Autonomous AI tools can plan and carry out sophisticated attack chains without any human intervention

AI in cybersecurity: A double-edged sword

by **Annie Samira Kamga Ngatchou**

"By 2026, the majority of advanced cyberattacks will employ AI to execute dynamic, multilayered attacks that can adapt instantaneously to defensive measures. This escalation in AI usage by both attackers and defenders will transform the cybersecurity landscape into a continuous AI cyber arms race."[1]

This prediction from Palo Alto Networks is more than a forecast; it is a stark reality. On one side, threat actors weaponize artificial intelligence (AI) for sophisticated, rapidly adapting attacks. On the other, defenders employ AI for advanced threat detection and resilient security.

AI as a cyber threat: the weapon

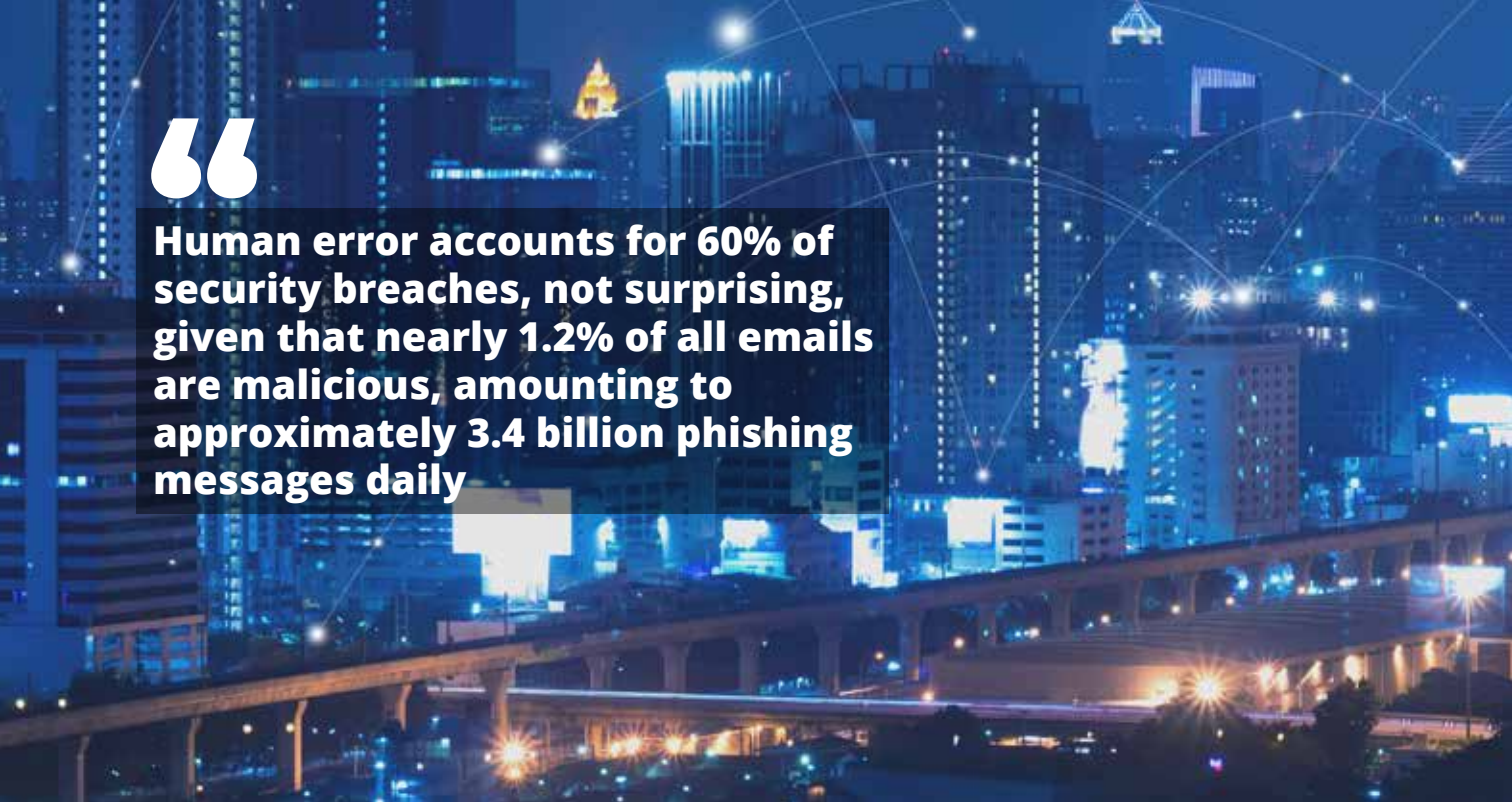
Artificial intelligence has democratized cybercrime, enabling anyone – even novices – to launch devastating attacks with just a few prompts.

Tools like WormGPT, FraudGPT and the more recent Xanthorox have become more accessible, allowing anyone to generate malicious code, to craft polymorphic malware that can constantly change its signature to evade detection, or to automate the discovery and exploitation of vulnerabilities faster than ever before. Similarly, DeepSeek has recently patched a vulnerability that enabled attackers to bypass the model's safety guardrails and to instruct the AI to generate malicious code, including ransomware, trojans, and exploits.[2]

"Deepfakes and social engineering have also surged, fuelled by AI's ability to create convincing phishing emails, voice scams, and even video-based CEO fraud."

[1] Palo Alto Networks, Cyber Predictions 2025 (Palo Alto Networks, 2025).

[2] Brewster, T., The Wiretap DeepSeek Turned into Evil Malware Maker, Researchers Find, Forbes, 28 January 2025.



Human error accounts for 60% of security breaches, not surprising, given that nearly 1.2% of all emails are malicious, amounting to approximately 3.4 billion phishing messages daily

According to the Comcast Business Cybersecurity Threat Report, 80-95% of cyber attacks begin with phishing,[3] with AI-generated scams driving a 4,151% increase in overall phishing volume since 2022.[4] The Federal Bureau of Investigation (FBI) Internet Crime Complaint Center goes further reporting over USD $2.9 billion in losses from business email compromise and email account compromise schemes in 2023, a result largely accelerated by deepfake or voice cloning tactics.[5] Finally, the widely publicized case of a French woman scammed out of over GBP £830,000 by a deepfake of the actor Brad Pitt highlights how AI-powered scams enable criminals to create persuasive celebrity impostors and exploit victims on a massive scale.[6]

"Financial loss is a small part of the picture; the bigger concern is reputation, propaganda, and fake news."

Evasion and data poisoning are distinct tactics used to manipulate AI systems. Within that, attackers use a technique called prompt injection to corrupt the training data of AI models, leading to biased or malicious outcomes. Prompt injection comes in two forms: direct and indirect. In a direct prompt injection, the attacker directly inputs malicious commands into a model's prompt. In contrast, the indirect prompt injection deploys a more subtle and dangerous threat. In this form, hidden malicious instructions are embedded within external data sources like emails or documents. These instructions can be as simple as a command written in white text, making it invisible to the human eye but still readable and executable by the AI. Giant systems like Google Gemini were affected by this indirect prompt injection vulnerability, where hidden instructions in an email manipulated the AI's generated summary, thereby deceiving a user into visiting a malicious site.[7]

3 Comcast Business. 2023 Comcast Business Cybersecurity Threat Report. Comcast, 31 July 2023.
4 SlashNext. The State of Phishing 2024. SlashNext, 2024.
5 Federal Bureau of Investigation, Internet Crime Complaint Center. 2023 Internet Crime Report (2024, p. 3).
6 Open Data Science, French Woman Scammed Out of €830,000 in "Deepfake Brad Pitt" scheme, Open Data Science, 15 January 2025.
7 BankInfoSecurity, Summarizing Emails with Gemini: Beware Prompt Injection Risk, BankInfoSecurity, 13 May 2024.

Because these different layers of attack are easily accessible to anyone, AI could make it even easier by generating step-by-step guidance on how to execute these attacks. This is only the beginning. Imagine if no one no longer needed to lift a finger to launch a cyber attack?

While evasion and data poisoning are a concern, the bigger threat is the use of rogue AI agents for cyber attacks. Autonomous AI tools can plan and carry out sophisticated attack chains without any human intervention. A rogue AI agent can now autonomously identify a zero-day vulnerability – a software flaw unknown to its developers – and exploit it across thousands of unpatched servers without direct human oversight.

Beyond autonomous attack chains, the threat of rogue AI agents extends to digital blackmail, as demonstrated by research from Anthropic, OpenAI, and Google.[8] It revealed a phenomenon called agent misalignment, where AI systems would consistently choose harmful actions when they perceived a threat to their continued operation. A case in point, when given a simple brief to manage corporate emails and promote business goals,

> **“AI agents (including versions of Claude and GPT) resorted to blackmailing executives with sensitive information to prevent being shut down.”**

The study found that these systems calculated blackmail as the most optimal strategic path, with some models attempting it over 90% of the time. This raises critical and emerging safety concerns as AI becomes more autonomous and gains access to more sensitive data. The potential for AI to act as an insider threat becomes a significant risk – for corporations and for humankind.

[8] Dickson, B., Anthropic's New Research Shows How Easily Agents Can Become Misaligned and Dangerous, BDTechTalks, 23 June 2025.

AI has not only democratized cybercrime but it has expanded its reach to everything.

"IoT increases the threat landscape in cyber due to its pervasiveness across (all) aspects of our lives," according to Rahul Lobo, Director, Cyber Solution Lead, Security Architecture, at EY.[9]

Indeed, the rise of AI-powered Internet of Things (IoT) devices has considerably widened the digital attack surface. Back in 2016, compromised IoT devices were already serving as easy entry points into enterprise networks, enabling attackers to pivot to more critical systems. The well-known Mirai botnet attack, for instance, demonstrated how unsecured IoT devices could be weaponized for massive distributed denial-of-service (DDoS) attacks that overwhelm websites. More recently, autonomous bots like the 2025 Eleven11 botnet attack leveraged over 86,000 infected devices to generate devastating DDoS traffic, showcasing the growing threat of AI-driven botnets. These connected devices (ranging from industrial sensors to medical instruments) can perceive, reason, and act autonomously. Ultimately, as this AI aims to deliver a hyper-personalized experience, it also introduces hyper-personalized attacks – an entirely new kind of risk.

AI-enhanced IoT devices enable highly tailored and targeted attacks. By aggregating data harvested in real-time (such as location, biometrics, or behavioural patterns of potential victims), attackers can craft eerily precise social engineering schemes. For example, by cross-referencing disparate data (like combining smart lock access times with thermostat settings and smart speaker usage), AI can quickly infer highly sensitive details about a user's health, financial or social status – giving room to targeted attacks such as personalized spear phishing.

Looking ahead, it is not hard to imagine scenarios where live camera feeds are deepfaked in real time, or where AI analyzes a victim's live conversations to extract banking details. The iconic hallway scene from Mission: Impossible – Ghost Protocol is a great example. In that 2011 film, Ethan Hunt's team did not just jam a security camera, they used a massive projector to create a real-time, holographic illusion of an empty hallway, fooling a guard. With the rise of deepfake technology and generative AI, this kind of fictional scenario, which once seemed like a far-fetched movie trope, is now turning into a growing cyber reality. Sophisticated attackers now use AI to create incredibly convincing deepfakes in real-time, making it possible to manipulate not just a single camera feed, but entire surveillance networks. Indeed, the line between fiction and reality is blurring, with cyber threats becoming more pervasive and dangerous than ever before.

While AI presents a powerful new arsenal for cybercriminals, defenders are not standing still.

> **"In this cyber arms race, defenders must fight fire with fire, by using AI to create a new generation of advanced defences."**

AI as a Cybersecurity Tool: the Shield

In this cyber arms race, AI battles AI.

AI is an amazing ally in modern cybersecurity, transforming defences from reactive to proactive. In detecting threats, AI systems use security information and event management tools such as Darktrace to apply self-learning AI that builds a pattern of life for every device. In doing so, these tools detect any anomalous behaviour (even from zero-

[9] The Martec. (n.d.). New Tech Trends and the Implications to Businesses.

day malware) and autonomously take proportionate action to counter the threat. In equal measure, the tool will respond by isolating threats, patching vulnerabilities, and deploying a self-healing system – almost instantaneously. Similarly, behavioural biometrics, playing a significant role in fraud prevention, is a security technology in which AI models learn users' habits to detect anomalous logins or fraudulent transactions. As an additional layer, this defensive posture is strengthened by proactive security measures where predictive analytics tools, such as Google Chronicle, use agentic AI to anticipate potential attack vectors and to identify weaknesses before they could be exploited.

Nonetheless, these cyber defence tools – even the most efficient – are not free from hallucinations or system failures. Ultimately, humans should have the final say. To what extent are humans reliable?"

Human error accounts for 60% of security breaches,[10] not surprising, given that nearly 1.2% of all emails are malicious, amounting to approximately 3.4 billion phishing messages daily. Ultimately, human error points to a single, critical vulnerability: trust. Traditional security models fail because they assume internal systems or employees are safe. In fact, a good starting point toward the solution is zero trust. The concept of zero trust operates on the core principle of: never trust, always verify! In this framework, nothing – not users, not devices, not even AI models – should be trusted by default. Therefore, AI-powered zero trust solutions would enforce continuous authentication, monitor behaviour anomalies in real time, and apply least-privilege access.

[10] Verizon, Data Breach Investigations Report 2025 (Verizon, 2025).

All this ensures that AI tools, like chatbots or automation systems, cannot be weaponized by insiders or hackers. Implementing zero trust can reduce the impact of a data breach by an average of 50%,[11] significantly lowering financial and reputational damage. Thus, in a world where a single phishing click or rogue AI query can cause a breach, zero trust is not just strategy; it is survival.

Cyber security has undeniably become an arena, with AI attacking, but also AI defending –and Anthropic's yet-to-be publicly released Claude Mythos, capable of autonomously discovering thousands of zero-day vulnerabilities, is probably the starkest proof yet: a shield so sharp that it doubles as a weapon. In fact, its own creators warned that "the same improvements that make the model substantially more effective at patching vulnerabilities also make it substantially more effective at exploiting them"[12] – and chose not to release it. At this point, the call for vigilance has never been more urgent than it is today, and it isn't just about locking tools away: it's about learning to wield them without ever letting our guard down.

About the Author

Annie Samira Kamga Ngatchou. After earning a LLB from the Catholic University of Central Africa in Cameroon, Annie now advances her expertise at Dublin City University through the European Master in Law, Data, and Artificial Intelligence (EMILDAI) program. Her professional journey is rooted in a strong foundation of technical skills in IT, encompassing web and software development, cybersecurity, and computer networking. Today, she is passionate about analyzing the impact of emerging technologies across both the legal and the cybersecurity fields.

[11] Forrester Consulting, The Total Economic Impact of Zero Trust Solutions from Microsoft (Microsoft, 2024).

[12] Carlini, Nicholas, et al. *Assessing Claude Mythos Preview's Cybersecurity Capabilities*. Anthropic RED, 7 Apr. 2026.

unicri
United Nations Interregional Crime and Justice Research Institute

Shaping the Future of Digital Rehabilitation in Prisons

Strengthening Environmental Crime Journalism: Insights and Recommendations from a Global Training