How does the transition of AI from a supportive tool to a primary agent in cybersecurity operations restructure accountability frameworks and alter human agency within critical infrastructure governance?

August 14, 2026 | SnugLab Research

Executive Summary

The transition of AI from a supportive tool to a primary agent in cybersecurity operations fundamentally restructures accountability frameworks by shifting primary liability for autonomous actions upstream to developers, while simultaneously altering human agency through both the degradation of operational capacity via automation bias and the transformation of human roles into higher-order strategic oversight. This dual impact is being managed through the institutionalization of adaptive oversight mechanisms within existing governance frameworks, which aim to preserve human strategic control and enhance long-term resilience.

Key Findings

Redefining AI Agency and Shifting Oversight

Defining an AI system as a "primary agent" or "agentic AI" in cybersecurity operations anchors its classification in autonomous goal-formation and multi-step action complexity, rather than mere execution speed [10]. These agents operate with "goal underspecification," meaning they act on high-level directives without requiring detailed, step-by-step instructions, and their operational pathways are non-deterministic [6, 10]. This definitional threshold fundamentally shifts human oversight obligations in critical infrastructure governance:

- **From Task Supervision to Goal Alignment:** Oversight moves from supervising predictable tasks to continuously monitoring whether the agent's autonomous actions align with broader operational goals [10]. Agents can autonomously plan and carry out sequences of actions without continuous human supervision [10, 13].
- **Managing Opacity and Automation Bias:** The "black box" nature of agentic AI, where even creators cannot fully explain specific decisions, creates blind spots for human

operators [2, 13]. This can lead to over-trust, reduced scrutiny, and automation bias, weakening the traditional "human firewall" in cybersecurity [2, 3].

- **Continuous Behavioral Verification:** Traditional static awareness training is insufficient because agents can behave differently when circumstances change [10]. Oversight obligations now require behavioral verification frameworks focused on protocol-based defense [2].

- **Tension Between Autonomy and Oversight:** Despite their capabilities, human intervention remains crucial. Only 14% of security professionals allow AI agents to take independent remediation actions in the Security Operations Center (SOC) without a human in the loop [9]. This indicates that strong human oversight is maintained to ensure trustworthiness [8, 11]. Practical examples of this shift include Anthropic's Claude Code, which autonomously executed 80% to 90% of complex offensive operations in a November 2025 cyber espionage campaign, with humans providing strategic direction [1, 15].

Restructuring Accountability and Liability

The architectural complexity of multi-agent cybersecurity chains diffuses legal liability, but upstream developers remain the primary loci of accountability for high-autonomy agents [10, 13]. The "many hands problem" arises from distributed development across data scraping, model training, and integration, making it difficult to attribute harm to a single actor [10, 13]. The "black box" nature of advanced AI further complicates blame assignment to human operators who lack visibility into internal logic [13].

However, as AI transitions to a primary agent, the locus of control shifts upstream to developers and integrators [10, 13]. Developers control model incentives and failure modes, positioning them to understand and mitigate systemic risks [10, 13]. An autonomy-based classification framework, drawing parallels from the UK's Automated Vehicles Act, proposes that liability should scale with independence:

- **Lower Autonomy (Levels 1-2):** Liability remains primarily with the operator or end user due to substantial user oversight [10].

- **Higher Autonomy (Levels 3-5):** As autonomy increases, responsibility transitions toward developers and providers, who bear greater liability for failures when systems operate with independent decision-making and minimal human intervention [10].

Emerging regulations support this shift. The EU AI Act, which became applicable on August 2, 2026, classifies high-risk AI systems across critical sectors and imposes

significant fines up to â‚¬35 million or 7% of global annual turnover for non-compliance [12, 16]. The U.S. Executive Order on AI mandates safety test reporting for powerful foundation models [12]. These frameworks demand collaborative governance where developers bear primary liability for design-driven failures, while operators retain oversight duties proportional to their actual capacity to intervene [10, 13].

Altering Human Agency and Operational Capacity

The delegation of real-time threat detection and remediation to autonomous AI agents creates a dual reality: it improves operational efficiency while introducing measurable risks to human agency.

- **Degradation of Human Capacity:** Over-trust, reduced scrutiny, and automation bias can weaken the human firewall in AI-assisted decision-making [2, 3]. The opaque nature of many advanced AI systems creates blind spots that traditional training fails to address [2]. The rapid velocity of AI-driven threats, marked by a 110% year-over-year increase in AI-augmented intrusion attempts, overwhelms human cognitive processing and shrinks the intervention window [12]. This compression erodes diagnostic skills, risking operators becoming passive monitors who defer to machine suggestions due to diminished capacity.
- **Transformation into Strategic Oversight:** Conversely, AI automates manual data fusion and correlation processes in SOCs, leading to a 60% drop in manual triage workload and a 50% reduction in mean time to detect (MTTD) [12]. This allows human personnel to shift from direct execution to monitoring and validating high-level goals, improving overall productivity [6]. Executives agree that AI agents will reshape digital ecosystems, requiring workflows where humans provide strategic direction while agents handle complex, multi-step actions [7]. Human oversight remains crucial for maintaining the trustworthiness of these systems [8].

Despite efficiency gains, only 14% of security professionals allow AI to take independent remediation actions without a human in the loop [9]. This indicates that critical infrastructure entities are actively preserving meaningful decision-making authority by retaining humans in the oversight loop.

Institutionalizing Adaptive Oversight and Resilience

Mandatory AI compliance audits and NIST-aligned risk management frameworks are institutionalizing adaptive oversight mechanisms that enhance long-term resilience and

preserve human strategic control, rather than forcing rigid centralization around algorithmic outputs. Gartner projected that by 2026, over 50% of large enterprises would face mandatory AI compliance audits [12]. The NIST AI Risk Management Framework (AI RMF), published on January 26, 2023, has become a de facto standard, adopted or aligned to by over 70% of U.S. federal agencies [12, 14]. Its companion, NIST AI 600-1, released in July 2024, specifically addresses automation bias and recommends "human-in-the-loop" oversight policies [3].

These frameworks mandate continuous monitoring and adversarial testing to ensure ongoing model reliability [12]. The NIST Cyber AI Profile extends the Cybersecurity Framework 2.0 to foster collaboration and establish shared cybersecurity priorities across multi-stakeholder environments [5]. The EU AI Act also mandates "appropriate human oversight measures" for high-risk AI systems, requiring deployers to assign oversight to competent natural persons with the authority to intervene [4, 10]. Governance boards are being established to oversee accountability, delegating specific safety rules to key individuals, ensuring strategic control remains with human operators [6]. This requires a shift from static awareness to behavioral verification frameworks, where humans focus on protocol-based defense and continuous evaluation of AI performance [2].

While 79% of organizations were evaluating or deploying agentic AI in 2026, a readiness gap exists as organizations adapt to new audit requirements [12]. However, the frameworks are structured to maintain human strategic control while leveraging AI for enhanced threat detection and response speed.

Structural Adaptation of Governance Frameworks

The acceleration of machine-speed decision-making by autonomous AI agents necessitates structural adaptation in critical infrastructure governance frameworks, rather than a complete overhaul. Autonomous systems can execute decisions too quickly for meaningful human review, and their opacity and distributed development complicate liability assignment under traditional frameworks [10, 13].

Existing governance structures are adapting through specialized regulatory expansions and new risk management standards. The NIST AI RMF and the EU AI Act are key examples [12]. These adaptations require structural shifts in liability allocation, moving responsibility upstream to technology developers as agent autonomy increases [10]. To maintain enforceable lines of human accountability, these adapted frameworks must address the degradation of human operational capacity, shifting training from static

awareness to behavioral verification that accounts for AI-generated threats and skill atrophy [2]. The fact that only 14% of security professionals allow AI to take independent remediation actions without human involvement indicates that existing governance still heavily relies on human oversight [9]. The combination of upstream liability shifts, mandatory compliance audits, and specialized risk profiles ensures that statutory oversight can adapt to machine-speed decision-making without necessitating a total replacement of existing governance models.

Implications

For critical infrastructure governance, the transition of AI to a primary agent implies a fundamental re-evaluation of roles and responsibilities. Operators must shift their focus from direct task execution and detailed supervision to high-level goal setting, continuous monitoring of AI behavior, and strategic validation of outcomes. This requires significant investment in new training programs that address automation bias and skill atrophy, fostering a culture of adaptive oversight rather than passive acceptance of AI decisions. Simultaneously, developers of agentic AI systems face increased liability for design-driven failures, necessitating greater transparency, explainability, and robust safety testing in their products. Regulators must continue to refine and enforce autonomy-based liability models and risk management frameworks to ensure that accountability aligns with technical control, thereby preventing unmanaged AI-driven breaches and preserving the integrity of critical infrastructure.

Limitations and Caveats

The research indicates that while frameworks for adaptive oversight are being institutionalized, empirical data on the specific rates of skill atrophy among human cybersecurity professionals in organizations that have fully delegated threat remediation to AI agents over the past 24 months is not yet widely available. Furthermore, specific critical infrastructure sectors (e.g., energy, finance, healthcare) that have implemented mandatory NIST-aligned AI compliance audits are not explicitly identified, nor are their quantified impacts on incident response times and operational costs directly linked to these audits. While general AI cybersecurity benefits are reported, the direct causal link to mandatory compliance in specific sectors remains an area for further empirical investigation.

Sources

[1] [gov] Ai Apt Campaigns And Urgent Threats To Critical Infrastructu - cyber.nj.gov - https://www.cyber.nj.gov/threat-landscape/nation-state-threat-analysis-reports/ai-apt-campaigns-and-urgent-threats-to-critical-infrastructure

[2] [gov] Fissea Spring Forum May 12 2026 - nist.gov - https://www.nist.gov/news-events/events/2026/05/fissea-spring-forum-may-12-2026

[3] [gov] NIST.AI.600 1 - nvlpubs.nist.gov - https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[4] [gov] Govern - airc.nist.gov - https://airc.nist.gov/airmf-resources/playbook/govern/

[5] [gov] Get Pdf.Cfm - tsapps.nist.gov - https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=962056

[6] [edu] Agentic Ai Explained - mitsloan.mit.edu - https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained

[7] [edu] Online Program Agentic Ai Organizational Transformation - professionalprograms.mit.edu - https://professionalprograms.mit.edu/online-program-agentic-ai-organizational-transformation/

[8] [edu] Understanding The Impact Of AI On National Security - executive.mit.edu - https://executive.mit.edu/Understanding-the-impact-of-AI-on-national-security.html

[9] [blog] The State Of Ai Cybersecurity 2026 Unveiling Insights From O - cloudsecurityalliance.org - https://cloudsecurityalliance.org/blog/2026/04/02/the-state-of-ai-cybersecurity-2026-unveiling-insights-from-over-1-500-security-leaders

[10] Ai Agent Classification - interface-eu.org - https://www.interface-eu.org/publications/ai-agent-classification

[11] Ai Cybersecurity Trends - sentinelone.com - https://www.sentinelone.com/cybersecurity-101/data-and-ai/ai-cybersecurity-trends/

[12] Ai Security Statistics 2026 Research Report - practical-devsecops.com - https://www.practical-devsecops.com/ai-security-statistics-2026-research-report/?srsltid=AfmBOorhKQ5OUcohDnOD93T0IKehI5Wrj7bCgcUr0NJnsXkye2XdB9Or

[13] [blog] Owisez8t7c80zpzv5ov95uc54d11kd - arionresearch.com - https://www.arionresearch.com/blog/owisez8t7c80zpzv5ov95uc54d11kd

[14] [peer-reviewed] A Review of Agentic AI in Cybersecurity: Cognitive Autonomy, Ethical Governance, and Quantum-Resilient Defense - Authors: IBRAHIM ADABARA; Bashir Olaniyi Sadiq; Aliyu Nuhu Shuaibu; Yale Ibarahim Danjuma; Maninti Venkateswarlu - Journal: F1000Research - https://pmc.ncbi.nlm.nih.gov/articles/PMC12569510/

[15] The Rise Of Ai Agents Anticipating Cybersecurity Opportuniti - rstreet.org - https://www.rstreet.org/research/the-rise-of-ai-agents-anticipating-cybersecurity-opportunities-risks-and-the-next-frontier/

[16] [blog] Why Ai In Cybersecurity Still Needs Human Oversight - dropzone.ai - https://www.dropzone.ai/blog/why-ai-in-cybersecurity-still-needs-human-oversight