Preventing AI-Driven Cyberattacks Targeting Critical Infrastructure

General Assembly 1

Background

Critical infrastructure including energy grids, water systems, telecommunications, transportation networks, healthcare facilities, and financial services underpins national security, economic stability and public welfare. These systems increasingly depend on digital technologies and networked control systems, making them attractive targets for cyberattacks. Emerging artificial intelligence (AI)-driven cyber threats amplify traditional risks by enabling attackers to automate reconnaissance, craft highly convincing phishing campaigns, mutate malware, and adapt attacks in real time, lowering the technical threshold for sophisticated intrusions. In a recent survey, 87 % of U.S. critical infrastructure organizations reported concern about AI-powered threats, including adaptive AI malware and automated hacking tools.

As AI capabilities advance further, experts warn that “agentic malware”, autonomous AI agents that can execute multi-stage attacks, could materialize within a few years, posing disruptive risks to power grids, ports, and other networked systems.

Current stance

States and multilateral institutions increasingly recognize cyber insecurity as a collective threat to peace and stability. Many governments are bolstering cyber defenses, investing in AI-enhanced threat detection and resilience measures, and cooperating on norms and information-sharing. NATO is testing AI tools to detect and respond to simulated attacks on power systems and grids, reflecting growing interest in defensive AI applications. At the same time, accountability and regulatory frameworks are lagging: there is no binding global treaty specifically addressing AI-enabled cyber threats to critical infrastructure, and enforcement of existing norms, such as those in the Budapest Convention on Cybercrime, varies widely. International sanctions regimes (e.g., EU measures targeting individuals responsible for cyberattacks) demonstrate political will to deter malicious activity.

Most affected states/areas:

Highly networked economies with extensive critical infrastructure are prime targets. The United States, with both advanced infrastructure and frequent state-linked cyber campaigns reported against it, remains a focal point. Singapore has publicly acknowledged cyber espionage targeting its infrastructure. Middle-income and developing countries with digitalization gaps, including in energy and water systems, also face rising threats. Urban centers globally are likewise vulnerable due to concentrated digital assets and interdependent systems.

Involved actors

A broad ecosystem engages this issue; think of National cybersecurity agencies (e.g., U.S. CISA, European Union Agency for Cybersecurity, also ENISA) coordinate defenses and incident reporting. More organisations include NATO (conducts AI security experiments and shares best practices), private sector cybersecurity firms (contribute threat intelligence, detection tools, and incident response), international forums such as the UN Group of Governmental Experts (GGE) on cybersecurity, the Paris Call for Trust and Security in Cyberspace (advocate norms against harmful cyber operations) and research institutions (explore AI-based defenses and simulated attack frameworks).

Note

AI-driven cyberattacks on critical infrastructure can disrupt essential services, endanger public safety, and undermine trust in digital and physical systems. Preventing these threats requires coordinated multilateral action, including norm development, capacity-building, information-sharing, resilient system design, and ethical AI governance.

Please let this be a guideline and feel encouraged to put forward your own proposal.

Sources

87 percent of US critical infrastructure organizations concerned about AI-Powered cyberthreats -- Security Today. (2024, May 30). Security Today. https://securitytoday.com/articles/2024/05/30/87-percent-of-us-critical-infrastructure-organizations-concerned-about-aipowered-cyberthreats.aspx
AI will supercharge cyber weapons within two years, experts warn. (January 7th, 2025). Axios. https://www.axios.com/2025/01/07/goldilock-agentic-malware-2027-doomsday
Mhill. (2025, May 8). *NATO tests AI's ability to protect critical infrastructure against cyberattacks*. CSO Online. https://www.csoonline.com/article/574281/nato-tests-ai-s-ability-to-protect-critical-infrastructure-against-cyberattacks.html
Timeline - Sanctions against cyber-attacks. (n.d.). European Council. https://www.consilium.europa.eu/nl/policies/sanctions-against-cyber-attacks/timeline-sanctions-cyber-attacks
AI-Driven Cybersecurity Testbed for Nuclear Infrastructure: Comprehensive Evaluation Using METL Operational Data. (December 1st, 2025). Arxiv. https://arxiv.org/abs/2512.01727