CARNEGIE
EUROPE

JULY 2026

When AI Agents Attack: Autonomous Cyber Operations and Europe's Governance Gap

Raluca Csernatoni
Patryk Pawlak

When AI Agents Attack: Autonomous Cyber Operations and Europe's Governance Gap

Raluca Csernatoni
Patryk Pawlak

July 2026

About the Carnegie Endowment for International Peace

In a complex, changing, and increasingly contested world, the Carnegie Endowment generates strategic ideas, supports diplomacy, and trains the next generation of international scholar-practitioners to help countries and institutions take on the most difficult global problems and advance peace. With a global network of more than 200 scholars across twenty countries, Carnegie is renowned for its independent analysis of major global problems and understanding of regional contexts.

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

Carnegie Europe
Rue du Congrès, 15
1000 Brussels, Belgium
P: +32 2 735 56 50
CarnegieEurope.eu

This publication can be downloaded at no cost at CarnegieEurope.eu.

TABLE OF CONTENTS

About the Authors

Raluca Csernatoni is a fellow at Carnegie Europe, where she works on European security and defense, with a focus on emerging and disruptive technologies like AI.

Patryk Pawlak is a visiting scholar at Carnegie Europe, where his fields of expertise include global governance of cyberspace and the impact of technology on foreign and security policy.

Acknowledgments

This paper was produced in the context of the EU Cyber Direct—EU Cyber Diplomacy Initiative project with the financial assistance of the EU. The contents of the paper are the sole responsibility of the authors and can under no circumstances be regarded as reflecting the positions of the EU or of any other institution.

The authors acknowledge the use of LLM tools for minor clarificatory editing in earlier versions of the draft. The argument, analysis, structure, and substantive content are entirely the authors' own.

Carnegie Europe

Carnegie Europe delivers interdisciplinary expertise and independent insights that bring together national, regional, and global perspectives and help European policymakers grasp and respond to global challenges. From Brussels, we focus on three themes: the European Union's relations with its partners and competitors; the risks of democratic backsliding in Europe and around the world; and Europe's efforts to meet today's most pressing global challenges, from climate change to the new frontiers of cyber diplomacy.

Introduction

In January 2026, a new online platform called Moltbook quietly appeared on the internet. Advertised as "the front page of the agent internet," the platform resembled a familiar social network. Users formed communities, debated ideas, and organized collaborative projects.[1] But Moltbook had one crucial difference: None of its users were human. The platform was designed exclusively for autonomous artificial intelligence (AI) agents—software systems capable of reading, reasoning, posting, and interacting with one another without direct human instruction. Humans were "welcome to observe." Within days of Moltbook's launch, more than 1.5 million agent accounts had been created.

The Moltbook experiment mattered because it showed what happens when autonomous agents become a population rather than isolated tools. It revealed both a technical failure and a governance problem.[2] On the first aspect, a misconfigured database on the Moltbook platform resulted in a data leak, which exposed 1.5 million application programming interface (API) authentication tokens, tens of thousands of email addresses, and private communications between agents.[3] Yet, the more important lesson was structural: Behind the autonomous systems stood just 17,000 human operators controlling an average of nearly ninety agents each. In such an environment, even a familiar security lapse has unfamiliar consequences, because identity, authorship, and accountability become much harder to trace when machines act at scale on behalf of a small number of humans.

The broader shift from isolated AI tools to agentic cyber capabilities has been accelerating in recent months. In 2025, China-linked groups used AI company Anthropic's large language model (LLM), Claude, against foreign governments and critical infrastructure in the first publicly reported large-scale AI-assisted operation.[4] Only a few months later, in April 2026, the same company announced that it would not make Claude Mythos Preview, its latest general-purpose frontier model, generally available, citing its ability to autonomously identify and exploit zero-day vulnerabilities across every major operating system and web browser.[5] Anthropic initially restricted access to the model via Project Glasswing, a consortium of firms and organizations that maintain critical software, including Amazon Web Services, Apple, Google, JPMorganChase, Microsoft, and Nvidia; in June 2026, access was expanded to an additional 150 organizations in fifteen countries.[6]

The initiative reflects a defenders-first rationale of granting privileged early access so that vulnerabilities can be found and fixed before a capability spreads more widely. Independent testing by the United Kingdom (UK) AI Security Institute found that Mythos Preview solved expert-level capture-the-flag challenges 73 percent of the time and was the first model to complete a thirty-two-step simulated network intrusion from start to finish, while it

struggled to penetrate a hardened, actively defended environment.[7] That a frontier developer now judges one of its general-purpose models too dangerous for public release is a worrying sign that the boundary between AI-assisted cyber operations and autonomous cyber capability is blurring.

This boundary has become a matter not only of technical restraint but also of national security control. As of this writing, a U.S. government export-control directive requires Anthropic to suspend access to its AI models Fable 5 and Mythos 5 for all foreign nationals, including the company's own employees, effectively disabling the models for all customers.[8] The move shows that access to frontier cyber-capable models can be abruptly reorganized by unilateral state authority, not just by developer safety policies or voluntary release decisions.

These episodes are stark, public demonstrations of the security risks that arise when autonomous AI agents are deployed at scale without proper governance, security measures, or even basic technical hygiene. They offer an early glimpse into a rapidly emerging reality: a digital ecosystem in which autonomous AI systems interact, cooperate, and potentially compete at machine speed, often with limited human oversight. In such environments, cybersecurity risks no longer stem exclusively from malicious humans using increasingly powerful tools. Rather, these risks arise from autonomous software systems capable of initiating actions, chaining decisions, and interacting across digital infrastructures on their own. Instead of assisting human attackers, these systems can increasingly function as actors themselves, able to conduct reconnaissance, identify vulnerabilities, write exploits, and execute attacks in iterative cycles.

This evolution does not simply accelerate existing threats; it changes the structure of the threat environment. Cybersecurity frameworks built around human adversaries, identifiable operators, and sequential attack chains are already stretched and will struggle even more to account for autonomous systems that operate continuously and at scale.

For the European Union (EU), these developments raise urgent governance issues that require a more agile, faster policy response than existing approaches. While the most advanced systems still struggle with long-term planning, state management, and error recovery, the strategic challenge for European policymakers is to adapt rapidly enough: not within decades but within years.[9]

Three trends are worth highlighting. First, so-called agentic AI represents a structural shift in the cyber threat landscape, as it introduces new attack surfaces and operational dynamics that differ from previous forms of AI-enabled cyber activity. Second, existing EU cybersecurity and AI governance frameworks only partly address these risks, leaving gaps in how autonomous systems are deployed, monitored, and secured. Third, the rapid integration of advanced AI models into the U.S. cyber strategy, combined with the EU's technological dependence on American AI infrastructure, complicates the bloc's ability to shape the governance of these technologies.

Taken together, these developments create a twofold challenge for Europe: rules that are not fit for purpose to deal with autonomous cyber operations at runtime; and dependence on a U.S. AI ecosystem whose strategic and normative priorities may diverge from European goals over time. If autonomous AI systems become a routine part of cyber conflict, Europe's problem is not regulatory volume but regulatory fit: The EU framework was built for human operators and static software, not for autonomous systems operating in trusted environments at machine speed.

The Agentic Turn in Cybersecurity

Agentic AI refers to software architectures that combine LLMs with planning loops, tool access, memory, and feedback mechanisms, enabling systems to pursue goals autonomously across digital environments.[10] Unlike traditional automation or AI-assisted tools that execute predefined instructions, agentic systems can observe their environment, choose their actions, evaluate outcomes, and iteratively adjust their behavior with limited human supervision. In cybersecurity contexts, this means that the system itself can orchestrate complex sequences of actions, from reconnaissance to exploitation, rather than only assisting a human operator.[11]

Cybersecurity has historically been organized around the assumption of a human adversary and the limitations around the speed and complexity of the attacks that this entails. That assumption shaped its defining features: intrusion timelines measured in hours and days, kill chains that move sequentially from reconnaissance to exfiltration, and attribution frameworks based on behavioral signatures.[12] For much of the past decade, the intersection of AI and cybersecurity has focused on a relatively limited set of issues, from machine learning–based detection of anomalies to, more recently, the use of generative models to speed up phishing, malware development, and social engineering at scale.[13] These are real and serious threats that share a common element: meaningful human involvement in the process, namely a person still set the objective and drove the operation forward.

The shift from AI-assisted tools to agentic systems challenges many traditional assumptions about cybersecurity and transforms the threat landscape in three structural ways.[14] First, when it comes to autonomy, attackers can delegate decisionmaking to software agents that can execute multistep operations independently. Second, with regard to scale, a single operator can deploy hundreds or thousands of agents that operate simultaneously across targets. Finally, in terms of cognitive attack surfaces, adversaries increasingly target the behavior of AI systems themselves—through prompt injection (disguising malicious inputs as legitimate prompts), jailbreaks (embedding instructions that allow a system to bypass its safety restrictions), or data manipulation—and do not only exploit software vulnerabilities.[15]

Technically, most agentic systems operate through a recurring decision loop. An LLM generates a plan; selects a tool, such as web browsing, code execution, or database queries; executes the action; and then evaluates the outcome before deciding on the next step. The system may also store information in its memory and retrieve external data sources. This architecture enables agents to chain together dozens of actions without human input, creating operational behavior that resembles a continuous cyber campaign, not a single scripted exploit.

What matters here is not only the automation of reconnaissance or exploitation. Highly autonomous cyber-capable systems also rely on operational capabilities: the capacity to establish infrastructure, coordinate across multiple instances, acquire compute and financial resources, evade detection and shutdown, and iteratively enhance their own performance. This broader operational stack is what transforms agentic AI from a useful tool into a potentially persistent cyber actor.[16]

Reinforcement learning—the training approach that makes agents goal-oriented—produces systems that are indifferent to the means they use to achieve their objectives.[17] Work by the Center for AI Standards and Innovation at the U.S. National Institute of Standards and Technology has shown that when agents have flexible tools, such as code execution, they can exploit loopholes and even game the surrounding rules and checks in ways that appear to be successful.[18]

As cybersecurity experts have repeatedly observed, AI agents operating under reinforcement learning are not programmed to respect the spirit of the constraints placed on them.[19] Rather, they are programmed to succeed at their task and exploit any available route to do so. This has direct and unsettling implications for the design of both enterprise security and regulatory governance. For example, an enterprise AI agent tasked with gathering threat intelligence from emails, internal documents, and the open web could encounter a malicious prompt embedded in a web page or attachment that instructs it to ignore prior restrictions, extract sensitive credentials, and send them to an external server. Because the agent is optimized to complete the task rather than understand the spirit of the security policy, it may treat that hostile instruction as a legitimate step toward its objective.

While agentic AI compresses detection timelines from months to seconds and lowers the barriers for offensive automation, it also removes the cognitive or logistical bottlenecks that constrained adversaries in the past and creates attribution problems that existing frameworks find difficult to address.[20] Although these frameworks can still function in such contexts, attribution is becoming more complex and more susceptible to machine-generated false signals or hallucinations, especially when agentic systems operate without direct human oversight.

Traditional cybersecurity infrastructure assumes that actions are attributable to human users or identifiable services. Agentic systems blur this distinction. Thousands of autonomous agents may act on behalf of a single individual or organization, often without clear external

markers that distinguish them from human users, thereby obscuring the perpetrator's identity.[21] Emerging proposals, including cryptographic agent identities and auditable action logs, suggest that future digital infrastructures may need mechanisms to verify whether an interaction originates from a human operator or an autonomous system.[22]

AI Turbocharging Cybercrime and Espionage

In November 2025, Anthropic disclosed an incident that researchers described as one of the first publicly reported cases in which an LLM was used as a central component of an automated cyber intrusion campaign.[23] Anthropic assessed with high confidence that the operation had been conducted by a Chinese state-sponsored group and described it as the first documented case of a large-scale cyber attack executed without substantial human intervention.[24] A threat actor had repurposed the company's AI assistant, Claude, to become the central component of an autonomous hacking campaign. Using prompt-based jailbreaking to frame the AI as a legitimate cybersecurity tool, the attacker directed Claude to carry out most of the intrusion process independently: reconnaissance, vulnerability detection, exploit development, initial access, and data exfiltration.[25] The AI carried out an estimated 80–90 percent of the operation without human guidance, working at machine speed and compressing what would have taken months of skilled human effort into just a few days.[26]

This was not an isolated incident. At around the same time, Google reported that threat actors were experimenting with AI-enabled malware and model misuse across the attack life cycle. In one case, the Russian government-backed cyber espionage group APT28 used PROMPTSTEAL, a data-mining tool that queried an open-source LLM via U.S. AI company Hugging Face against Ukrainian targets. Separately, Google observed state-sponsored actors, including from China, Iran, and North Korea, misusing its AI assistant Gemini for reconnaissance, phishing, command-and-control development, and data-exfiltration support.[27] OpenAI reported similar patterns of state-linked misuse across its platform.[28]

The overlap among multiple frontier model providers shows that the weaponization of commercial AI is not a flaw unique to any one company's safety system. It is increasingly a structural aspect of the current landscape, in which the same capabilities that make LLMs useful for software development, code generation, and system analysis also make them effective tools for offensive cyber operations.

Hence, the significance of the November 2025 incident goes beyond the specific breach it permitted. Several structural aspects of the campaign require ongoing policy attention to anticipate future trends. The attack did not exploit a technical flaw in Claude's code but a social-engineering method applied directly to the model. Traditionally, this method—jailbreaking—has been a skilled craft that has outstripped the defensive efforts of model providers. But with recent progress in AI models, this practice is shifting from a specialist technique to a widespread capability, whereby models can automatically generate effective multistep jailbreak prompts, reducing the skill barrier.

Cyber criminals circumvent ethical and safety restrictions by using new tools to leverage or create LLMs for malicious ends.[29] Powerful black-hat generative AI models, such as WormGPT, make it much easier to create realistic attacks at scale.[30] The Anthropic case shows that sophisticated state actors have already begun to incorporate prompt-based AI manipulation into their offensive cyber operations, demonstrating that these techniques are moving from experimentation into real-world campaigns.

The case also reveals the force-multiplying effect at the core of AI-enabled offensive operations. Although the episode does not suggest the full democratization of offensive cyber capabilities, since it involved a highly capable state-linked actor, it does show how AI can compress time, reduce labor demands, and amplify the operational reach of even a small number of human operators. States with limited cyber talent may increasingly be able to deploy sophisticated intrusion capabilities by combining a few operators with powerful AI agents. Nonstate actors, criminal enterprises, and ideologically motivated groups may also face lower barriers over time as jailbreak techniques, open models, and malicious tooling continue to diffuse beyond state use.[31]

Perhaps most significantly in terms of governance, the Anthropic case exposed the limitations of current AI safety frameworks as security measures. Safety alignment, which involves training LLMs to decline harmful requests, proved inadequate against persistent adversarial prompting and jailbreaking.[32] In this respect, defenders cannot depend on models to regulate themselves. This has serious implications for how regulators, deployers, and developers perceive the relationship between AI safety and cybersecurity, which are related but separate fields, and advances in one do not ensure security in the other.

The tension between safety commitments and operational demands reached a crisis in February 2026, when the U.S. secretary of defense demanded that Anthropic grant unrestricted military access to Claude or risk losing a $200 million government contract.[33] Anthropic refused to cross two established redlines: fully autonomous lethal targeting without human oversight, and domestic mass surveillance. Consequently, the Department of Defense designated Anthropic a supply-chain risk to national security.[34]

The significance of this case lies in the governance questions it raises. The dispute centered on future permissible uses of frontier AI, especially autonomous weapons and domestic surveillance, and on who sets the terms of downstream control once a model is embedded in

classified military environments. The case therefore highlights how quickly provider safeguards come under pressure when procurement leverage and operational dependence shape the conditions of use. It also shows that safety alignment is inseparable from procurement power, state pressure, and corporate leverage over downstream military use.

The politics of military AI are increasingly fought through legislation and doctrine as well as contract language, blacklist threats, and supply-chain designations.[35] In the case of Anthropic, the speed with which a competitor moved into that space sharpened the point. The military contract awarded to OpenAI mere hours after Anthropic was banned shows how quickly state demand and market competition can reshape the practical governance environment around frontier models.[36]

For European deployers that rely on U.S.-developed models, the central lesson is about durability and control. When the practical terms of deployment are set through U.S. procurement power and security doctrine, European governance frameworks require stronger operational safeguards of their own. The issue is not only that AI agents can be weaponized for cyber operations. It is also that developer safeguards can be bypassed, while downstream use is shaped by state power and procurement leverage rather than by provider commitments alone. The lesson is that European deployers increasingly depend on frontier models that are developed, governed, and politically pressured outside Europe. If model providers can be compelled to relax or reinterpret safety commitments under U.S. national security pressure, then Europe cannot treat provider-side alignment as a stable substitute for its own operational controls, procurement conditions, and runtime oversight.

Attack Machines and the Guardrail Illusion

The Anthropic case also highlights a wider range of guardrail failures that have become alarmingly common in the deployment of commercial AI agents. In February 2026, Microsoft confirmed that a bug in its 365 Copilot Chat product had led the AI assistant to summarize confidential emails for weeks, circumventing enterprise data loss prevention (DLP) policies and sensitivity labels explicitly set to prevent such access.[37] The vulnerability affected emails stored in users' Sent Items and Drafts folders across Microsoft's enterprise customer base. The AI had not been directly instructed to process confidential material. It processed it anyway, not out of malicious intent, but because a coding error allowed it to bypass the governance layer between the model and the data. The point is not that Copilot went rogue but that enterprise controls built for conventional software can fail when an AI assistant is granted broad semantic access across communications and documentation systems.

The Microsoft Copilot incident highlights a broader issue. DLP tools aim to prevent the deliberate or accidental sharing of sensitive information. However, AI assistants operate differently.[38] To be effective, they need to access and process large amounts of data across emails, documents, chats, and cloud storage. Once an AI system is integrated into everyday workplace systems, it can effectively see everything those systems can access.[39] Merely attaching existing DLP rules to these tools creates gaps, because the controls were not designed for AI-driven access and summarization.

A safer approach would be to design permissions, monitoring, and governance with AI in mind from the outset, rather than patch them after deployment. In that setting, the central risk is that an AI assistant acts on the basis of semantic access instead of clearly bounded user intent. A workplace agent that can read confidential emails, draft messages, summarize documents, and search internal repositories operates inside the organization's trusted environment with unusually broad contextual visibility. This creates a wider attack surface: A malicious instruction hidden in an email, an attachment, a shared document, or a web page can shape how the assistant interprets its task and what information it retrieves or transmits.

In this respect, the real governance problem is architectural. It concerns how access is scoped, how agent actions are logged, whether agent identities are distinguishable from human users, and whether authorization can be enforced dynamically as agents move across systems. In a Copilot-like environment, the danger is that the assistant's broad access to trusted workplace data turns a hidden malicious instruction into an operational command.

One of the clearest examples of this practice is prompt injection.[40] Attackers can embed hidden instructions in the data an enterprise assistant is meant to read, such as a document, an email, or a web page. Because LLMs process text as both information and potential instructions, the assistant may interpret those adversarial cues as part of the user's request. In a Copilot-like setting, that could mean a seemingly routine summarization task pulls in sensitive material from protected emails, follows a malicious instruction in a document, or includes confidential content in an outward-facing response. In practice, this collapses the traditional boundary between data and code, creating an attack surface that conventional cybersecurity controls were not designed to monitor.

The Moltbook ecosystem highlighted this structural vulnerability in a different way. At the center of the platform was OpenClaw, an AI agent framework designed to operate with extensive permissions on users' local devices. It could read and modify files, access email and messaging applications, store login credentials, and interact independently with external services. It concentrated, in a single architecture, the capabilities that make agentic systems especially difficult to secure: broad access to sensitive local data, exposure to untrusted external inputs, and the capacity to take outbound action without continuous human supervision.

This is what security researcher Simon Willison has called the "lethal trifecta" of autonomous agents: a specific, threefold risk when AI can read sensitive internal information, take in material from the open internet, and send messages or files outside an organization.[41] In such a setting, a malicious instruction embedded in a web page, shared document, or message can do more than mislead the model. It can redirect the system's behavior across trust boundaries, prompting it to retrieve local files, expose credentials, or transmit sensitive material outward under the appearance of completing legitimate tasks.

The point is not simply that the agent becomes a more efficient tool for a human attacker. The architecture itself creates the conditions for a loss of control, because the same system that interprets information also holds the permissions needed to act on it. Thus, the risk is not limited to more efficient human attackers. Autonomous cyber systems present a distinct loss-of-control issue: Misalignment, adversarial exploitation, and cascading multiagent failures could transform systems designed as tools into semi-independent threat actors that are difficult to attribute, redirect, or shut down.

Europe's Governance Gap

If model-level safety training cannot reliably prevent misuse, the regulatory logic of delegating cybersecurity obligations primarily to AI developers—the implicit assumption in the EU's 2024 AI Act and 2026 Cybersecurity Package—is insufficient on its own.[42] A developer may build a model with strong safety alignment and still have that model weaponized. This means that deployers—the organizations that embed AI systems into their operations and grant them access to sensitive data—cannot treat developer safety certification as a substitute for their own security obligations.

The AI Act partly addresses this challenge, but less directly than is often assumed. The act does not regulate agentic AI as a distinct category. In effect, it regulates specific uses and contexts by prohibiting a narrow set of unacceptable practices and imposing obligations on high-risk systems listed in the regulation. Importantly, AI systems used exclusively for military, defense, or national-security purposes fall outside the act's scope. That carve-out matters greatly here, because some of the most strategically consequential autonomous cyber capabilities may emerge precisely in the domain least covered by the union's flagship AI law. The Cybersecurity Package is another effort to strengthen the EU's cyber defenses, particularly across supply chains, crisis coordination, and critical infrastructure. These are important advances. Yet, they do not fully address the problem posed by agentic AI.

A credible European response requires more than legislative updates. It also needs practical mechanisms: shared reporting channels for anomalous agent behavior, reporting procedures in case of significant incidents with the use of AI, and common standards for classifying failures that involve autonomous systems. Such a response requires stronger coordination with national computer security incident-response teams (CSIRTs) and sectoral information-sharing and analysis centers (ISACs), led by the EU's cybersecurity agency, ENISA. Without better visibility of how these systems behave in practice, the EU will struggle to distinguish isolated malfunctions from the early signs of a broader autonomous campaign.

This points to a wider weakness in the EU's current approach. Europe still places too much weight on model developers and too little on the organizations that deploy these systems in real-world settings. Existing rules focus largely on product safety, regulatory compliance, and organizational resilience. Yet, the core governance problem with agentic systems lies in runtime behavior: what they are allowed to access, what actions they can take, how those actions are recorded, and how quickly a malicious actor can redirect these agents against the systems in which they are embedded. In other words, the main risk no longer only concerns defective products or insecure components; it also stems from trusted systems acting in untrusted ways after deployment.

This is where the Cybersecurity Package remains anchored in an earlier threat model. The package centers on familiar cyber risks: insecure components, foreign suppliers, and external network attacks. Those risks are still real and important. Agentic AI, however, introduces a different kind of vulnerability. Because these AI systems operate with permission, the danger comes less from breaking in than from steering a trusted system in the wrong direction. An agent may also retain context across multiple sessions, allowing harmful instructions or patterns of behavior to accumulate gradually over time. This makes the threat harder to detect with conventional cybersecurity tools, which are typically designed to identify discrete suspicious events instead of slow, distributed failures that unfold across many interactions.

For that reason, rules centered on perimeter defense, certification, and entity resilience leave an important gap. They offer only limited guidance on how to monitor agent behavior in real time, restrict what autonomous systems can do once deployed, and respond when their behavior shifts within authorized environments. Europe's cyber framework has become stronger, but it is still not fully designed for systems that can act with a degree of autonomy within the very organizations meant to trust them.

A parallel issue concerns the emerging AI tool chain. Agentic systems typically depend on multiple external components, including APIs, third-party plug-ins, browsing tools, and cloud-based model services. Each dependency expands the attack surface and creates new supply-chain risks.[43] A compromised plug-in, manipulated API response, or poisoned retrieval database could alter an agent's behavior without directly breaching the host system. Existing supply-chain frameworks focus primarily on hardware and software components, but the security of autonomous agents increasingly depends on the integrity of these dynamic tool ecosystems.

A related weakness concerns the enabling services on which autonomous cyber operations depend to function at scale. Malicious actors do not rely solely on the agent itself. They also need access to model APIs, cloud computing, external tools, and, often, payment and hosting services. These dependencies create practical governance choke points. For that reason, oversight cannot stop at the certification of the model or system itself. It also needs to cover who can obtain high-risk access to these services, how suspicious usage patterns are detected, and under what legal authority hostile actors can be cut off from the computing, hosting, and financial infrastructure that sustains their operations.

The EU's certification framework, similarly, is focused on information and communication technology (ICT) products and services, but it does not specify how to certify or regulate agentic AI systems, which present a fundamentally different challenge from conventional software.[44] Unlike a standard product, which can be tested, assessed, and approved because it behaves the same way every time, an agentic system's behavior varies with context, cannot be fully predicted in advance, and changes as the system acquires new tools and capabilities. A system that passes a security assessment today may behave differently, and less safely, in another environment next month.

Add to this the risk that interactions with external, untrusted content can alter how a system acts, and it becomes clear why a certification regime designed for static products might be structurally inadequate for governing these systems. Independent evaluators have reached a parallel conclusion: As the most cyber-capable models increasingly exhaust undefended test environments, assessment must shift toward hardened, actively monitored ranges—an approach difficult to square with one-off certification at the point of market entry.[45]

The Transatlantic AI-Cyber Divide

The guardrail failures and espionage campaigns described above do not occur in a geopolitical vacuum. They are accelerating just as the United States realigns its approach to offensive cyber operations and increasingly integrates autonomous AI into that strategy. The stance of the administration of U.S. President Donald Trump has been described by cyber researcher Stephanie Pell as a "gloves-off" approach to foreign rivals that target U.S. networks.[46] This approach has been driven largely by disclosures of Chinese intrusions into U.S. telecommunications infrastructure and other critical systems.[47]

The transatlantic problem is not only that the United States is moving faster than Europe. It is also that Washington and Brussels are moving in different directions. The U.S. approach increasingly treats frontier AI as an operational asset to be integrated into cyber and military power, even at the cost of weakening provider-imposed safeguards. The European approach, by contrast, remains centered on regulation, cyber resilience, risk management, and legal

restraint. That creates a structural tension. Europe heavily depends on technologies produced in a system whose strategic logic it does not share and cannot control. In short, the core issue is strategic divergence.[48]

This divergence will manifest itself in at least three ways. First, European public- and private-sector actors may depend on U.S. models whose safety settings, access conditions, or downstream security and defense uses are shaped by U.S. procurement and national-security imperatives. Second, European legal obligations on data protection, due diligence, and infrastructure security may sit uneasily alongside a more permissive U.S. posture toward offensive cyber operations and contractor involvement. Third, the faster the United States embeds agentic AI into its cyber doctrine, the greater the risk that Europe becomes strategically dependent without being normatively aligned.

U.S. Offense

U.S. intelligence agencies have been conducting offensive operations with greater transparency, as U.S. officials and strategic documents have become more explicit in publicly acknowledging, signaling, and justifying offensive cyber activity. The increasing involvement of private-sector actors in such government-supported activity, however, marks a significant departure from previous doctrine.[49] This shift raises important questions about the legal frameworks that regulate state responsibility, the line between state-sponsored and private cyber operations, and the risk of escalation in an environment in which attribution is highly contested. The governance challenge is further intensified by the possibility that contractors, proxies, and commercial security firms may become early adopters of offensive autonomous agents, obscuring the distinction between state responsibility, commercial experimentation, and deniable cyber actions.

Trump's Cyber Strategy for America, presented in March 2026, marks a more aggressive posture toward using AI-enabled cyber tools to detect, divert, and deceive threat actors as well as promoting agentic AI to securely scale network defense and disruption.[50] The strategy also signals a broader willingness to treat offensive cyber capabilities as a more routine instrument of statecraft, including through AI-enabled operations designed to preempt, disrupt, and impose costs on adversaries. In this respect, cyber diplomacy will need to ensure that AI systems, particularly generative AI and agentic AI, advance not only innovation but also global stability. The United States is moving toward faster operational adoption of AI-enabled cyber capabilities, including offensive uses, while Europe's approach remains more closely tied to restraint.

Consequently, the challenge for Europe is one of interlocking technological and political dependence on an ecosystem shaped by U.S. doctrine, procurement choices, and operational priorities. For instance, cloud providers, telecommunications firms, and cybersecurity vendors with transatlantic footprints operate in a legal environment that is becoming increasingly ambiguous.[51] European data-protection and cybersecurity regulations impose

obligations that could conflict with participation in U.S.-directed offensive operations, and the extraterritorial dimensions of this tension have not been adequately addressed by either side. The involvement of private-sector actors in offensive cyber operations poses direct risks for European companies, whose infrastructure may be used, consciously or unconsciously, as conduits for such activities.

A similar debate over the U.S. "defend forward" doctrine might offer some lessons in this regard. This stance shows the value of acting early, persistently, and close to the source of cyber threats instead of waiting to respond to damage that has already been done.[52] Yet, this only carries Europe so far. The new U.S. cyber strategy sits within a more unilateral "America first" framework, which makes it a weaker model for allies seeking common rules, shared governance, and coordinated burden sharing. At the same time, the EU cyber posture has structural features that become liabilities in an AI-accelerated threat environment.

European Defense

Europe's cyber model still leans heavily toward regulation, resilience, and post-incident response. Those are important strengths, but AI-enabled cyber operations move faster, adapt more quickly, and can unfold inside trusted systems before conventional warning and response mechanisms are activated. A posture built mainly on absorbing shocks and responding after the fact becomes harder to sustain when attacks are scaled at machine speed. Europe must make a more serious effort to invest in both defensive and strategic cyber capabilities that incorporate AI systems.[53] While the EU has increased its overall defense spending, it has not yet paid similar attention to AI-enabled cybersecurity and defense. The draft Cloud and AI Development Act presented by the European Commission in June 2026 mentions cybersecurity as one of the key sectors for pioneering projects in frontier AI, but the whole process from the act's proposal to its adoption and implementation will take years.

In an environment in which adversaries already use AI to accelerate cyber operations, Europe cannot rely exclusively on traditional defense methods. The challenges are speed, scale, and timing. AI-enabled cyber operations can compress the interval between reconnaissance, exploitation, and disruption, leaving less time for human-led detection and response. The solution involves developing advanced detection, response, and resilience capabilities supported by AI, so that defenders can identify anomalous behavior earlier, contain intrusions faster, and recover more effectively when attacks unfold at machine speed. The strategic argument is straightforward: If attackers are using AI to move faster, defenders also need AI-enabled capabilities to preserve strategic stability and protect critical infrastructure. Dependence on U.S. intelligence sharing creates a single point of strategic vulnerability, both technically and politically, given the volatility of transatlantic relations.

If Europe depends too heavily on U.S. intelligence, cloud services, and frontier models to understand and respond to AI-enabled threats, its own room for autonomous action narrows in moments of crisis. In this regard, the EU could lean into asymmetric resilience—meaning

a strategy that does not try to replicate the full offensive cyber posture of the United States or China. That would require Europe to accept that it will not match U.S. or Chinese offensive cyber capacities and to invest heavily in AI-driven defensive systems, redundancy, and recovery capabilities. It would also mean prioritizing the EU's ability to absorb attacks, restore essential functions quickly, and reduce the strategic payoff of disruption. This is more consistent with the European strategic culture and the EU's institutional strengths in regulation and standardization.

Most immediately, the shift in U.S. doctrine leaves European allies facing a widening capability gap with no clear path to burden sharing. The EU and the North Atlantic Treaty Organization (NATO) have deepened their cyber cooperation, but the integration of AI-enabled offensive capabilities into allied doctrine remains in its early stages.[54] If autonomous AI systems become the main means of offensive cyber operations in the coming years, the EU's current posture—primarily defensive, institutionally fragmented, and reliant on U.S. intelligence sharing and technological capabilities—may be structurally inadequate.[55] But the problem is not only lower EU capabilities. It is also a mismatch between the speed and autonomy of the emerging threat environment and a European framework still organized around slower coordination and reactive responses. The fundamental question is whether EU governance structures are equipped for this new environment.

Toward a More Robust EU Approach

The EU's dedicated security frameworks for autonomous AI systems must go beyond traditional product certification. Current certification-based approaches are designed mainly to assess systems before they are deployed: whether a product meets certain standards at the point of market entry, whether risks have been documented, and whether formal compliance obligations have been satisfied. Those tools are useful, but they do not capture the way an autonomous system behaves after deployment, once it interacts with new data, external tools, human users, and adversarial inputs in real time. These systems do not behave like conventional software; they can act independently, adapt to new situations, and operate at high speed. This means oversight cannot end with approval or market entry. It must also include runtime monitoring, clear limits on what agents are allowed to do, logging of agent actions, incident reporting for autonomous failures, and mechanisms to suspend or restrict systems whose behavior changes in operation.

The increasing normalization of offensive cyber operations, including through careful framing in the EU, will also have to be assessed against the normative foundations of the existing international framework for cyber governance.[56] The EU's positions at the United Nations have stressed the view that when cyber operations take place in the context of armed conflict or produce effects comparable to those of conventional attacks, international humanitarian law continues to apply.[57] This includes core principles such as distinction, proportionality,

and precaution in attack, regardless of whether the effect is delivered through code or kinetic force. A U.S. stance that considers offensive cyber action a routine tool of statecraft and blurs the boundaries between state and private actors matters for Europe, not only as an issue of alliance management, but also as one of governance. If, indeed, agentic AI becomes the preferred tool for state-level offensive cyber activity, the urgency of the EU's regulatory and capacity gaps intensifies sharply.

The autonomy of cyber capabilities also complicates questions of international law and state responsibility.[58] Existing frameworks assume that cyber operations are conducted by human-controlled actors whose intent and direction can be attributed to a state. Autonomous agents complicate this assumption. If an AI system independently identifies targets, develops exploits, or escalates an operation beyond its original parameters, determining responsibility becomes much more complex. While states remain legally responsible for the cyber operations they deploy, the evidentiary and attribution challenges associated with autonomous systems could make diplomatic responses and accountability mechanisms far more difficult to enforce.

What is more, the strategic challenge for states extends beyond frontier capability and attribution to include uneven resilience, meaning that the capacity to prevent, absorb, and recover from cyber disruption is distributed unequally across society and critical sectors. Utilities, hospitals, municipalities, and small critical-infrastructure operators are likely to be the most vulnerable points in an agentic threat environment, because they often have fewer technical resources, less redundancy, and weaker incident-response capacity than large state or private-sector actors. They also require targeted support beyond generic compliance obligations.

The Due Diligence Guidance for Responsible AI of the Organisation for Economic Co-operation and Development offers a useful framework to identify, evaluate, prevent, and mitigate adverse impacts across the whole AI value chain, including deployers, developers, infrastructure providers, and financiers.[59] Continuous monitoring, clear restrictions on what autonomous systems can do, and meaningful human supervision for sensitive or high-impact actions are essential. ENISA, with its enhanced mandate and resources, could play a key role in developing practical guidance and standards in this field.

But due diligence alone will not be enough. Europe also needs an operational doctrine, procurement criteria, and minimum runtime security requirements for agents deployed in critical sectors. Otherwise, the EU risks building a sophisticated accountability framework around a category of systems whose most dangerous behavior emerges only after they have been deployed in live environments, connected to everyday infrastructures, and allowed to undertake autonomous operations across real systems and data flows.

Finally, considering the EU's historical role setting global standards, it could consider reprising that role for AI norms. The bloc could push aggressively for international constraints on autonomous cyber weapons, analogous to frameworks for autonomous lethal-weapons

systems. But this only works if it is backed by credible deterrence, which is linked to Europe's capability gap. As some states adopt more assertive stances with offensive cyber capabilities, the risk of escalation and collateral damage grows.

In this respect, the EU should use its diplomatic influence to advocate clear rules that protect civilian infrastructure and prevent destabilizing behaviors in cyberspace. The union should also call for clear redlines on autonomous cyber operations that target civilian critical infrastructure, hospitals, and systems involved in, for instance, nuclear command, control, and communications. At a minimum, high-risk deployments ought to require political approval and documented risk assessments, rather than be left to operational discretion or vendor-level safety commitments.

Conclusion

The governance challenge posed by agentic AI extends well beyond cybersecurity. What this paper has described is, in many ways, only the beginning of a broader regulatory reckoning driven by the spread of autonomous agents throughout the digital ecosystem. Agentic AI does not merely accelerate existing cyber threats; it also has the potential to disrupt regulatory frameworks in areas far removed from network security. In each of these fields, the regulatory gap is structural: Frameworks designed for human agency struggle to manage systems capable of acting independently.

The implication is that Europe will need to review almost its entire digital regulatory framework—not just its cybersecurity laws but also its financial regulations, platform governance, labor protections, and media policies—in light of the realities of autonomous AI. Moreover, the speed and scale at which agentic systems operate suggest that enforcement and countermeasures may need to be partly automated, raising a new and largely unexplored public policy question: how to govern the deployment of defensive agentic AI against hostile agentic AI.

This shift does not mean that human operators will disappear from the picture. States, criminal organizations, and private-sector actors will continue to set objectives and deploy capabilities. But the operational layer of cyber conflict may increasingly be delegated to autonomous software systems that can act faster, at greater scale, and with less direct supervision than previous generations of cyber tools. As this transition unfolds, many of the governance mechanisms that underpin today's cybersecurity frameworks, from attribution practices to certification regimes, will face growing strain.

For the EU, the challenge is that its existing frameworks remain anchored in a model of cybersecurity centered on human adversaries and relatively predictable software systems. Agentic AI introduces a different operational logic. Closing this governance gap will require more than incremental adjustments to existing legislation. It will first need a clearer European view of the proper role of agentic systems in cybersecurity and beyond: which functions can be delegated to autonomous agents, where autonomy should be tightly circumscribed, and which decisions should remain under meaningful human control.

Those broader governance choices need to come first, because they determine what operational rules are needed and where stricter limits should apply. Security oversight must extend beyond the design and certification of AI systems to the runtime behavior of autonomous agents that operate in real-world environments. Organizations that deploy these systems will need clearer obligations for monitoring, logging, and restricting the capabilities of AI agents that interact with sensitive data and infrastructure.

In practice, that means linking high-level political choices about the acceptable role of AI agents to concrete requirements for deployment: what systems agents may access, what actions they may take, what forms of human approval are required, and when autonomous operations must be slowed, suspended, or blocked altogether. Technical standards must also evolve to address questions that were previously peripheral to cybersecurity policy, including how autonomous agents are authenticated, how their actions are recorded, and how malicious manipulation of agent behavior is detected.

As Washington increasingly integrates AI-enabled capabilities into its cyber strategy, the EU risks finding itself both technologically dependent and strategically misaligned. In such a landscape, regulatory leadership alone will not guarantee influence over how these systems are used or governed. The EU needs more than better coordination among existing rules. It needs a tighter governance model for autonomous cyber systems and a more explicit strategy for reducing dependence on U.S. frontier AI providers. Otherwise, the union risks regulating the margins of a technological ecosystem whose most consequential security choices are being made elsewhere.

The Moltbook experiment offered an early glimpse of a digital environment populated largely by autonomous agents interacting with one another. What appeared as a curiosity may soon become a common feature of internet infrastructure. As autonomous systems take on increasingly consequential roles in cybersecurity, defense, and the economy, the central policy challenge will be to ensure that the expansion of machine agency does not outpace the human institutions responsible for governing it.

Notes

1 "A Social Network for AI Agents," Moltbook, https://www.moltbook.com/?ref=kinakomemo.com.

2 Chris Stokel-Walker, "Moltbook, the Viral Social Network for AI Agents, Has a Major Security Problem," FastCompany, February 3, 2026, https://www.fastcompany.com/91485597/moltbook-the-viral-social-network-for-ai-agents-has-a-major-security-problem.

3 Gal Nagli, "Hacking Moltbook: The AI Social Network Any Human Can Control," Wiz, February 2, 2026, https://www.wiz.io/blog/exposed-moltbook-database-reveals-millions-of-api-keys.

4 "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign," Anthropic, November 2025, https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf.

5 Chiara Barbeschi and Tarik Fayad, "Anthropic's Mythos Moment: How Frontier AI Is Redefining Cybersecurity," World Economic Forum, April 20, 2026, https://www.weforum.org/stories/2026/04/anthropic-mythos-ai-cybersecurity/.

6 "Expanding Project Glasswing," Anthropic, June 2, 2026, https://www.anthropic.com/news/expanding-project-glasswing.

7 "Our Evaluation of Claude Mythos Preview's Cyber Capabilities," AI Security Institute, April 13, 2026, https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities.

8 "Statement on the US Government Directive to Suspend Access to Fable 5 and Mythos 5," Anthropic, June 12, 2026, https://www.anthropic.com/news/fable-mythos-access.

9 Jam Kraprayoon et al., "Highly Autonomous Cyber-Capable Agents: Anticipating Capabilities, Tactics, and Strategic Implications," Institute for AI Policy and Strategy, March 2026, https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/69b1709a79b9076e980f8dc9/1773236378439/Highly+Autonomous+Cyber-Capable+Agents_+_+Anticipating+Capabilities%2C+Tactics%2C+and+Strategic+Implications.pdf.

10 Beth Stackpole, "Agentic AI, Explained," MIT Sloan School of Management, February 18, 2026, https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained.

11 "Agentic AI in Cybersecurity," Rapid7, https://www.rapid7.com/fundamentals/agentic-ai/.

12 Nilantha Prasad et al., "A Survey of Cyber Threat Attribution: Challenges, Techniques, and Future Directions," *Computers and Security* 157 (2025), https://doi.org/10.1016/j.cose.2025.104606.

13 Mohamed Amine Ferrag et al., "Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities," *Internet of Things and Cyber-Physical Systems* 5 (2025): 1–46, https://doi.org/10.1016/j.iotcps.2025.01.001.

14 Sahaya Jestus Lazer et al., "A Survey of Agentic AI and Cybersecurity: Challenges, Opportunities and Use-case Prototypes," arXiv, January, 8, 2026, https://arxiv.org/html/2601.05293v1; Lucia Stanham, "What Is Agentic AI," CrowdStrike, May, 1, 2025, https://www.crowdstrike.com/en-us/cybersecurity-101/artificial-intelligence/agentic-ai/; and "Technical Blog: Strengthening AI Agent Hijacking Evaluations," Center for AI Standards and Innovation, National Institute of Standards and Technology (NIST), January 17, 2025, https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations.

15 Thilo Hagendorff, Erik Derner, and Nuria Oliver, "Large Reasoning Models Are Autonomous Jailbreak Agents," *Nature Communications* 17 (2026), https://www.nature.com/articles/s41467-026-69010-1.

16 Kraprayoon et al., "Highly Autonomous."

17 Tom Everitt et al., "Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective," arXiv, March 26, 2021, arXiv:1908.04734.

18 Maia Hamin and Benjamin Edelman, "Cheating on AI Agent Evaluations," NIST Center for AI Standards and Innovation, December 2, 2025, https://www.nist.gov/caisi/cheating-ai-agent-evaluations.

19 Victoria Krakovna et al., "Specification Gaming: The Flip Side of AI Ingenuity," Google DeepMind, April 21, 2020, https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

20 Jonathan Reed, "AI Cybersecurity Solutions Detect Ransomware in Under 60 Seconds," IBM, https://www.ibm.com/think/insights/ai-cybersecurity-threat-detection-ransomware.

21 Carey Frey, "Why Agentic AI Forces a Reckoning with Identity," Palo Alto Networks, https://www.paloaltonetworks.com/perspectives/why-agentic-ai-forces-a-reckoning-with-identity/.

22 "Software and AI Agent Identity and Authorization," NIST National Cybersecurity Center of Excellence, https://www.nccoe.nist.gov/projects/software-and-ai-agent-identity-and-authorization.

23 Maya Derrick, "How Anthropic Disrupted a World-First AI Cyber Espionage," *AI Magazine*, February 5, 2026, https://aimagazine.com/news/how-anthropic-disrupted-ai-cyber-espionage.

24 "Disrupting," Anthropic.

25 "Technical Blog," Center for AI Standards and Innovation.

26 "Disrupting," Anthropic.

27 "AI Misuse Exposed as OpenAI Details Global Disinformation and Scam Networks," Geneva Internet Platform, February 26, 2026, https://dig.watch/updates/openai-threat-report-chatgpt-misuse.

28 "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors," OpenAI, February 14, 2024, https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/.

29 Nate Nelson, "ChatGPT Jailbreaking Forums Proliferate in Dark Web Communities," Dark Reading, September 12, 2023, https://www.darkreading.com/application-security/chatgpt-jailbreaking-forums-dark-web-communities.

30 Mohamed Fazil Mohamed Firdhous et al., "WormGPT: A Large Language Model Chatbot for Criminals," Institute of Electrical and Electronics Engineers, March 18, 2024, https://ieeexplore.ieee.org/document/10453752.

31 Mariarosaria Taddeo, "Agentic AI Is the Hacker's New Accomplice," *Financial Times*, November 27, 2025, https://www.ft.com/content/9966d9e8-7fd3-4324-b57e-f02763795d29.

32 "Agentic Misalignment: How LLMs Could Be Insider Threats," Anthropic, June 20, 2025, https://www.anthropic.com/research/agentic-misalignment; and Matt Sutton and Damian Ruck, "Indirect Prompt Injection: Generative AI's Greatest Security Flaw," Alan Turing Institute, November 1, 2024, https://cetas.turing.ac.uk/publications/indirect-prompt-injection-generative-ais-greatest-security-flaw.

33 Lily Jamali, "US Threatens Anthropic with Deadline in Dispute on AI Safeguards," BBC, February 24, 2026, https://www.bbc.co.uk/news/articles/cjrq1vwe73po.

34 "Where Things Stand with the Department of War," Anthropic, March 5, 2026, https://www.anthropic.com/news/where-stand-department-war.

35 Raluca Csernatoni, "Governing Military AI amid a Geopolitical Minefield," Carnegie Europe, July 17, 2024, https://carnegieendowment.org/russia-eurasia/research/2024/07/governing-military-ai-amid-a-geopolitical-minefield.

36 Hadas Gold, "OpenAI Strikes Deal with Pentagon Hours After Trump Admin Bans Anthropic," CNN, February 28, 2026, https://edition.cnn.com/2026/02/27/tech/openai-pentagon-deal-ai-systems.

37 Liv McMahon, "Microsoft Error Sees Confidential Emails Exposed to AI Tool Copilot," BBC, February 19, 2026, https://www.bbc.co.uk/news/articles/c8jxevd8mdyo.

38 "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," NIST, July 2024, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

39 "Learn About Using Microsoft Purview Data Loss Prevention to Protect Interactions with Microsoft 365 Copilot and Copilot Chat," Microsoft, https://learn.microsoft.com/en-us/purview/dlp-microsoft365-copilot-location-learn-about.

40 Beliz Kaleli et al., "Fooling AI Agents: Web-Based Indirect Prompt Injection Observed in the Wild," Unit 42, March 3, 2026, https://unit42.paloaltonetworks.com/ai-agent-prompt-injection/.

41 Simon Willison, "The Lethal Trifecta for AI Agents: Private Data, Untrusted Content, and External Communication," *Simon Willison's Weblog*, June 16, 2025, https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/.

42 "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence," Official Journal of the European Union, July 12, 2024, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng; and "Proposal for a Regulation of the European Parliament and of the Council on the European Union Agency for Cybersecurity (ENISA)," European Commission, January 20, 2026, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52026PC0011.

43 "Artificial Intelligence and Machine Learning: Supply Chain Risks and Mitigations," Australian Signals Directorate, October 16, 2025, https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/artificial-intelligence-and-machine-learning-supply-chain-risks-and-mitigations.

44 "Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity)," Official Journal of the European Union, June 7, 2019, https://eur-lex.europa.eu/eli/reg/2019/881/oj.

45 "Our Evaluation," AI Security Institute.

46 Stephanie K. Pell, "Trump 2.0: What Cybersecurity Shifts Lie Ahead?," Brookings, December 9, 2024, https://www.brookings.edu/articles/trump-2-0-what-cybersecurity-shifts-lie-ahead/.

47 "Treasury Sanctions Company Associated with Salt Typhoon and Hacker Associated with Treasury Compromise," U.S. Department of the Treasury, January 17, 2025, https://home.treasury.gov/news/press-releases/jy2792.

48 Matthias Schulze, "How European and Allied Cybersecurity Strategies Are Shifting from Defence to Offence," Binding Hook, January 27, 2026, https://bindinghook.com/how-european-and-allied-cybersecurity-strategies-are-shifting-from-defence-to-offence/.

49 Adam Sella, "U.S. Weighs Expanding Private Companies' Role in Cyberwarfare," *New York Times*, January 14, 2026, https://www.nytimes.com/2026/01/14/us/politics/us-cyberwarfare-private-companies.html.

50 "President Trump's Cyber Strategy for America," The White House, March 2026, https://www.whitehouse.gov/wp-content/uploads/2026/03/president-trumps-cyber-strategy-for-america.pdf.

51 Alexandra Paulus, "Europe's Cybersecurity Depends on the United States: Europe Can and Must Do More," German Institute for International and Security Affairs (SWP), November 5, 2025, https://www.swp-berlin.org/en/publication/europes-cybersecurity-depends-on-the-united-states.

52 Max Smeets, "U.S. Cyber Strategy of Persistent Engagement & Defend Forward: Implications for the Alliance and Intelligence Collection," *Intelligence and National Security* 35, no. 3 (2020): 444–453, https://doi.org/10.1080/02684527.2020.1729316.

53 Paulus, "Europe's Cybersecurity."

54 "European Union and NATO Hold the First Structured Dialogue on Cyber," European External Action Service, October 4, 2024, https://www.eeas.europa.eu/eeas/european-union-and-nato-hold-first-structured-dialogue-cyber-0_en.

55 "EU Policy on Cyber Defence," European Commission, November 10, 2022, https://www.eeas.europa.eu/sites/default/files/documents/Comm_cyber%20defence.pdf.

56 Antoaneta Roussi, "Europe Needs Cyber Weapons, Says EU Tech Chief," *Politico*, February 13, 2026, https://www.politico.eu/article/europe-needs-offensive-cyber-power-says-eu-tech-chief/; and Louise Marie Hurel et al., "Global Compendium on Responsible Cyber Behaviour," Royal United Services Institute, March 2025, https://static.rusi.org/global-compendium-rcb.pdf.

57 Patryk Pawlak and Thomas Biersteker, "EU Cyber Sanctions and Norms in Cyberspace," EU Institute for Security Studies, October 2019, https://www.iss.europa.eu/sites/default/files/EUISSFiles/cp155.pdf.

58 Katie A. Johnston, "Artificial Intelligence and the 'Armed Attack' Threshold in International Law," *International Law Studies* 107, no. 45 (2026): 45–75, https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=3135&context=ils; and Bérénice Boutin, "State Responsibility in Relation to Military Applications of Artificial Intelligence," *Leiden Journal of International Law* 36, no. 1 (2023): 133–150, https://doi.org/10.1017/S0922156522000607.

59 "OECD Due Diligence Guidance for Responsible AI," Organisation for Economic Co-operation and Development, February 19, 2026, https://www.oecd.org/en/publications/oecd-due-diligence-guidance-for-responsible-ai_41671712-en/full-report/component-4.html.

CARNEGIE
EUROPE

Rue du Congrès, 15
1000 Brussels, Belgium
CarnegieEurope.eu