Shumaker

DECEMBER 12, 2025 | PUBLICATION

"When Artificial Intelligence Becomes the Hacker: Legal Risks and Compliance Strategies for Autonomous Cyber Threats"

INDUSTRY SECTOR

Technology

SERVICE LINE

Technology, Data Privacy, Cybersecurity & AI

RELATED PROFESSIONALS

Lloyd J. Wilson

MEDIA CONTACT

Wendy M. Byrne
wbyrne@shumaker.com

In mid‑September 2025, Anthropic’s Threat Intelligence team reported detecting and disrupting what it assesses as the first large-scale cyber-espionage campaign orchestrated primarily by an Artificial Intelligence (AI) system. The operation was assessed with high confidence to be a Chinese state-sponsored group that leveraged Anthropic’s *Claude Code* tool to conduct autonomous intrusion activity to target approximately 30 global organizations, and succeeding in a small number of cases.[1] This is a significant shift as it raises concerns that cyberattacks can be amplified in both velocity and scale as AI systems replace human attackers. Proactive governance, contractual safeguards, and framework aligned controls are essential when the hacker is an algorithm.

I. Turning Point: AI as a primary attacker

Anthropic reports AI executed 80 percent to -90 percent of the tactical work (reconnaissance, exploit development, lateral movement, credential harvesting, data parsing, and documentation), with humans intervening at a few decision points for strategic authorization gates.[2] What is unique is that this allows for operational autonomy at scale for hackers. The threat actors were successful in bypassing guardrails by allegedly role-playing as employees of legitimate cybersecurity firms to convince *Claude* that its actions were part of authorized penetration testing.[3] Once the guardrails were bypassed, they used a custom orchestration framework built around the Model Context Protocol (MCP), an open standard that enables AI models to interact with externals tools and systems. MCP acted as a bridge between the AI and various technical utilities such as network scanners, exploit frameworks, and browser automation tools.[4] By leveraging MCP, the attackers could decompose complex multi-stage attacks into smaller, routine technical tasks that appeared to *Claude* to be benign when viewed individually. *Claude* then executed the tasks autonomously at rates that would have been physically impossible for humans to perform and chained each of the tasks together into full attack sequences without revealing the malicious context.[5] This orchestration allowed the AI to function as an autonomous penetration-testing engine, coordinating multiple sub-agents and tools to progress through reconnaissance, exploitation, and data exfiltration phases with minimal oversight.

For legal teams, this development reframes questions of attribution, causation, duty of care, contractual allocation of risk, and disclosure through a lens where the "actor" is an algorithm.

II. Attribution and Liability: the algorithm as the "actor"

a. Attribution and Causation. Traditional cybercrime frameworks assume human intent; here, agentic AI executed most attack steps, with humans approving stages. In the event of an attack, lawyers should expect disputes over whether the misuse was foreseeable and whether controls sufficed to prevent AI weaponization (access controls, kill-switches, human-in-the-loop, etc.).

b. Plaintiffs and regulators will scrutinize whether companies using agentic tools implemented risk-proportionate controls (adversarial testing, tool privilege boundaries, approval steps for sensitive actions, etc.). Anthropic's report itself shows how prompt-like orchestration can steer agent behavior; combined with public advisories on prompt injection (a leading cause of AI incidents), this strengthens arguments that misuse was foreseeable.

c. Product Liability. Vendors may face design-defect or failure-to-warn claims if tools capable of autonomous actions are deployed without adequate guardrails or warnings about jailbreak, prompt injection, and agent misalignment. Plaintiffs or regulators could cite patterns identified by security researchers and industry reports showing an increase in real-world AI security failures (e.g., prompt injection, jailbreak exploits, etc.) to argue that these risks were well recognized and foreseeable.

d. Contractual Exposure. Where security representations/warranties or AI-specific addenda exist, misuse (even by an agent) can trigger breach or indemnity claims, especially if vendors did not disclose known limitations or maintain update processes consistent with recognized governance and security standards (e.g., NIST AI Risk Management Framework (RMF), ISO/IEC 42001, etc.).

III. Regulatory Frameworks: harmonizing a patchwork

a. The European Artificial Intelligence Act (AI Act). The EU AI Act is being implemented in phases and imposes risk-based obligations for high-risk AI, post-market monitoring and incident reporting, as well as obligations for general purpose models.[6] Breaches involving misuse of agentic AI can implement provider/deployer duties as a result of the AI Act.

b. S. Enforcement Posture. The FTC is actively pursuing "Operation AI Comply" to police unfair/deceptive practices that harm consumers, signaling there's no AI exemption from existing law.[7]

c. Global Governance Standards. Jurisdictions across the globe are creating their own AI regulatory policies. The National Institute of Standards and Technology (NIST) has provided a voluntary AI Risk Management Framework to help companies incorporate trustworthy considerations into their AI products.[8]

IV. Criminal and Civil Enforcement

a. Cybercrime Investigations. Even if the automated AI agent executed the attack steps of the cyberattack, companies can face inquiry into whether their accounts or tools were misused or whether their controls were inadequate. Anthropic's case demonstrates the potential for agentic misuse at a massive scale.

b. Privacy and Breach Laws. European regulations, specifically the General Data Protection Regulation (GDPR) sets strict rules for organizations to handle personal data of EU residents when AI models process or store personal data in training sets, model memory, or embeddings. If an AI-driven incident results in a personal data breach, GDPR requires controllers to notify the supervisory authority within 72 hours of awareness (unless it is unlikely to endanger individuals' rights) and to inform affected individuals without undue delay when the breach poses a high risk.

c. Cross-border and Disclosure. Factual scenarios related to jurisdiction and attribution may become legally complex when agentic operations originate abroad or traverse clouds/vendors.

V. Contractual and Insurance Considerations

a. Organizations should revisit master service agreements (MSAs) and statements of work (SOWs) to incorporate AI-focused security and governance provisions. Contracts should include clear representations and warranties confirming that vendors conduct robust adversarial testing and maintain documented processes for safe model updates and rollback.

b. Agreements should mandate administrative safeguards, such as kill-switch capabilities and human approval for privileged actions, to prevent uncontrolled autonomy. Audit rights and incident cooperation clauses should align with emerging regulatory expectations, such as the EU AI Act's documentation and post-market monitoring duties.

c. Companies should review their cyber insurance policies for exclusions and sub-limits for AI-driven incidents. Some underwriters may request evidence of controls aligned to ISO/IEC 42001 or NIST AI RMF before covering agentic workflows.

VI. Conclusion

a. Anthropic's case marks a turning point where agentic AI can compress attack timelines and scale campaigns while reducing human involvement, which reshapes the duty-of-care expectations and disclosure risk when "the hacker" is an algorithm. Legal exposure may span negligence, product liability, contractual breach, privacy obligations, and security disclosures. The path forward is AI governance, framework-aligned controls, and human oversight.

If you would like more information on legal exposure and compliance strategies related to autonomous AI-driven cyber threats, please contact Lloyd Wilson.

Whether you are reassessing governance programs, updating contractual safeguards, or implementing framework-aligned controls to mitigate AI misuse, Shumaker's Technology, Data Privacy, Cybersecurity & AI

Service Line provides forward-looking, practical guidance to help organizations stay secure and compliant as the threat landscape evolves.

[1] Anthropic, *Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign* (Nov. 17, 2025), https://www.anthropic.com/news/disrupting-AI-espionage.

[2] *See Id*.

[3] *See Id.*

[4] *See Id.*

[5] *See Id.*

[6] *Implementation Timeline*, **Artificial Intelligence Act** (last visited Dec. 10, 2025), https://artificialintelligenceact.eu/implementation-timeline/.

[7] Fed. Trade Comm'n, *FTC Announces Crackdown on Deceptive AI Claims and Schemes* (Sept. 25, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes

[8] Nat'l Inst. of Standards & Tech., *AI Risk Management Framework* (last visited Dec. 10, 2025), https://www.nist.gov/itl/ai-risk-management-framework.