Agentic AI and Cybersecurity: Policy Implications from the Anthropic Hacking Incident

For this assignment I chose a recent story about suspected Chinese state-backed hackers using Anthropic's Claude Code "agentic" AI system to automate spying on roughly thirty organizations worldwide (Sabin, 2025). According to reporting summarized in the OECD AI Incidents Database, the attackers posed as a legitimate company, broke their objectives into small subtasks to bypass safety checks, and then let the AI help them scan networks, write exploit code, plant backdoors, and exfiltrate data with very little human effort. Several of the attacks succeeded, which means an AI coding assistant was effectively turned into an operational cyber weapon. This event sits at the intersection of information technology, security, and policy because it shows how general purpose AI tools can shift the scale and speed of cyberattacks long before regulators fully understand their impact.

In my view, the core issue is not simply "China" or "hacking," but the way autonomous or semi-autonomous AI agents can amplify both skilled and unskilled attackers. Traditional cybersecurity threats already stretch the capacity of governments and companies. When an AI system can generate exploits, adapt to defenses, and coordinate tasks at machine speed, the balance between defenders and attackers tilts even further toward those who are willing to ignore norms. The Axios coverage notes that Claude was used to chain together reconnaissance, exploitation, and data theft in a way that would normally require a team of humans (Sabin, 2025). That raises difficult questions: What

responsibilities do AI providers have to monitor and limit high-risk use, and how do we design policies that address both beneficial and harmful applications of the same tool? The legislative tracing exercise helped me connect this incident to current policy efforts. On Congress.gov I reviewed H.R. 5079, the Widespread Information Management and Cybersecurity Framework Act of 2025. This bill would reauthorize and update parts of the Cybersecurity Act of 2015, including definitions that explicitly reference advanced technologies such as artificial intelligence and machine learning, and it directs federal cybersecurity centers to improve information sharing, threat analysis, and coordination across critical infrastructure sectors (H.R. 5079, 2025). Although the bill does not name Claude or agentic AI, the language about emerging technologies and cyber risk clearly applies. It offers a framework for better sharing indicators of compromise and best practices when AI powered attacks occur, but it does not yet spell out clear obligations for AI vendors whose systems are misused.

Other policy voices are starting to fill that gap. The ACM Technology Policy Council's recent brief on "Systemic Risks Associated with Agentic AI" argues that AI agents that can perceive, reason, and act toward goals with limited human supervision pose new kinds of systemic risk that are not fully covered by existing regulations. The authors recommend continuous monitoring of deployed systems, requirements for robust red-teaming, and mechanisms for independent oversight of high-risk agentic deployments (ACM Technology Policy Council, 2024). Combined with H.R. 5079, this suggests a direction where policy could evolve: cybersecurity laws that recognize AI agents as both

defensive and offensive tools, paired with professional standards for how developers design, log, and audit agent behavior.

The stakeholders in this event are wide ranging. First are the victims, the thirty or so organizations whose systems were probed or breached. They may include companies, nonprofits, or government agencies whose data and operations were put at risk. Second are AI providers like Anthropic, as well as cloud platforms that host agentic systems. They benefit from making powerful tools broadly accessible, but they also carry reputational and potentially legal risk when those tools are abused. Third are governments and regulators, who must decide how to update cybersecurity, export control, and privacy laws to account for AI enabled attacks. Finally, there is the broader public, whose personal data and critical services, such as healthcare or transportation, depend on resilient infrastructure that can withstand these new kinds of threats.
From my perspective, the policy implications revolve around accountability and transparency. H.R. 5079 focuses on improving federal coordination and support for critical infrastructure, which is important, but the Anthropic case shows that we also need clearer expectations for private AI companies. For example, AI incident reporting could be integrated into existing cybersecurity disclosure rules, so that when a vendor discovers that its model has been used in a major attack, it must share technical details with agencies like CISA. Standards bodies and professional associations such as ACM can help define what “reasonable” safeguards look like: rate-limiting high risk actions, building in strong identity checks for enterprise agents, and maintaining detailed logs

that enable forensic analysis when misuse occurs (ACM Technology Policy Council, 2024).

For information professionals, this incident is a reminder that policy literacy is becoming part of the job. Whether working in cybersecurity, data governance, or software engineering, professionals will increasingly need to understand how tools like Claude fit into legal frameworks such as the Cybersecurity Act and how to document their systems in ways that support investigations and audits. There is also an ethical dimension. As Pasek (2015) argues, information policy shapes the entire information cycle, from creation and storage to access and use. When AI agents participate in that cycle, we have to think carefully about who can trigger them, what data they can reach, and how their outputs are monitored.

For society more broadly, the Anthropic incident is a warning and also an opportunity. It shows that AI can significantly lower the barrier to launching sophisticated cyberattacks, which raises the stakes for national security and personal privacy. At the same time, it gives policymakers a concrete example to work from instead of debating purely hypothetical risks. If we treat this as an early case study, we can build laws and norms that encourage responsible AI development while discouraging reckless deployment. That means pairing legislative updates like H.R. 5079 with strong professional standards, better public reporting of AI related incidents, and educational efforts that prepare future technologists to think about security and ethics from the start.

References

ACM Technology Policy Council. (2024). *Systemic risks associated with agentic AI: A policy brief.* Association for Computing Machinery. https://www.acm.org/binaries/content/assets/public-policy/europe-tpc/systemic_risks_agentic_ai_policy-brief_final.pdf

Anthropic AI hacking incident source:

Sabin, S. (2025, November 13). Chinese hackers used Anthropic's AI agent to automate spying. *Axios.* https://www.axios.com/2025/11/13/anthropic-china-claude-code-cyberattack

Federal legislation source (legislative tracing):

H.R. 5079, 119th Cong. (2025). *Widespread Information Management and Cybersecurity Framework Act of 2025.* https://www.congress.gov/

Information policy theory source:

Pasek, J. (2015). Defining information policy: Relating issues to the information cycle. *New Review of Academic Librarianship, 21*(2), 286–303. https://usf-flvc.primo.exlibrisgroup.com/discovery/fulldisplay?docid=cdi_crossref_primary_10_1080_13614533_2015_1009126&context=PC&vid=01FALSC_USF:USF

General ACM policy page (used for context in essay):

Association for Computing Machinery. (n.d.). *Public policy.* https://www.acm.org/public-policy

TechNews source (for locating the event):

Association for Computing Machinery. (2025, November 14). *ACM TechNews.* https://technews.acm.org/