Cybercrime in the age of artificial intelligence (AI)

by **Marta Janus**

From deepfakes to automated scams, AI is changing the nature and speed of cybercrime

Rapid advancements in AI technology gave rise to a new era. Like the popularization of the Internet a few decades ago, the widespread adoption of AI-based solutions revolutionized the way we perform our day-to-day tasks, be it in professional or personal settings. As beneficial as these developments are for healthcare, science, and business, we must remember that powerful technologies are equally valuable for threat actors.

Over the last two years, AI has altered the nature and magnitude of cybercriminal activities. It enabled adversaries to automate the development of malware and attack tools, as well as the attacks themselves. It transformed online scams into state-of-the-art engagements in which it is almost impossible to call out deception. It helped spread misinformation in political campaigns, shifting public opinion at adversaries' will. Welcome to 2026, where AI has made cybercrime more successful, efficient, and scalable than ever before.

The art of digital deception

Phishing and digital scams are almost as old as the Internet itself, and despite continuous efforts to raise awareness, they have remained very profitable for cybercriminals. Even imperfect attempts, ridden with typos and grammar mistakes, can yield decent results – all thanks to the power of emotional manipulation and the human tendency to act before thinking. However, the days in which a suspicious email or website could be spotted by a peculiar choice of words and its overall sloppiness are gone now.

> **“Cybercriminals no longer need language, design, or programming skills, as AI chatbots will craft flawless texts, convincing imagery, and professionally looking web interfaces for them.”**

Moreover, AI can help create highly customized campaigns where the phishing content is tailored to each victim and includes personal touches such

Over the last two years, AI has altered the nature and magnitude of cybercriminal activities. It enabled adversaries to automate the development of malware and attack tools, as well as the attacks themselves

as references to the victim's online history, making it much more believable.

Deepfake voice and video take the scam industry to a whole new level, making the worst nightmares come true. Cybercriminals can impersonate anyone with only a handful of photos sourced from the person's social media. The larger the victim's online presence is, the better the quality of the deepfake videos will be. What is worse, these deepfakes can be fully interactive and used in real-time camera calls, as happened in recent financial scams targeting Wire and Plastic Products (WPP) and an unnamed company in Hong Kong.

In the first case, WPP executives were invited to what appeared to be a video meeting with their CEO, in which the CEO requested the transfer of a substantial amount of money. The whole interaction turned out to be a deepfake, and luckily, the WPP staff called it out.[1] The Hong Kong company was not this lucky: after a camera call with deepfake versions of other team members, an initially suspicious employee was convinced to make a payment of about $25M to fraudsters.[2]

Businesses are not the only victims of deepfake scams. Cybercriminals also target individuals, usually by impersonating a family member in distress or posing as a potential romantic partner. These attacks, which have long been successful, levelled up significantly in the deepfake era. In one of the recent high-profile romance scams, a French woman paid over a million dollars to scammers who posed as a famous actor needing help.[3]

When malware learns to code itself

From performing reconnaissance to designing attack tooling and scenarios, to executing the attacks themselves – everything can now be streamlined and automated. This greatly enhances the speed and precision of attacks, allowing cybercriminals to scale faster and cheaper than ever.

It is not uncommon for contemporary malware to bear the signs of being AI-generated. More sophisticated malware families use AI to generate harmful code on the fly and therefore stay under the radar of security solutions. Recently discovered PromptLock ransomware does not contain overtly malicious functionality in itself, but instead downloads an OpenAI GPT model and prompts it to generate scripts that perform the file encryption.[4] Another example is LameHug – an infostealer relying on Alibaba's Qwen model to generate system commands for the extraction of sensitive data.[5] In this way, malicious activity can be easily overlooked by traditional security scanners.

The underground economy

Deep in the Internet's basement, dark web marketplaces have also been transformed by AI. Shifting from ransomware-as-a-service, the underground chatter is now very much focused on deepfake creation services and methods for bypassing AI safety guardrails. Compromised AI accounts provide anonymous access to jailbroken systems capable of generating any content imaginable, essentially turning legitimate AI tools into weapons of digital deception.

1 Nick Robins-Early, "CEO of World's Biggest Ad Firm Targeted by Deepfake Scam", The Guardian, May 2024.
2 Heather Chen, Kathleen Magramo, "Finance Worker Pays Out $25 Million After Video Call with Deepfake 'Chief Financial Officer'", CNN, February 2024.
3 Laura Gozzi, "AI Brad Pitt Dupes French Woman out of €830,000", BBC, January 2025.
4 ESET Research, "ESET Discovers PromptLock, the First AI-Powered Ransomware", August 2025.
5 Vitaly Simonovich, "Cato CTRL™ Threat Research: Analyzing LAMEHUG – First Known LLM-Powered Malware with Links to APT28 (Fancy Bear)", CATO Networks, July 2025.

“The market dynamics reveal something unsettling about technological adoption curves. Criminal enterprises often embrace new technologies faster than legitimate businesses because they are not bound by regulatory compliance, ethical considerations, or customer trust concerns.”

They can move at a great speed, even with limited resources, creating a perfect storm of innovation applied to nefarious purposes.

Electoral integrity and disinformation

One of the most challenging issues that can have severe consequences going much further than financial or reputational loss is the use of deepfake technology in political campaigns. Different election cycles across countries saw an increasing number of incidents in which AI was used to create convincingly looking yet utterly untruthful propaganda. Foreign adversaries no longer need extensive infrastructure and resources to sow discord. A single compelling piece of synthetic content can create a snowball effect in which legitimate users amplify misinformation faster than fact-checkers can respond.

Fabricated images showing scenes of demographic support and deepfake audio clips that vilify prominent political figures are just a few examples of misinformation circulating on social platforms during the 2024 US presidential election.[6] When such content is shared by influencers, millions of

[6] Renee Barnes, Aimee Riedel, Lucas Whittaker, Rory Mulcahy, “Disinformation and Deepfakes Played a part in the US Election. Australia Should Expect the Same”, The Conversation, November 2024.

users might be led to believe the content is authentic. The line between truth and fiction becomes heavily blurred.

The uncomfortable truth about progress

Like with any previous ground-breaking technology, advances in AI are a double-edged sword. Every breakthrough that helps society can also be used by malicious actors to harm society. The more powerful the technology, the more potential for harm once it ends up in the wrong hands. The scary thing about AI-powered criminality is not only the extensive capabilities that AI brings in; it is also the fact that it scales infinitely without scaling linearly in resources.

> **“Sophisticated cybercrime campaigns, which once required expensive and time-consuming preparations, are fast becoming accessible to pretty much anyone - no skills required, just foul intentions.”**

More than an incremental improvement, we are looking at an unprecedented paradigm shift that requires us to think well beyond traditional defensive strategies.

Adapting old detection methods that follow familiar ways of thinking is usually the first step in tackling emerging threats. However, in the case of AI, it might not be enough. When criminals can generate millions of targeted phishing emails in no time, even the best spam filters will struggle. Deepfakes of any kind are increasingly sophisticated and, in many cases, impossible to detect; the approach in which AI-generated content is tagged with a special watermark has already proven very easy to circumvent. We can expect to be flooded with synthetic content – malicious or misleading – on a scale never experienced before, and more likely than not, our current approach to security will fail us. In a world where it is impossible to tell harmful from benign and fake from real, it might be easier to focus on certifying the authenticity of original harmless content while treating anything else as suspicious by default.

About the Author

Marta Janus is a principal researcher and founding team member at HiddenLayer, a startup focused on securing AI systems. Her work centers on investigating attacks against AI and studying the evolving AI threat landscape. Before joining HiddenLayer, Marta spent over a decade as a security researcher at leading anti-virus companies, where she developed extensive expertise in threat intelligence, malware analysis, and reverse engineering. That background has made her a prolific voice in the cybersecurity field: she's authored over three dozen publications across HiddenLayer, BlackBerry, Cylance, Securelist, and DARKReading. She's also a regular presence at industry conferences, recently delivering talks on AI security at BSidesSF, Hacktivity, CanSecWest, 44Con, OWASP Global AppSec, and Defcon's AI Village.

DANGEROUS LIAISONS

Assessing the Nexus Between Terrorism and Criminal Activities in Africa

This new publication examines the interactions between terrorist actors and criminal economies, drawing on analysis of selected groups affiliated with the Islamic State in Iraq and the Levant (ISIL/Da'esh) and Al-Qaida. The study focuses in particular on West Africa, with field research conducted in Benin, Côte d'Ivoire and Nigeria.

DOWNLOAD THE PUBLICATION