2021). The Verizon 2023 DBIR found that 74% of breaches involved human error, with many attributed to missed alerts that were technically visible but cognitively invisible.

Pattern Normalization. Humans naturally build mental models of "normal" and unconsciously filter out matching data. Sophisticated attackers exploit this by mimicking legitimate traffic, conducting reconnaissance during high-volume periods, and spacing malicious activities to avoid clustering.

Attention Decay. Research shows significant performance degradation after 20-30 minutes of monotonous monitoring. Accuracy declines within the first hour of a shift, with error rates spiking during overnight and weekend coverage.

4.3 What the Human Eye Missed

Consider these eight authentication log entries from a corporate Active Directory environment, routine Monday morning logins:

Timestamp	Event ID	User	Source IP	Status
08:47:23	4624	jsmith	10.0.15.142	Success
08:51:07	4624	mwilliams	10.0.15.89	Success
08:52:34	4624	agarcia	10.0.22.201	Success

08:54:19	4624	klee	10.0.15.203	Success
08:54:22	4624	jsmith	10.0.22.47	Success
08:55:41	4624	rpatel	10.0.15.167	Success
08:57:03	4624	tchen	10.0.15.88	Success
08:58:56	4624	lnguyen	10.0.22.156	Success

At first glance: unremarkable. Successful logins, standard usernames, internal IP addresses, morning timestamps. Everything looks normal.

Look again. User "jsmith" appears twice, at 08:47:23 from 10.0.15.142, and at 08:54:22 from 10.0.22.47. **That's a red flag.** The same user authenticated from two different network segments within seven minutes. Unless jsmith teleported between buildings, this pattern suggests compromised credentials or a pass-the-hash attack.

Did you catch it on first read? Most people don't. When presented with routine-looking data, human cognition seeks confirmation of normalcy rather than hunting for exceptions. We process the visual pattern ("these all look like successful logins"), not the semantic content ("wait, that username appeared before").

An AI system would flag this instantly. Every time. Regardless of how many thousands of similar tables it had processed that day. AI doesn't get tired. It doesn't normalize patterns. It catches what human cognition inevitably misses.

4.4 The Compounding Effect

The jsmith anomaly appeared in just eight entries. Real analysts review thousands per hour across multiple consoles while juggling tickets, meetings, and alert queues.

Overlooked anomalies compound: a missed credential compromise enables lateral movement. Undetected reconnaissance in week one enables targeted attacks in week three. Each unnoticed anomaly becomes part of the analyst's mental model of "normal," making future detection even harder. Attackers deliberately space activities across shifts and days to exploit exactly this pattern.

This is why SIM applied to cybersecurity is so powerful: AI extends human attention without replacing human judgment.

5. Applying SIM to Security Operations

With SIM's principles established, here's how the three phases work in a real security operations center.

5.1 Observe Phase

AI monitors log streams and flags potential anomalies for human review. It doesn't take action or prioritize, it simply surfaces candidates for analyst attention.

During this phase, analysts learn the AI's detection patterns: which anomalies it catches reliably, which it misses. The organization documents false positive rates and establishes baseline metrics. Trust builds incrementally as analysts verify the AI catches things they would have missed, like the jsmith duplicate. Edge cases get identified before the AI gains any decision authority.

5.2 Collaborate Phase

As reliability metrics justify progression, the AI starts proposing severity classifications and response priorities. Analysts retain approval authority but no longer start from raw logs.

The AI presents prioritized alert queues with preliminary triage. Analysts approve, modify, or reject, with structured feedback captured for model refinement. Override patterns create an audit trail. Escalation pathways get tested for low-confidence cases.

5.3 Autonomy with Oversight Phase

AI handles routine triage within defined parameters. Analysts focus on complex threats, strategic assessment, and exception handling, higher-value work than first-pass log review.

Low-confidence alerts automatically route to human review. Random sample audits verify ongoing accuracy. Override capability remains immediately available for emerging threat patterns. Human expertise concentrates on threat hunting, incident response, and strategic improvements.

The expected result: reduced alert backlog, faster response times, and analysts freed from repetitive triage to focus on complex investigations. Early case studies of AI-augmented SOCs report significant improvements, one implementation achieved a 34% reduction in investigation time and 75% improvement in response times (Elastic, 2025).

6. Addressing the Human Factor

BCG's finding that 70% of AI implementation challenges stem from people and process, not technology, demands frameworks that explicitly address human factors.

6.1 Building Trust Through Progression

Employee resistance to AI typically stems from two fears: loss of control and unclear accountability. SIM's phased model addresses both.

In Observe, employees experience AI as a subordinate tool, it generates options, they evaluate. Agency preserved, familiarity built. In Collaborate, employees develop judgment about when to accept, modify, or reject recommendations, supervisory skills for their evolving role. In Autonomy with Oversight, they transition to exception handling and quality assurance, higher-value work that leverages AI rather than competing with it.

This mirrors how trust develops with human interns: not through immediate delegation, but through demonstrated reliability in progressively complex tasks.

6.2 Reframing the Error Conversation

The AI discourse obsesses over AI errors while ignoring substantial human error rates. Research documents significant human limitations in high-volume processing tasks:

Task Domain	Human-Only Challenge	AI + Human Oversight
Security alert triage	25-30% of alerts uninvestigated; >50% false positives overwhelm analysts	Automated filtering reduces noise; humans focus on verified threats
Data entry/processing	1-4% error rate typical for manual entry (Garza et al., 2024)	Automated validation catches errors before propagation
Pattern detection in logs	Cognitive fatigue after 20-30 min; anomalies blend into noise	Consistent detection regardless of volume or time-on-task

The insight: SIM doesn't promise perfection, it makes errors visible and manageable. Just as organizations expect and manage intern mistakes, SIM creates structures for identifying, correcting, and learning from AI errors while capturing the benefits of tireless, consistent processing.

7. Learning from Automation History

Today's AI anxiety echoes historical resistance to technologies now considered unremarkable:

Technology	Initial Fears	Today
Automated doors	Injury, malfunction	Standard everywhere

ATMs	"People won't trust machines with money"	Primary banking interface
Elevators	"Unsafe without operators"	Trusted globally

In each case, trust emerged not from perfection but from visibility, control, and the ability to intervene. SIM replicates this pattern: observation builds familiarity, collaboration establishes control, supervised autonomy maintains intervention capability.

8. Workforce Transformation

SIM carries significant workforce implications. Workers with AI skills now command a 43% wage premium, up from 25% in 2023. The AI skills gap is viewed as the biggest integration barrier (Deloitte, 2026). Most workers believe AI will change their jobs, though fewer actually fear replacement.

SIM provides a practical evolution pathway: employees develop supervision competencies in Observe, refine judgment and feedback skills in Collaborate, and transition to exception handling in Autonomy. For security operations, this means analysts evolve from log reviewers to threat hunters and strategic defenders, roles that leverage AI rather than competing with it.

9. Limitations and Future Work

This paper presents SIM as a conceptual framework derived from established principles in AI governance, organizational change management, and cybersecurity operations. Several limitations should be acknowledged.

Empirical Validation. SIM has not yet been tested through controlled implementation studies. While the framework draws on documented industry challenges and established best practices, its effectiveness requires validation through pilot deployments across diverse organizational contexts. Future research should examine implementation outcomes, measure transition criteria reliability, and assess workforce adaptation patterns.

Context Dependency. The framework assumes organizations have baseline AI capabilities and governance structures. Smaller organizations or those in regulated industries may require modified approaches. The cybersecurity application, while illustrative, represents one domain; other applications (legal review, medical diagnosis support, financial analysis) may require domain-specific adaptations.

Measurement Challenges. Defining precise transition criteria between phases remains an implementation challenge. Organizations will need to develop context-specific metrics for

accuracy thresholds, trust indicators, and governance checkpoints. Future work should develop standardized assessment instruments.

Longitudinal Effects. The long-term effects of AI-human collaboration structures on workforce skills, job satisfaction, and organizational culture require extended study. SIM's phased approach is designed to support positive adaptation, but outcomes may vary based on implementation fidelity and organizational factors.

Despite these limitations, SIM provides a structured starting point for organizations seeking to operationalize human-AI collaboration. The framework's alignment with NIST AI RMF offers regulatory grounding, while the intern metaphor provides intuitive guidance for practitioners navigating AI integration.

10. Conclusion

Ethan Mollick gave us a powerful metaphor: treat AI like an intern. The Supervised Intern Model gives enterprises a conceptual framework to make that metaphor operational, phased deployment with explicit criteria, NIST-aligned governance, and change management that addresses the human factors behind most AI failures.

The cybersecurity application shows why this matters. Human analysts face cognitive limits that create exploitable detection gaps. That overlooked jsmith duplicate? AI would have caught it instantly. But AI needs the structured oversight and graduated autonomy that SIM prescribes. Neither human-only nor AI-only approaches suffice; supervised collaboration captures the strengths of both.

SIM offers enterprises a structured path to scale AI safely, extend human attention, and address cognitive limits through human-AI collaboration. The AI intern doesn't replace humans, it makes them more effective.

Organizations that implement this framework position themselves to capture AI's substantial value while maintaining accountability, workforce trust, and regulatory readiness. The question is no longer whether to adopt AI, but how to do it systematically. SIM provides a foundation for that systematic approach, with empirical validation as an important next step.

References

Jalalvand, F., Chhetri, M. B., Nepal, S., & Paris, C. (2024). Alert Prioritisation in Security Operations Centres: A Systematic Survey on Criteria and Methods. ACM Computing Surveys, 57(2), Article 42. https://dl.acm.org/doi/10.1145/3695462

Boston Consulting Group. (2024). Where's the Value in AI? AI Adoption in 2024. BCG Press. https://www.bcg.com/press/24october2024-ai-adoption-in-2024-74-of-companies-struggle-to-achieve-and-scale-value

Cymulate. (2025). Alert Fatigue: How to Fix SOC Overload. https://cymulate.com/cybersecurity-glossary/alert-fatigue/

Deloitte. (2026). State of AI in the Enterprise: The Untapped Edge. Deloitte AI Institute. https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-generative-ai-in-enterprise.html

Elastic. (2025). AI and the 2025 SIEM landscape. Elastic Blog. https://www.elastic.co/blog/ai-2025-siem-landscape

IBM Security. (2024). Cost of a Data Breach Report 2024. IBM Corporation. https://www.ibm.com/reports/data-breach

IDC & FireEye. (2021). The State of Security Operations. Cybersecurity Dive. https://www.cybersecuritydive.com/news/security-alert-analyst-SOC-idc-fireeye/595111/

Information Services Group. (2025). State of Enterprise AI Adoption Report 2025. ISG. https://isg-one.com/state-of-enterprise-ai-adoption-report-2025

Garza, M. Y., Williams, T., Ounpraseuth, S., Hu, Z., Lee, J., Snowden, J., Walden, A. C., Simon, A. E., Devlin, L. A., Young, L. W., & Zozus, M. N. (2024). Error Rates of Data Processing Methods in Clinical Research: A Systematic Review and Meta-Analysis. PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC10775420/

Mollick, E. (2023, May 20). On-boarding your AI Intern. One Useful Thing. https://www.oneusefulthing.org/p/on-boarding-your-ai-intern

Mollick, E. (2024). Co-Intelligence: Living and Working with AI. Portfolio/Penguin.

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://www.nist.gov/itl/ai-risk-management-framework

National Institute of Standards and Technology. (2024). AI RMF: Generative Artificial Intelligence Profile. NIST AI 600-1. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

Stanford Institute for Human-Centered Artificial Intelligence. (2025). AI Index Report 2025. Stanford University. https://hai.stanford.edu/ai-index/2025-ai-index-report

Verizon. (2023). Data Breach Investigations Report. Verizon Communications. https://www.verizon.com/business/resources/reports/dbir/