



| Assessment |

How GenAI is Changing the Cyber Threat Landscape

A-2024-03-06a

Classification: TLP:CLEAR

Criticality: Low

Intelligence Requirements: Artificial Intelligence, Deep & Dark Web, Threat Actor, Geopolitical

March 6, 2024

Scope Note

ZeroFox Intelligence is derived from a variety of sources, including—but not limited to—curated open-source accesses, vetted social media, proprietary data sources, and direct access to threat actors and groups through covert communication channels. Information relied upon to complete any report cannot always be independently verified. As such, ZeroFox applies rigorous analytic standards and tradecraft in accordance with best practices and includes caveat language and source citations to clearly identify the veracity of our Intelligence reporting and substantiate our assessments and recommendations. All sources used in this particular Intelligence product were **identified prior to 9:00 AM (EST) on March 1, 2024**; per cyber hygiene best practices, caution is advised when clicking on any third-party links.

Assessment | How GenAI is Changing the Cyber Threat Landscape

Executive Summary

Generative Artificial Intelligence (GenAI) models offer vast potential in enabling mass automation, economic growth, scientific advances, and communication. However, there are also numerous security risks arising from its development and deployment, including ethical concerns, societal impacts, and use by nefarious cyber actors for malicious purposes. Adoption of GenAI-enabled technologies by corporate enterprises will also very likely introduce new vulnerabilities into organizations' infrastructure and undermine secure-by-design principles.

With emerging technologies and tools getting widespread and significant media and public attention, much of the discussion around the threat from GenAI is misleading, unclear, and unhelpful.

At least in the short to medium term, GenAI will very likely serve as a force for both good and evil; the rapid development of new technologies will be added to the repository of tools used by threat actors to conduct their attacks as well as to the arsenals security personnel use to mitigate threats.

What is GenAI?

Much of cyber security-related data is categorical in nature, meaning it can be divided into groups or categories. Traditionally, discriminative machine learning models have been used for categorizing this data, which are trained by learning the optimal decision boundaries that separate different classes within the input data. When given supervised machine learning tasks, these models excel at drawing clear distinctions between classes, including image classification and sentiment analysis.

Instead, generative machine learning models are trained to learn the underlying patterns (or distributions) of the input data to generate new, similar data instances. By learning these patterns, GenAI models can be leveraged to create synthetic content—including image, voice, and video generation, as well as data augmentation. In the case of Large Language Models (LLMs), by modeling the probabilistic structures of large language datasets, the model can predict the most likely next word (or sentence), facilitating text creation.

The development and user uptake of GenAI-based technologies increased significantly in 2023, with tools like ChatGPT garnering mass-media attention and public use.

Cybersecurity vendors have also sought to capitalize on new capabilities, swiftly incorporating GenAI-based features into their offerings to enable better detection and analysis of cyber threats. Cybercriminals have also leveraged GenAI-based tooling to increase the scope, scale, and efficiency of their attacks.

Threat Overview

In the short term, attacks leveraging GenAI will almost certainly increase in volume and lower the barriers to entry for threat actors, increasing the threat and potential impact of attacks. The greatest impact on the cyber threat landscape will most likely come from

threat actors leveraging GenAI-based tools to enhance the efficiency, efficacy, and scale of tactics, techniques, and procedures (TTPs) used in existing attacks.

- GenAI-enabled tooling will very likely be of use—to varying degrees—to most threat actors.
- Cyber threat actors of all capabilities, from less sophisticated hacktivists to nation-state actors, are very likely leveraging GenAI tools to enhance their attacks.

AI-enabled tools will very likely be leveraged by threat actors to make their existing operations more efficient and effective. AI's ability to process and analyze exfiltrated data at pace and with greater accuracy is very likely to make cyberattacks more impactful.

- By analyzing, categorizing, and summarizing data at pace, threat actors will likely be able to identify information of high value for exfiltration and, potentially, extortion. This will likely make attacks more effective in the short term. It will, however, almost certainly increase costs for threat actors to run such code.

However, threat actors' technical capabilities and resources will almost certainly serve as a limiting factor to their ability to harness the full potential of GenAI-enabled tools in their attacks. Those that are able to gain access to vast and high-quality training data, expertise, and resources stand to benefit most from the proliferation of GenAI-enabled technologies and will likely drive the most significant changes to the overall cyber threat.¹

- The most sophisticated cyber threat actors—such as nation-states and advanced persistent threats (APTs), as well as mature cybercriminals—will almost certainly drive the greatest developments in TTPs by leveraging GenAI-based tooling, given their access to the resources, expertise, and data required to benefit from the potential uplift in capability.
- Lower-capability cybercriminal and hacktivist actors will likely be able to leverage GenAI-based tools to lower barriers to entry for conducting attacks and raise the baseline efficacy of attacks. However, this will likely have an evolutionary rather than revolutionary impact on the threat from these actors.

¹ Rapid developments in GenAI are driving an increase in pre-trained Gen AI that does not need large volumes of data to get effective results.

ZeroFox anticipates that GenAI and machine learning-enabled tooling will become increasingly democratized, with threat actors and security personnel rapidly adopting and developing these services to achieve their offensive and defensive aims.

GenAI-enabled tools are evolving at a rapid pace, making detection particularly difficult. Once the purview of data scientists and mature engineering teams, these models will increasingly be developed and deployed by nefarious individuals with wider skill sets, with increased data feeding these models to drive faster and more effective cyberattacks.

Use of LLMs for Social Engineering

One of the greatest threats to organizations globally from threat actors leveraging GenAI is through social engineering. These models enhance the efficacy of campaigns by making them more effective, efficient, and difficult to detect, as well as opening up attack vectors to those that would otherwise lack the skill sets or language competencies to participate effectively. ZeroFox has observed threat actors leveraging LLMs to conduct phishing campaigns, enabling non-native speakers to capture the complexity and nuance of natural language in attacks.

LLMs have very likely raised the baseline for convincing, low-sophistication mass phishing attacks. This will highly likely increase over the next two years as models evolve and uptake increases. LLMs and GenAI more broadly will very likely make it difficult for everyone, regardless of their level of cybersecurity understanding, to assess whether an email or password reset request is genuine.

- Generic mass phishing attacks have traditionally been easily identifiable, with translation, spelling, and grammatical mistakes that often reveal inauthentic and malicious intent behind messages.
- LLMs enable threat actors to better capture natural, native language, very likely raising the minimum standard of many “spray and pray” phishing campaigns.
- However, LLM-created content can still appear unusual to readers, with atypical words or phrases generated.

Many of the typical red flags associated with phishing campaigns will remain, such as lures to encourage victims to engage with malicious links and attachments and a

time-sensitive element or urgency in taking action.

LLMs have very likely made it easier for threat actors to conduct spear phishing campaigns, lowering the barriers to targeting specific users with crafted lures under the guise of being a known contact. Typically, spear phishing requires deep knowledge of the target and an added layer of reconnaissance. As LLMs develop, automating high-quality phishing at scale will become easier, providing criminals with a powerful new tool for automating identity theft and fraud.

- LLMs only need a few email or social media samples to learn to impersonate a target's writing, which makes it easy to quickly create customized, believable phishing emails.
- Threat actors can feed inputs to LLMs—such as scraping open source repositories or breached information—to personalize email content to individuals based on the target's personal and political interests, sensitive family information, or other personal details.
- Threat actors can leverage basic prompt engineering to circumvent safeguards installed in LLMs that restrict the creation of content useful for malicious purposes.

Use of Deepfakes for Malicious Purposes

While deepfake technology can be designed and used for benign purposes such as art, marketing, and advertising, threat actors are increasingly manipulating the facial appearance of subjects in images or videos using generative models for malicious purposes. Synthetically-generated videos can now exhibit natural expressions, speech, and mannerisms that can be difficult to distinguish from reality.

The threat from deepfakes likely remains on an upward trajectory as threat actors continue to experiment with deepfake tools for a wide range of purposes, including—but not limited to—financial fraud, extortion and harassment, authentication bypass, and disinformation. However, evidence of financially-motivated threat actor campaigns successfully leveraging sophisticated deepfake content is currently limited, and the frequency of publicly-reported attacks remains low.

Cybercriminals on the deep and dark web (DDW) have begun commoditizing GenAI-enabled tooling in as-a-service offerings, providing buyers of all capabilities access to innovative services. Multiple threat actors, over a sustained period, have a history of offering deepfake services for sale in DDW communities.

- Positive reputation threat actors have been observed advertising deepfake services on DDW forums such as lip sync services, face swaps, and voiceovers.
- The price for such services is typically low, with some offered for as little as USD 150 for a minute of deepfake content.
- Other services include developing landing pages for retailers, portals, gambling sites, and graphic design, as well as video and audio manipulation.

The threat from deepfakes to most organizations is highly nuanced and widely oversimplified, misdirected, and sensationalized. Security teams must evaluate the threat from deepfakes to their organizations on a case-by-case basis and consider the prioritization of resources to mitigate them as well as other types of direct threats that currently pose a considerably higher risk.

- Despite garnering significant media attention, the direct threat to most organizations from deepfakes is currently assessed to be low and limited to several specific cases outlined below.
- The quality of deepfakes is improving, but most cases remain easily identifiable as fake by most viewers.

Sextortion and Harassment

While currently rare, organizations should be wary of sextortion campaigns leveraging deepfaked explicit material of employees for financial gain or to damage the organization's reputation. ZeroFox has observed an increase in threat actors collecting images and videos from victims' social media accounts and video chats, as well as requesting media directly from victims to create sexually explicit but fake content. Threat actors then use this content to blackmail victims—either to extort financial payment or sexually-themed images and videos from the victim—threatening to share the manipulated explicit media with family members and friends in the event of non-compliance.

- In January 2024, explicit deepfake images of American singer-songwriter Taylor Swift spread across the social media platform X (formerly Twitter), highlighting the damaging impact and prevalence of non-consensual, deepfake pornography.²
- Reporting indicates increased use of child and teen material, suggesting that some threat actors have very likely changed their tactics in an attempt to elicit payments with greater success.

Fraud

Successful deepfake scams for fraudulent purposes remain rare but are very likely on an upward trajectory, with the potential to significantly impact a wide number of organizations globally. The increased threat stems mainly from deepfake video and voice-cloning technology leveraged by threat actors to socially engineer victims into transferring money or sensitive information.

- In February 2024, threat actors stole HK\$ 200 million (approximately USD 26 million) from a multinational company by tricking a Hong Kong-based employee into transferring the funds.³ The scammers reportedly used publicly-available footage of the company's Chief Financial Officer to conduct a deepfake video conference with the targeted employee.
- In January 2024, deepfake videos of Anatoly Yakovenk, co-founder of the Solana blockchain platform, were leveraged as part of scams attempting to capitalize on a legitimate Solana marketing campaign.⁴ Deepfake-related scams have become increasingly regular within the cryptocurrency market, as threat actors seek to capitalize on typically weak security awareness among blockchain users and the highly speculative nature of cryptocurrencies.

Authentication Bypass

ZeroFox has identified multiple instances of DDW actors seeking or offering deepfake services to bypass authentication security measures, most likely for financial gain. Threat actors in these forums have also sought credible deepfake verification bypass services and clues for exploiting such techniques.

² [hXXps://www.wired\[.\]co\[.\]uk/article/deepfake-porn-is-out-of-control](https://www.wired[.]co[.]uk/article/deepfake-porn-is-out-of-control)

³ [hXXps://www.theregister\[.\]com/2024/02/05/hong_kong_deepfaked_cfo/](https://www.theregister[.]com/2024/02/05/hong_kong_deepfaked_cfo/)

⁴ [hXXps://www.theverge\[.\]com/2024/1/3/24024262/youtube-twitter-x-crypto-solana-deepfake-scam](https://www.theverge[.]com/2024/1/3/24024262/youtube-twitter-x-crypto-solana-deepfake-scam)

- On October 3, 2023, actor “akabembi” requested guidance regarding the exploitation of deepfake and AI for bypassing verification security measures. The actor was interested in swapping passport photographs and gave an example of a video of a person speaking in Russian and software that performs lip sync and simultaneous translation.
- Responses included threat actors sharing toolkits and useful sources of information, explicitly mentioning DeepFaceLive and Obs Virtual Camera to emulate deep fake models.⁵ Other tools included Manycam and Splitcam (virtual cameras), Cloner Pro APK, Roop Cam, and Avatarify, as well as a “Deepfake Offensive Toolkit” which allegedly makes real-time, controllable deep fakes ready for visual camera injection.
- One actor also shared a Chinese-speaking forum dedicated to deepfakes named deepfakelab, which focuses on deepfake models, deepfake tutorials, materials, and AI speech-modeling tutorials.

Mis- and Disinformation

The primary threat to most organizations from mis- and disinformation resulting from deepfakes and other synthetically-generated content is reputational damage. This content can serve as a powerful disinformation tool, with the synthetic content circulated on social media typically involving high-profile individuals and brands in fabricated scenarios due to the ease of acquiring readily-available data on the subject. Tools such as Midjourney, Dall-E, and Sora enable users to generate images and videos purely via a user-provided text prompt. These can:

- Create a false narrative that impacts the entity’s reputation or livelihood.
- Be shared for propaganda purposes, representing politically-charged ideas through manipulated content that may contain false information to intentionally mislead the public.
- Be shared for satirical purposes.

While deepfakes typically remain largely satirical and low in overall sophistication, synthetically-generated images, audio, and videos have exacerbated the difficulty in identifying trustworthy information and discerning real from fake content—particularly on social media platforms. GenAI has supercharged this issue, as prior upper limits of

⁵ [hXXps://obsproject\[.\]com/forum/resources/obs-virtualcam.1744/](https://obsproject[.]com/forum/resources/obs-virtualcam.1744/)

simple transposing or retouching images have disappeared; generated images can involve radical transformations of original images and even become entirely dissociated from them.

- ZeroFox identified multiple AI-generated images circulating on social media following the outbreak of the Israel-Hamas war. While some instances of an image or video are flagged on the platform as being synthetically generated (such as via community notes), these images are widely circulated and reposted, with other iterations of the same synthetic content having no flag.

One of the broadest and most complex examples of GenAI-based mis- and disinformation exists within the context of elections. Each year, synthetic content becomes increasingly prevalent in the lead-up to and during election cycles, raising concerns over the ability of GenAI tools to influence public discourse and interfere with the electoral process. While such a scenario remains unlikely—primarily due to the lack of sophisticated content that could cause significant and widespread influence—the threat is likely to continue growing in line with the advancement of GenAI tools and their output. Notably, synthetic content does not need to be malicious at its source to influence a voter's opinion; an attempt to parody a political candidate can quickly be manipulated or used out of context and shared widely to spread false narratives.

- Deepfakes and cheapfakes (the process of manipulating media without leveraging AI) were shared widely leading up to Pakistan's February 2024 elections and Turkey's May 2023 elections, aiming to sway voter opinion. ZeroFox has observed similar tactics being leveraged in the run-up to the U.S. presidential elections in November 2024.
- In January 2024, deepfake audio calls designed to impersonate President Joe Biden were reported to be targeting New Hampshire-based residents, advising them against voting in January's presidential primary and to save their vote for the November general election.⁶ Following a law enforcement investigation, the calls were discovered to originate from Texas-based telecommunications companies, very likely in an attempt to misinform voters and interfere with the election process.

⁶ [hXXps://edition.cnn\[.\]com/2024/01/22/politics/fake-joe-biden-robocall/index.html](https://edition.cnn.com/2024/01/22/politics/fake-joe-biden-robocall/index.html)

- Before Slovakia's general election in September 2023, a deepfake audio clip involving politicians allegedly discussing how to commit electoral fraud spread across social media platforms, casting doubt among voters.

AI for Malware Development

The greatest threat from AI-enabled tools facilitating malware development stems from highly-capable and resourced threat actors creating supervised machine learning models to train malware on quality exploit data that enables it to evade detection. In the short to medium term, this will very likely remain the purview of highly-resourced nation-state operatives with access to considerable malware repositories and personnel with sophisticated cyber and data science expertise.

- GenAI-enabled tools can piece together detection evasion capabilities of existing samples, rather than producing something inherently unique.
- The majority of cyber threat actors almost certainly lack a repository of malware samples large enough to train models to produce malware capable of circumventing security protocols.
- In parallel, security vendors will very likely be training machine learning models to detect and prevent the execution of such content. As such, malware generated this way would need to be continually maintained to consistently evade evolving detection measures.

The threat from malware produced by commoditized LLMs is low. ZeroFox has identified no evidence to suggest that code produced by LLMs is more effective than code produced by capable human actors. While LLMs can be leveraged to produce malware, the code is typically simple, low complexity, and likely to be detected by the majority of comprehensive detection solutions. For a threat actor to leverage an LLM to produce malware of an effective standard, they would very likely need to already possess a sufficient level of technical acumen that they could write the code themselves.

- While many LLMs have restrictions that prohibit malicious content creation, carefully-selected prompt engineering can allow users to circumvent these restrictions.
- Some "jailbroken" LLMs such as WormGPT and FraudGPT can produce malware,

but the code remains likely to be of a low standard and easily detectable.

- While LLMs are very unlikely to be capable of producing effective malware on their own, capable threat actors may be able to leverage LLMs to enable them to accelerate the pace at which they can create malware.

| New Vectors to Attack Corporate Infrastructure

Currently, many organizations are rapidly adopting AI-enabled technologies into their infrastructure to improve customer experience—and likely unwittingly introducing new vulnerabilities into their environments and undermining secure-by-design principles.

LLM-based chatbots incorporated into enterprise offerings pose a significant risk for organizations. These tools essentially implement a “black box” into organizations’ infrastructure: a tool for which the host organization has no knowledge of its inner workings. As chatbots and other LLM-based interfaces are more widely adopted, it is not clear how organizations would be able to implement effective vulnerability mitigation procedures if and when vulnerabilities are identified in these tools.

Prompt injection vulnerabilities are of increasing concern, enabling threat actors to use crafted prompts to bypass filters or manipulate the LLM output to perform unauthorized actions. Prompt injection attacks can lead to the leaking of sensitive data and other security breaches, cause an LLM to give out unintended content that can be of reputational risk to companies, and grant threat actors unauthorized access.

- This vulnerability is exacerbated by LLMs being increasingly equipped with plug-ins, enabling the tool to call on up-to-date information and external services via APIs.
- GenAI models trained on datasets, including confidential information, could inadvertently result in leaks of sensitive and proprietary information such as source code.

Recommendations

Prevent

- Incorporate synthetic media education into existing cybersecurity training, including examples designed to increase workforce awareness.
- Review compliance procedures for financial transactions, providing greater latitude to challenge senior leadership requests.
- Document and track executive exposure in open and closed sources, and reduce digital footprints to minimize the availability of media that fuels impersonation.
- Consume ZeroFox AI and Synthetic Media Intelligence for ongoing awareness and recommendations for defending against synthetic media-enabled threats.
- Conduct a thorough risk assessment of GenAI-enabled technologies adopted into corporate infrastructure, including potential vulnerabilities that may result.

Detect

- Leverage deepfake detection technologies, such as Sensity, DuckDuckGoose, Reality Defender, deepware, or tools from Microsoft and Intel.
- Monitor corporate social media and websites for signs of manipulation.
- Inspect images used by third-party profiles for distortions, indistinct and blurry backgrounds, and other visual artifacts often found in synthetic images.

Respond

- Create a crisis response plan to neutralize and contain any incidents of deepfake impersonation or misinformation and disinformation targeting the organization.

Protect

- Apply watermarks and/or use blockchain technology to verify media content.
- Explore digital provenance solutions through the Coalition for Content Provenance and Authenticity (CP2A) or the Content Authenticity Initiative (CAI).

Appendix A: Traffic Light Protocol for Information Dissemination

	Red	Amber
WHEN SHOULD IT BE USED?	Sources may use TLP:RED 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.	Sources may use TLP:AMBER when information requires support to be effectively acted upon but carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
HOW MAY IT BE SHARED?	Recipients may NOT share TLP:RED with any parties outside of the specific exchange, meeting, or conversation in which it is originally disclosed.	Recipients may ONLY share TLP:AMBER information with members of their own organization and its clients, but only on a need-to-know basis to protect their organization and its clients and prevent further harm. Note that TLP:AMBER+STRICT restricts sharing to the organization only.
	Green	Clear
WHEN SHOULD IT BE USED?	Sources may use TLP:GREEN when information is useful for the awareness of all participating organizations, as well as with peers within the broader community or sector.	Sources may use TLP:CLEAR when information carries minimal or no risk of misuse in accordance with applicable rules and procedures for public release.
HOW MAY IT BE SHARED?	Recipients may share TLP:GREEN information with peers and partner organizations within their sector or community but not via publicly accessible channels.	Recipients may share TLP:CLEAR information without restriction, subject to copyright controls.

Appendix B: ZeroFox Intelligence Probability Scale

All ZeroFox intelligence products leverage probabilistic assessment language in analytic judgments. Qualitative statements used in these judgments refer to associated probability ranges, which state the likelihood of occurrence of an event or development. Ranges are used to avoid a false impression of accuracy. This scale is a standard that aligns with how readers should interpret such terms.

Almost No Chance	Very Unlikely	Unlikely	Roughly Even Chance	Likely	Very Likely	Almost Certain
1-5%	5-20%	20-45%	45-55%	55-80%	80-95%	95-99%