When AI becomes the attacker

1 MARCH 2026
Julian ter Hedde



Artificial Intelligence is rapidly transforming the character of cyber conflict. In November 2025, Anthropic revealed that its AI coding assistant, Claude Code, had been manipulated by a Chinese state-sponsored hacking group and was used in attempted cyber infiltrations of thirty global targets (Anthropic, 2025). Just months later, in early 2026, a cybercriminal successfully exploited Claude again to assist in stealing 150 GB worth of sensitive data from the Mexican government (Knobloch, 2026). Although cyber espionage and state-sponsored hacking have been part of international politics for quite some time, the arrival of autonomous AI systems marks a new era. Unlike traditional cyber tools, AI systems can independently analyse

vulnerabilities, generate attack strategies and execute operations with high speed and on a large scale (Brundage et al., 2018).

In cyber conflict, autonomy refers to the ability of AI systems to perform operational tasks without continuous human supervision (NIST, 2023). Autonomous cyber systems can analyse networks, identify weak points, generate exploit code and adapt their own behaviour dynamically (Brundage et al., 2018). Humans define objectives, but the system autonomously selects and executes operational steps to achieve them. The human role changes from a direct executor to just overseeing the AI (Scharre, 2018). Although this is more efficient, it reduces human control over tactical decisions, as the AI adapts dynamically, which makes predicting its behaviour harder.

This development raises the central question of this article: "What are the implications of autonomous AI-driven cyberattacks for organizational security, and are current cyberdefence tools able to adequately counter these threats?" This question is particularly relevant to the European Union, which relies heavily on foreign digital infrastructure and AI technologies. As cyber capabilities become increasingly important to geopolitical competition, technological dependence creates structural vulnerabilities (Farell & Newman, 2019).

This article argues that autonomous AI systems deeply alter cyber conflict by increasing the speed, scale, and autonomy of cyber attacks. On top of that, it reinforces geopolitical asymmetries between technological powers and dependent actors such as the European Union. To mitigate the risks and increase cyber resilience, policymakers must push for AI-enabled defence strategies.

What made the Claude incident particularly interesting is the fact that the system itself was never breached in the traditional sense (Anthropic, 2025). No firewall was broken, and no internal infrastructure was compromised. Instead, the attackers exploited the system through legitimate access, strategically manipulating its capabilities to serve their operational goals. In this manner, the Chinese state-sponsored group was able to convince Claude to analyse target infrastructures, identify vulnerabilities and generate exploit code that was able to adapt to their specific systems. The AI was able to do this in a span of time that would normally require substantial human expertise, allowing attackers to make their methods more efficient and on a larger scale. In early 2026, investigators reported that a hacker manipulated Claude again to assist in a prolonged cyberattack on Mexican government networks, exfiltrating roughly 150 GB of sensitive data (Knobloch, 2026).

This incident illustrates a broader structural shift in cyber conflict: it fundamentally alters the escalation dynamics by compressing decision-making timelines and increasing the risk of unintended escalation (Horowitz et al., 2018). Traditionally, cyber attacks required human operators to find vulnerabilities, develop exploits or execute attacks. AI systems can now execute these tasks autonomously by analysing network structures, gener-

ating code, and adapting that code swiftly to defensive measures (Brundage et al., 2018). These capabilities increase the speed, scale and effectiveness of cyberattacks immensely, and drastically lower the barriers to conducting complicated cyber operations (Horowitz, 2018). As a matter of fact, AI agents can scan thousands of systems simultaneously, identifying weaknesses across large networks (Trajkovska, Del Becaro, & Mijalkov, 2024). Unlike human hackers, they can operate without fatigue or resource limitations, continuously. This allows cyberattacks to be conducted faster and at a greater scale than ever before. This drastically reduces the time available for defenders to respond or interpret an attack's intent, which could result in defensive systems launching automatic countermeasures before policymakers can assess whether an incident was espionage, sabotage, or a malfunction. Machine-driven interactions can intensify conflict without deliberate human authorization (Scharre, 2018). Autonomous defensive AI may interpret normal network activity as hostile reconnaissance and respond aggressively, triggering retaliatory cycles. Cyber operations often occur in the grey zone (between the thresholds of peace and full-scale war), and these automated responses can escalate political tensions unintentionally (Lindsay, 2015). The speed and autonomy of AI make states lose effective control over the pace and consequences of cyber confrontations, increasing instability. This brings us to a second challenge in the grey zone: attribution. AI-driven cyberattacks significantly complicate attribution, making it increasingly difficult to identify perpetrators (Rid & Buchanan, 2015). Attribution in cyber conflict was already difficult because of anonymization tools, proxy actors (for example, by means of VPNs) and global infrastructure. Autonomous AI systems can amplify this problem by dynamically modifying their code, generating unique attack patterns and operating across distributed systems, which makes linking attacks to actors based on technical indicators very difficult (Brundage et al., 2018). When attribution is uncertain, deterrence weakens because targets cannot confidently identify the perpetrators (Libicki, 2009). Incorrect attribution may result in escalation with the wrong actor (Lindsay, 2015). Besides that, autonomous systems provide attackers with plausible deniability because the accused actor can blame unintended system behaviour to cause the attack, which erases accountability and makes international law in the field of cyber conflict outdated (Schmitt, 2017).

This process has major geopolitical implications. Cyber capabilities are increasingly central to strategic competition between major powers, particularly the EU, the US, and China (Hobbs, 2020). These states are in a global technological arms race, viewing AI as a strategic capability that enhances both offensive and defensive cyber power. They use cyber tools to conduct espionage, disruption and project influence without conventional military confrontation. AI enhances these capabilities significantly, reinforcing geopolitical asymmetries. Because this is part of the grey zone and relies on technological obfuscation, attribution becomes politically and

technologically contested. At the moment, the US and China dominate global AI development, cloud infrastructure, and digital platforms. Those advantages provide strategic leverage, which allows states to project cyber power globally. But the competition also creates destabilising incentives, as rapid development reduces safety testing and increases the risk of unintended vulnerabilities (Brundage et al., 2018). Competitive pressures also discourage transparency and cooperation, because states want to preserve their advantages (Cummings, 2017). Often, offensive capabilities are developing faster than defensive countermeasures, which creates vulnerabilities that adversaries may exploit (Libicki, 2009). As a result, the AI arms race increases cyber insecurity, even as states attempt to strengthen their defensive capabilities.

Furthermore, because autonomous AI also lowers the barriers for complicated hacks, less capable states are now also able to conduct operations that were previously limited to the technologically advanced states (Brundage et al., 2018). Thus, the arrival of autonomous AI increases systemic instability by expanding the number of actors that are capable of launching cyberattacks. Technological leaders still retain structural advantages due to superior computational capacities, data access, and infrastructure (Farell & Newman, 2019). This creates interesting consequences: AI simultaneously democratises offensive cyber capabilities while still reinforcing existing geopolitical asymmetries.

In contrast, Europe faces dual vulnerability: exposure to AI-driven cyber threats and structural dependence on foreign digital infrastructure and AI technologies. Control over digital infrastructure enables both surveillance and coercion, which is described as "weaponized interdependence" (Farrell & Newman, 2019). States that control key technical components can exploit these for strategic advantage. Existing cyber defence systems are currently insufficient against autonomous attacks, and Europe's reliance on foreign AI providers limits its strategic autonomy (Trajkovska, Del Becaro, & Mijalkov, 2024).

The EU has already recognized cybersecurity and digital sovereignty as priorities. The EU Cybersecurity Strategy addresses the need to strengthen resilience, improve cyber defence capabilities, and reduce technological dependence (European Commission, 2024). Besides that, the AI Act seeks to regulate AI systems and reduce risks that are associated with their deployment (European Commission, 2024). However, these regulations are mostly regulatory and defensive in nature; they aim to reduce the risks within the EU market, but do not address the geopolitical dimension of AI-driven cyberconflict, which creates a gap between sovereignty and technological capacity. While the AI Act governs how AI systems are used and developed, it does not provide Europe with technological autonomy or offensive and defensive capabilities comparable to China or the US (Farrell & Newman, 2019). As a result, the EU remains dependent on foreign AI providers, which limits the ability to control critical technologies and

respond independently to AI-driven cyber threats. In this geopolitically competitive reality, the EU's regulation alone is not enough. Without comparable technological capabilities to the major tech powers, Europe risks becoming a rulemaker without technological power, which leaves it vulnerable in an AI-driven cyber conflict environment.

Organisations must adapt their cybersecurity strategies to address autonomous AI threats. Traditional security approaches, which rely on human analysis and reactive defence, are insufficient against autonomous attackers. Companies and organisations have to instead adopt AI-enabled defensive systems capable of detecting and responding to threats in real time (Trajkovska, Del Becaro, & Mijalkov, 2024). Zero-trust architectures are an effective approach. These systems continuously verify access and limit hackers' ability to move within networks (Nist, 2023). AI-enabled threat detections can also identify anomalies and respond swiftly. As autonomy reshapes cyber conflict, control over artificial intelligence itself is emerging as a fundamental determinant of power, security, and sovereignty in the digital age.

References

Anthropic. (2025, August). Threat intelligence report: August 2025.

Anthropic. (2025, November). Disrupting the first reported AI-orchestrated cyber espionage campaign: Full report.

Cummings, M. L. (2017). Artificial intelligence and the future of warfare. *Chatham House.*

European Commission. (2020). EU cybersecurity strategy for the digital decade.

European Commission. (2024). Artificial Intelligence Act.

European Union Agency for Cybersecurity (ENISA). (2023). ENISA threat landscape 2023. *Publications Office of the European Union.*

Farrell, H., & Newman, A. L. (2019). Weaponized interdependence: How global economic networks shape state coercion. *International Security, 44(1), 42–79.*

Hobbs, C.,(2020). Europe's digital sovereignty: From rulemaker to superpower in the age of US–China rivalry. *European Council on Foreign Relations.*

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, 1(3), 36–57.

Knobloch, A. (2026). Claude: AI chatbot used for cyberattack on Mexican government. *Heise.*

Libicki, M. C. (2009). Cyberdeterrence and cyberwar. *RAND Corporation.*

Lindsay, J. R. (2015). Tipping the scales: The attribution problem and deterrence. *Journal of Cybersecurity,* 1(1), 53–67.

National Institute of Standards and Technology (NIST). (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). *U.S. Department of Commerce.*

Rid, T., & Buchanan, B. (2015). Attributing cyber attacks. *Journal of Strategic Studies,* 38(1-2), 4-37.

Scharre, P. (2018). Army of none: Autonomous weapons and the future of war. *W. W. Norton.*

Schmitt, M. N. (2017). Tallinn manual 2.0 on the international law applicable to cyber operations. *Cambridge University Press.*

ter Hedde, J. (2025). Clouded independence: Europe's struggle for digital sovereignty. *EPIS Thinktank for Foreign & Security Policy.*

Trajkovska, E., Del Becaro, T., & Mijalkov, B. (2024). Prevention of cybercrime in the age of artificial intelligence (AI) within the European Union. In *Social changes in the global world* (11th International Scientific Conference, Faculty of Law, Goce Delcev University).

EPIS