Science, Technology Assessment, and Analytics



SCIENCE & TECH SPOTLIGHT:

MALICIOUS USE OF GENERATIVE AI

GAO-26-108695, December 2025

WHY THIS MATTERS

The use of generative AI is growing rapidly, and it continues to be applied in new ways across the public and private sectors. But generative AI can also be used in destructive ways. This includes producing harmful content, obtaining sensitive information, or carrying out malicious instructions. Even with safeguards in place, no current generative AI systems are immune to such misuse.

KEY TAKEAWAYS

- Attackers have many different malicious techniques that are effective against generative AI systems.
- AI safeguards and defenses require continuous development and significant resources to maintain.
- Policymakers face challenges developing timely solutions to address rapidly evolving uses and threats.

THE TECHNOLOGY

What is malicious use? Attackers or other users can cause generative artificial intelligence (AI) systems to produce harmful content, disclose sensitive information, or carry out other instructions that defy their intended purposes and built-in safeguards. This includes cybercriminals who can also modify generative AI systems to enable malicious use.

Harmful content includes information or advice to build weapons, undertake criminal activities such as cyberattacks, help users to harm themselves, or produce deepfakes or other damaging content. Content may also be inadvertently harmful when an AI system fails to discourage user interest in self-harm. In addition, AI systems can be tricked into producing intentionally biased answers or disclosing proprietary business or sensitive personal information.

When generative AI is paired with other forms of AI that can autonomously make and adjust plans for users, such as agentic AI, it could enable other AI systems to complete complex and open-ended malicious instructions. For example, paired generative AI systems could autonomously create and deliver phishing emails.

How does it work? Generative AI systems can be developed with safeguards to prevent undesired and harmful use or protected by additional software. However, the National Institute of Standards and Technology (NIST) and others have found that no AI systems, generative or otherwise, can be fully secured.

Misuse is possible because generative AI cannot easily distinguish harmless requests from malicious instructions or use. Further, attackers can use many creative techniques to make generative AI ignore safeguards or carry out malicious instructions, such as entering prompts intended to make a system ignore its safeguards against misuse and disguising malicious instructions as harmless inputs. For example:

Roleplaying. Attackers ask generative AI systems to adopt characteristics that facilitate access to harmful or sensitive content, such as the “Do Anything Now” (DAN) attack. This tricks a system into acting as if it can do anything, allowing it to ignore some safeguards. While systems have been programmed to stop DAN attacks, roleplaying methods can still be successful particularly when combined with other techniques.

Crescendo. Attackers steer generative AI systems to create harmful content using small, seemingly benign steps. By shifting focus gradually, the attacker exploits the system’s tendency to comply and avoids direct requests that could trigger safeguards.

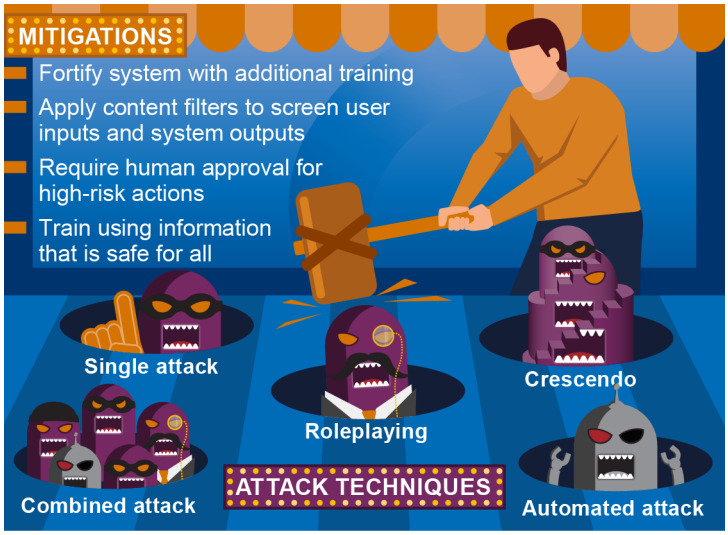
Automated attacks. These methods use two or more generative AI systems that together refine the phrasing of an

initial harmful request to find prompts that bypass safeguards. These attacks can iteratively modify and judge the effectiveness of new prompt phrasings until they are successful.

How mature is it? Many techniques for malicious use of generative AI already exist, and tactics are rapidly evolving. For example, researchers recently claimed to have successfully circumvented the safeguards of a new system a single day after its release, thereby obtaining instructions to build an incendiary weapon. Additionally as a readily available technology, one academic study reported that generative AI could reduce malicious users' costs of conducting phishing cyberattacks by more than 95 percent.

How can malicious efforts be countered? Security efforts to test generative AI vulnerabilities are well established in industry and defense techniques are rapidly evolving. NIST has identified mitigation techniques such as reinforcing safeguards through human feedback, filtering of user instructions, and using separate generative AI technology to detect malicious attacks.

Figure 1. Selected Attack Techniques and Mitigations for Malicious Use of Generative AI



Source: GAO analysis (data); GAO illustration (blob features/graphic elements/foam finger/mallet/robot); Kavya/stock.adobe.com (blobs); Wei/stock.adobe.com (person). | GAO-26-108695

CHALLENGES

- Attackers and other users such as cybersecurity researchers are continuing to find new methods to manipulate generative AI. Developers will need to continuously monitor and address vulnerabilities to prevent additional misuse or harmful content.
- Incentives for generative AI developers are primarily focused on improving performance, rather than addressing security issues such as malicious use.

POLICY CONTEXT AND QUESTIONS

- How could the development of AI benchmarks or auditing tools promote research in security and help address vulnerabilities?
- What guidelines can best ensure generative AI systems are used responsibly, and to what extent do generative AI systems follow existing guidance?
- As generative AI is integrated into IT systems supporting federal agencies, businesses, and others, what actions are needed to mitigate risks posed by malicious use?

SELECTED GAO WORK

Artificial Intelligence: Generative AI Training, Development, and Deployment Considerations, GAO-25-107651.

Science and Tech Spotlight: AI Agents, GAO-25-108519.

SELECTED REFERENCE

Vassilev A, Oprea A, Fordyce A, Anderson H, Davies X, and Hamin M. (2025) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Artificial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2025. https://doi.org/10.6028/NIST.AI.100-2e2025.

GAO SUPPORT:

The Government Accountability Office (GAO) meets congressional information needs in several ways, including by providing oversight, insight, and foresight on science and technology issues. GAO staff are available to brief on completed bodies of work or specific reports and answer follow-up questions. GAO also provides targeted assistance on specific science and technology topics to support congressional oversight activities and provide advice on legislative proposals.

For more information, contact: Karen L. Howard, PhD, HowardK@gao.gov

Public Affairs: Sarah Kaczmarek, Managing Director, Media@gao.gov

Congressional Relations: A. Nicole Clowers, Managing Dir., CongRel@gao.gov

This document is not an audit product and is subject to revision based on continued advances in science and technology. It contains information prepared by GAO to provide technical insight to legislative bodies or other external organizations. This document has been reviewed by Sterling Thomas, PhD, the Chief Scientist of the U.S. Government Accountability Office.

This work of the United States may include copyrighted material, details at https://www.gao.gov/copyright.

Staff Acknowledgments: Katrina Pekar-Carpenter (Assistant Director), Nathan Hanks (Analyst-in-Charge), Eli Duggan, Nathan Hamm, Rachael Johnson, Mark Kuykendall, and Joe Rando.

Source (header photo): Maksim Kabakou/stock.adobe.com (GAO adaptation). | GAO-26-108695