

## Public safety threat report: 2025 Artificial intelligence threat landscape

**Disclosure protocol:** GREEN: Restricted to the community

**Date of publication:** 04 November 2025

The Public Safety Threat Alliance (PSTA) threat intelligence team actively monitors and evaluates the threats to public safety. In this report, we examine the growth in artificial intelligence use in the public safety threat landscape. Our team used both open and closed sources as part of our investigation, including information from our ActiveEye Managed Detection and Response team, private and trusted vendors, and government reporting. We used numerous intelligence analytical techniques in the assessment of threat intelligence provided in this report.

### Key points

- Artificial Intelligence (AI) monetization surged in 2025 with threat actors selling jailbroken AI prompts and self-hosted LLMs through subscription based offerings for malware generation and phishing content.
- Nearly **44%** of AI misuse observed on the dark web involved phishing, contributing to phishing overtaking every other technique as the top initial access vector against public safety in 2025.
- Detection methods should adapt as threat actors weaponize AI, requiring defenders to pivot toward behavioral analytics and validating communications to defend against AI-enhanced phishing attacks.

### Executive summary

Artificial intelligence (AI) is now a foundational component of modern cyber operations, lowering the technical barrier for threat actors, while accelerating the scale of malicious activity. The PSTA has observed a **50%** increase in AI-related discussions and resources across dark web forums since early 2025, with adversaries of all skill levels integrating AI into phishing and malware development. This mirrors a broader global trend that found a **47%** rise in AI-enabled cyberattacks across all industries, indicating AI has moved from the experimental to the operational stage.

Phishing accounted for roughly **44%** of observed AI misuse on the dark web, followed by malware generating and deepfake content creation. Cybercriminals are monetizing these capabilities through subscription-based offerings, jailbroken models, and self-hosted large language models (LLMs) that operate without safeguards. This commercialization transformed AI misuse into a profitable and scalable underground industry.

TLP: GREEN

Public safety is a downstream target of these AI developments. AI-driven phishing kits and other social engineering tools have made credential theft easier, faster, and more convincing, contributing to phishing overtaking all other initial access techniques in 2025. The PSTA assesses with high confidence that AI misuse will continue to expand through 2026, requiring defenders to shift and prioritize other methods of defense.

## The rising use of malicious AI

Artificial intelligence (AI) has significantly lowered the technical barrier to become a malicious cyber actor. We observed a rise in threat actors using AI to generate phishing content and malicious code, marking a shift from experimentation to completely adopting AI. AI-related resources and discussions among threat actors of varying skill levels on the dark web increased by **50%** this year compared to 2024. This illustrates how quickly threat actors are scaling these capabilities to enhance their impact.

This trend also mirrors global reporting. Open source analysis indicates a **47%** increase in AI-enabled cyberattacks worldwide in 2025,<sup>1</sup> suggesting that the activity observed on the dark web reflects a broader transformation across the global threat landscape. The UK National Cyber Security Centre assessed that AI capabilities are being adopted across multiple tiers of cybercrime, from low-level threat actors to nation-state actors using large language models to enhance their operational impact.<sup>2</sup> This showcases how AI became a foundational component of threat actors' toolkits and will continue to be an essential capability at every level of sophistication.

Dark web marketplaces have evolved to monetize multiple forms of AI misuse. Traditional uses, such as malware generation, phishing content, and deepfakes remain common but 2025 introduced subscription-based services that resell these capabilities at scale. Threat actors are now offering "Prompt-as-a-Service" models that sell pretested jailbreak prompts and automated phishing workflows. This demonstrates that AI misuse is shifting to become a revenue generating enterprise within the cybercriminal economy, proliferating the availability and capability of AI services.

Phishing accounts for **43.8%** of observed AI misuse, followed by malware generation at **25%**, deepfake media at **25%**, and "Prompt-as-a-Service" or other emerging uses totaling **6.3%** (See Figure 1). This distribution highlights that adversaries primarily exploit AI for social engineering. This is due to AI helping cybercriminals produce believable, scalable, and adaptive phishing content that lowers the effort required to deceive victims, creating higher initial access success rates.

---

<sup>1</sup> <https://sqmagazine.co.uk/ai-cyber-attacks-statistics>

<sup>2</sup> <https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027>

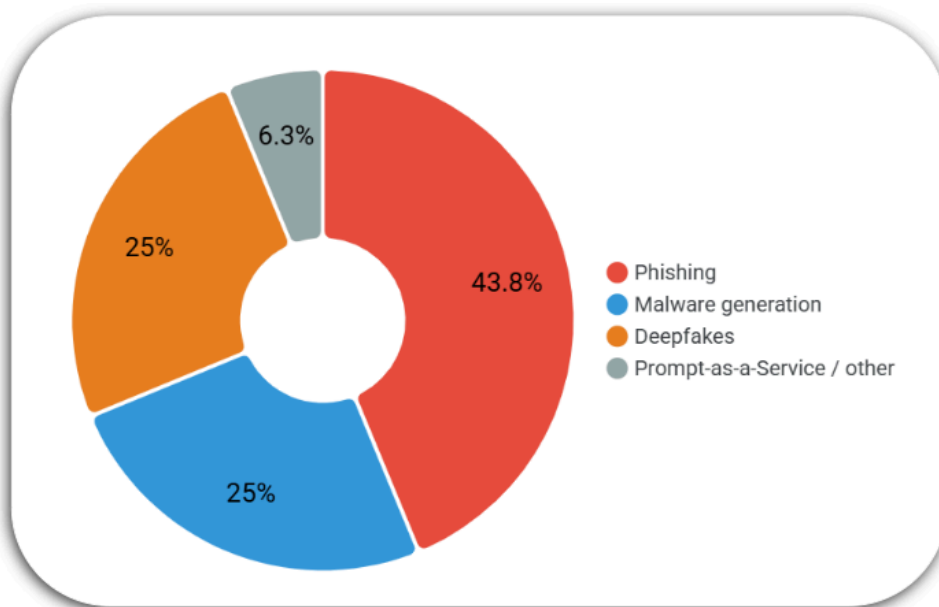


FIGURE 1: Observed AI misuse on dark web marketplaces in 2025

## Types of AI misuse

Threat actors are increasingly jailbreaking legitimate generative AI models and selling compromised prompts on the dark web. The PSTA identified multiple dark web listings advertising jailbroken prompts that bypass model safeguards to generate restricted outputs, including malware code and phishing lures. This method directly lowers the skill threshold to produce high quality malicious material without technical knowledge or access to open source LLMs. The proliferation of jailbroken prompts has accelerated the adoption of AI tradecraft, transforming generative AI model prompts into a criminal service.

Threat actors are also deploying their own private generative AI to operate without restrictions. These trained models allow adversaries to maintain full control, tailor datasets, and adjust outputs to generate malicious content. We observed an increase in threat actors advertising private environments where LLMs were retrained to create phishing messages, obfuscate malware, and analyze stolen data with no moderation limits. These self hosted models represent a growing strategic capacity within the cybercriminal ecosystem, providing threat actors with additional capabilities that they otherwise wouldn't have.

The operational impact of these developments is evident in the underground economy. Threat actors are using both private LLMs and jailbroken models with automated phishing kits, voice bots, and social engineering scripts to create stronger tradecraft capable of generating and delivering malicious content (see Figure 2). Regardless of the technique used, the ultimate goal of a threat actor using AI remains similar—generate phishing lures, generate malware, and automate attack workflows.

Techniques	Example
Self-hosted LLMs	Threat actors running their own instances for malware & phishing generation; no restrictions or safeguards
Jailbroken GPTs	Threat actors sell/use prompts that are "jailbroken" to bypass commercial AI safeguards; enables pay-to-play for script kiddies
Automation chains	Threat actors combine AI and other tools to automate their workflow, such as ransomware victim communication

FIGURE 2: Top AI techniques used by threat actors

Phishing has emerged as the most significant consequence of AI misuse for public safety, representing nearly half of AI-enabled activity observed on the dark web in 2025. This aligns with phishing becoming the primary initial access vector impacting public safety this year. This was primarily driven by the proliferation of AI-created phishing prompts which surged over **1200%** globally since the rise of generative AI use in 2022.<sup>3</sup> The proliferation of AI-enhanced phishing is the primary driver for malicious artificial intelligence impacting public safety.

## Future outlook

As AI becomes embedded in threat actor operations, traditional detection methods will struggle to keep pace with social engineering, which can bypass common defense mechanisms in public safety networks. Public safety organizations should emphasize behavioral and contextual detection to identify anomalies in emails and other communication methods that could compromise security mechanisms and networks. Continuously increasing phishing training and emphasizing situational awareness can effectively reduce human error and mitigate the risks of AI-assisted phishing.

The misuse of AI is expected to grow as LLMs and generative AI become more advanced, a trend observed in recent years. Phishing continues to be a tried-and-true method, as it remains one of the few ways to gain unauthorized access in a landscape heavily fortified by cyber defenses. Overall, AI has fundamentally reshaped the threat landscape by accelerating the attack tempo, lowering the skill threshold for cybercriminals, and amplifying deception, with these impacts now extending into the public safety sector.

<sup>3</sup> <https://www.mckinsey.com/about-us/new-at-mckinsey-blog/ai-is-the-greatest-threat-and-defense-in-cybersecurity-today>





## Appendix A: Assessment and response standard operating procedures

### Levels of analytic confidence

High confidence	Moderate confidence	Low confidence
Generally indicates judgments based on high-quality information, and/or the nature of the issue makes it possible to render a solid judgment. A "high confidence" judgment is not a fact or a certainty, however, and still carries a risk of being wrong.	Generally means credibly sourced and plausible information, but not of sufficient quality or corroboration to warrant a higher level of confidence.	Generally means questionable or implausible information was used, the information is too fragmented or poorly corroborated to make solid analytic inferences, or significant concerns or problems with sources existed.

## Appendix B: Traffic light protocol for disclosure

As part of the PSTA, agencies and other members are encouraged to share their own cybersecurity threat experiences to improve the awareness and readiness of the overall group. Submitting agencies should stipulate the level of disclosure required for their submissions according to the PSTA Traffic Light Protocol (TLP), based upon the [CISA Traffic Light Protocol guidance](#), which helps all members submit and leverage insights while being respectful of the submitting agency's preferences.

 <p><b>RED:</b> Restricted to the immediate PSTA participants only</p> <ul style="list-style-type: none"> <li><b>When should it be used?</b> Sources may use <b>TLP: RED</b> when information cannot be effectively acted upon by additional parties, and could lead to impacts on a party's privacy, reputation, or operations if misused.</li> <li><b>How may it be shared?</b> Recipients may not share <b>TLP: RED</b> information with any parties outside of the specific exchange, meeting, or conversation in which it was originally disclosed. In the context of a meeting, for example, <b>TLP: RED</b> information is limited to those present at the meeting. In most circumstances, <b>TLP: RED</b> should be exchanged verbally or in person.</li> </ul>	 <p><b>GREEN:</b> Restricted to the community</p> <ul style="list-style-type: none"> <li><b>When should it be used?</b> Sources may use <b>TLP: GREEN</b> when information is useful for the awareness of all participating organizations as well as with peers within the broader community or sector.</li> <li><b>How may it be shared?</b> Recipients may share <b>TLP: GREEN</b> information with peers and partner organizations within their sector or community, but not via publicly accessible channels. Information in this category can be circulated widely within a particular community. <b>TLP: GREEN</b> information may not be released outside of the community.</li> </ul>
 <p><b>AMBER:</b> Restricted to participants' organizations</p> <ul style="list-style-type: none"> <li><b>When should it be used?</b> Sources may use <b>TLP: AMBER</b> when information requires support to be effectively acted upon, yet carries risks to privacy, reputation, or operations if shared outside of the organizations involved.</li> <li><b>How may it be shared?</b> Recipients may only share <b>TLP: AMBER</b> information with members of their own organization, and with clients or customers who need to know the information to protect themselves or prevent further harm. <b>TLP: AMBER+STRICT</b> Restricts sharing to the organization only.</li> </ul>	 <p><b>CLEAR:</b> Disclosure is not limited</p> <ul style="list-style-type: none"> <li><b>When should it be used?</b> Sources may use <b>TLP: CLEAR</b> when information carries minimal or no foreseeable risk of misuse, in accordance with applicable rules and procedures for public release.</li> <li><b>How may it be shared?</b> Subject to standard copyright rules, <b>TLP: CLEAR</b> information may be distributed without restriction.</li> </ul>

MOTOROLA, MOTO, MOTOROLA SOLUTIONS and the Stylized M Logo are trademarks or registered trademarks of Motorola Trademark Holdings, LLC and are used under license. All other trademarks are the property of their respective owners. ©2024 Motorola Solutions, Inc. All rights reserved.

**TLP: GREEN**