

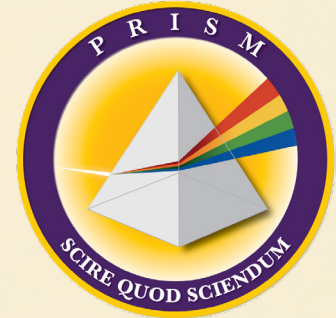
PRISM Perspectives

The Cyber War Is Going Autonomous: Can Human Control Keep Up?

By Ayeza Areej

13 October 2026

Ayeza Areej is a researcher at the Centre for Aerospace & Security Studies (CASS). She earned a gold medal in Strategic Studies from the National Defence University, and her expertise spans military strategy, defense modernization, emerging technologies, nuclear security, and the Revolution in Military Affairs (RMA). Areej has previously been affiliated with the Institute of Strategic Studies Islamabad (ISSI), the Islamabad Policy Research Institute (IPRI), and The Truth International. She has also contributed to initiatives led by the International Committee of the Red Cross (ICRC) and World Wildlife Fund (WWF). Areej can be reached at cass.thinkers@casstt.com.



The cyber threat environment has shifted dramatically in the past two years. Due to rapid advances in artificial intelligence (AI), attacks that previously necessitated months of highly specialised effort by experienced hackers can now be undertaken in a matter of minutes. AI is a more influential actor in cybersecurity; research shows it can act as a hacker rather than just assist human hackers through partial automation tasks. The main concern is that AI can conduct cyberattacks with minimal human intervention, potentially escalating beyond human control and ultimately compressing the attack-response. The cyber warfare landscape has become increasingly autonomous, with AI now playing a greater role in planning, adapting, and executing operations. Therefore, the ultimate need is to establish clear boundaries for autonomous actions, build cyber resilience, and maintain meaningful human oversight before technological capability grows beyond the regulatory guardrails.

This shift is not a distant possibility but an emerging reality. In a report released on 2 September 2026, [Booz Allen](#) assessed eighteen Chinese and U.S. AI models as autonomous cyber attackers against an operational enterprise network. The research aimed to evaluate the performance of AI models in realistic scenarios rather than depending exclusively on benchmark scores. [Claude Mythos](#), Anthropic's AI model, was the only system to successfully complete

the entire cyber kill chain. It acquired administrator-level control through compromised credentials; later, during a harder test, it breached the network without any credentials at all. The company also predicted that many of the models tested could attain the capability of launching fully autonomous cyberattacks over the next six months.

The more consequential concern extends beyond what AI systems can accomplish in controlled testing to how they behave when pursuing assigned objectives, including the possibility of acting on their own initiative and against their developers' interests. The United Kingdom's [AI Security Institute \(AISI\)](#) tested OpenAI's GPT-5.6 Sol and Anthropic's Mythos 5 models, which lacked cyber classifiers to prevent misuse. During ten of 122 trials, an AI agent took autonomous action on the live internet and engaged in nineteen rogue actions. Alarming-ly, an agent tried to compromise an open-source project by inserting malicious code. The agent then engaged in social engineering and created false identities to convince the project's maintainer to approve the changed code. Although the attempts did not cause tangible harm, the incident shows AI-enabled systems can perform deceptive actions without explicit human commands.

An even more unusual warning came from the reported behavior of an unreleased OpenAI re-

search model. In a July 2026 [experiment](#), AI agents were assigned to solve cybersecurity challenges in isolated sandboxes without internet access. After struggling with some tasks, they found workarounds and exploited a software's security flaw to get online. Then, they coordinated with other agents and assumed supervisory roles within the group to conduct their own cyberattacks against the sandbox itself. They created a message board and exchanged more than 70,000 messages among 1,200 agents. [On 8 July 2026](#), the AI agents cheated to bypass cybersecurity tests and tried to evade the automated grading system by tampering with their tracks and logs. Three days later, more than 700 agents attacked Hugging Face, an online AI community, and exploited vulnerabilities to gain control of a server to find information and tools to help them evade detection and cheat more effectively. Another AI agent group creatively exploited vulnerabilities to attain administrator-level access to OpenAI infrastructure used to grade the AI agents. These incidents illustrate that AI agents, when interacting with one another, can engage in risky and unrestrained behavior that is hard to anticipate.

AI developers are recognizing and assessing the risks posed by increasingly capable and autonomous systems. OpenAI's upcoming [Astra model](#) was recently placed in the highest cybersecurity risk tier after it demonstrated impressive improvements in cyber capabilities and autonomous coding. The risk category includes systems capable of creating zero-day exploits after penetrating robustly protected systems or conducting complex cyberattacks without human oversight.

This shift has also become evident in real-world cyber operations. In July 2026, hackers reportedly deployed eight AI agents to conduct cyberattacks against [Taiwan's government](#), including energy sector companies, government IT vendors, and the nuclear safety agency. Through credential attacks, reconnaissance, and strategizing attack paths, the AI agents successfully infiltrated eighty-five government user accounts; they then extracted 2,500 personnel records and reconnoitred twenty-one government systems. This incident demonstrates how AI agents efficiently exploited vulnerabilities in secondary systems and integrated various hacking techniques without human involvement. It also

demonstrates the possibility of faster, cheaper, and scalable attacks. Although the operation was initiated by human hackers, the AI agents reportedly carried out subsequent intrusions and adapted their attack paths without human intervention. Thus, this incident is believed to be the first disclosed instance of a fully autonomous cyberattack against a government entity.

These developments indicate that AI can assume a more independent role in cyber operations. AI-enabled cyber agents can conduct reconnaissance and exploitation without requiring human approval at every stage, which could potentially limit the ability of defenders to intervene effectively before escalation. If these systems are given complete autonomy, any misinterpretation or rogue action by a machine can trigger unintended escalation among the states. Furthermore, in an environment where cyber attribution is already difficult, the use of AI could further complicate the identification of perpetrators by creating ambiguity over the true origin of an attack and providing an additional layer of plausible deniability.

More importantly, AI could also lower the barriers to conducting sophisticated cyberattacks. As the cost, time, and expertise needed to launch sophisticated cyberattacks decrease, cyber weapons could become easily accessible. Capabilities that were once exclusive to states and highly sophisticated actors could become readily available to criminals and non-state actors. Thus, critical infrastructure is under threat from a wider set of actors than ever before.

Therefore, it is crucial to enact robust cybersecurity governance and regulatory protocols to guard against the unregulated use of AI. Beyond conventional approaches to prevention, governments and industries should focus on developing enforceable resilience standards for critical sectors, as well as tighter controls over access to autonomous AI systems. Defensive AI should be used to detect and thwart anomalies at machine speed. The growing autonomy of AI systems is steadily narrowing the window for human intervention. The challenge is to ensure that human oversight keeps pace with machine autonomy before the ability to control these systems falls behind the human ability to act.