ZEROFOX Intelligence

Brief

Q2 2023: Synthetic Media Update

B-2023-06-16a

Classification: TLP:CLEAR

Criticality: MEDIUM

Intelligence Requirements: Synthetic Media, Deepfakes

June 16, 2023

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 12:00 PM (EDT) on June 15, 2023; per cyber hygiene best practices, caution is advised when clicking on any third-party links.

| Brief | Q2 2023: Synthetic Media Update

| Key Findings

- The widespread impact of synthetic media-driven campaigns almost certainly remained limited within the cyber threat landscape in Q2 2023. Nonetheless, threat actors continued to look for ways to leverage synthetic media for fraud, impersonation, extortion and harassment, and the spread of disinformation.

- Accessible and easy-to-use open source (OS) tools continued to proliferate, further contributing to the low barriers to entry for creating synthetic content. However, ZeroFox Intelligence observed no new capabilities or tools that significantly changed the overall threat to individuals and organizations.

- Threat actors capitalized on the widespread interest in synthetic media tools to commit fraud and deliver malware by using the topic as a lure.

- The threat from synthetic media will likely remain broadly consistent over the next quarter; however, ZeroFox Intelligence anticipates that synthetic media-driven disinformation campaigns will increase towards the end of 2023 in the lead-up to the 2024 U.S. presidential election.

Overview

ZeroFox Intelligence observed no significant change in the threat from synthetic media in Q2 2023. Threat actors continued to seek ways to leverage synthetic media for a wide range of purposes, including—but not limited to—financial fraud, extortion and harassment, disinformation, and impersonation. However, the widespread impact of synthetic media-driven campaigns almost certainly remained limited within the context of the cyber threat landscape. The number of publicly reported attacks observed to have successfully leveraged synthetic content was consistent with the previous quarter and occurred sporadically, with the majority of threat activity remaining largely rudimentary and experimental in nature. Notably, ZeroFox Intelligence assesses that many security incidents relating to synthetic media very likely continue to go unnoticed or underreported due to the general novelty of the threat.

The sustained hype surrounding rapidly developing, state-of-the-art artificial intelligence (AI) technologies—particularly generative language tools, such as ChatGPT and Google Bard—ensured that synthetic media-related capabilities remained high-profile throughout Q2 2023. Accessible and easy-to-use OS tools continued to proliferate, further contributing to the low barriers to entry for creating synthetic content while also providing threat actors with a greater variety for manipulating media to their desired effect. However, ZeroFox Intelligence observed no new capabilities or tools that significantly changed the overall threat to individuals and organizations in Q2 2023; the majority of new tools observed were largely created to enhance design and productivity.

Deepfake Threats

The threat from deepfake tools in Q2 2023 almost certainly remained consistent with the previous quarter, with no significant change in capabilities observed. Broadly, deepfakes continued to be shared across social media platforms on a large scale, though most content remained largely satirical and low in overall sophistication and typically involved public figures. However, campaigns observed during the quarter demonstrated the complex and multifaceted threat posed by deepfake audio and video content when leveraged with malicious intent.

> Sextortion Campaigns

On June 5, 2023, the Federal Bureau of Investigations (FBI) issued an advisory that warned of an uptick in deepfakes used by malicious actors during sextortion schemes since at least April 2023.[1] Threat actors were observed collecting images and videos from victims' social media accounts and video chats, as well as requesting media directly from some victims to create sexually explicit yet fake content. Similar to traditional sextortion campaigns, threat actors demanded either 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.

Notably, the use of sexually explicit deepfakes by threat actors for extortion and harassment is not novel. However, the reported uptick in activity also included 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 remain rare, with only a small number of similar reported attacks occurring in recent years. In May 2023, deepfake video technology was used in a China-based scam to target an individual within the technology sector, resulting in a loss of approximately USD 620,00.[2] In this case, the victim was very likely targeted by the threat actor due to their perceived access and ability to transfer large sums, with social engineering almost certainly conducted to identify and collect data from specific individuals that the victim would recognize and trust. The victim willingly transferred the sum after receiving a video call from a threat actor via WeChat—a Chinese messaging, social media, and mobile payment app—who leveraged face-swapping techniques to pose as a friend of the victim and requested financial assistance for a work-related project. In an attempt to legitimize the request, the threat actor sent a screenshot of a fraudulent payment that had been transferred to a third party, making it seem as though they had made a genuine business transaction.

> Elections Influence

Turkey's recent elections highlight the growing threat posed by synthetic media within

[1] hXXps://www.ic3[.]gov/Media/Y2023/PSA230605
[2] hXXps://www.reuters[.]com/technology/deepfake-scam-china-fans-worries-over-ai-driven-fraud-2023-05-22/

a sociopolitical context and how it can be leveraged to skew the information landscape. Deepfakes and cheapfakes (the process of manipulating media without leveraging AI) played a central role in Turkey's presidential elections in May 2023, which was subject to high levels of misinformation and disinformation throughout.[3] Most notably, manipulated video footage was used directly by Turkey's president, Recep Tayyip Erdoğan, in an attempt to influence public sentiment against his main opposition candidate, Kemal Kılıçdaroğlu. At a campaign rally on May 7, 2023, Erdoğan shared with his supporters a Kılıçdaroğlu campaign video that had been edited to depict members of the Kurdistan Workers' Party—widely designated as a terrorist organization—showing direct support to Kılıçdaroğlu. The video was later debunked and proven to be a cheapfake that combined two separate videos with different backgrounds and content, with Kılıçdaroğlu later claiming that Russia was responsible for providing the altered footage in Erdoğan's favor.[4] [5] Separately, one of the best-known opposition candidates, Muharrem İnce, withdrew from the elections following the circulation of an alleged sex tape on social media, which İnce claimed was a deepfake.[6]

While it is unclear to what extent the involvement of synthetic media impacted the overall outcome of the election, it is evident that synthetic media has the potential to have a direct influence on the electoral process. In addition, the open use of synthetic media at a state level to influence public opinion during the elections suggests that countries with traditionally greater levels of censorship or corruption could be more susceptible to synthetic media-driven disinformation in the future; Turkey ranks 165 out of 180 countries on the Reporters Without Borders world press freedom index, which measures the degree of freedom available to journalists and media outlets globally.[7]

---

[3] hXXps://www.euronews[.]com/next/2023/05/12/ai-content-deepfakes-meddling-in-turkey-elections-experts-warn-its-just-the-beginning

[4] hXXps://www.dw[.]com/en/fact-check-turkeys-erdogan-shows-false-kilicdaroglu-video/a-65554034

[5] hXXps://www.reuters[.]com/world/middle-east/erdogan-rival-accuses-russia-deep-fake-campaign-ahead-presidential-vote-2023-05-12/

[6] hXXps://www.theguardian[.]com/world/2023/may/11/muharrem-ince-turkish-presidential-candidate-withdraws-alleged-sex-tape

[7] hXXps://rsf[.]org/en/country-t%C3%BCrkiye

Synthetic Media Tools Leveraged as Topical Lures

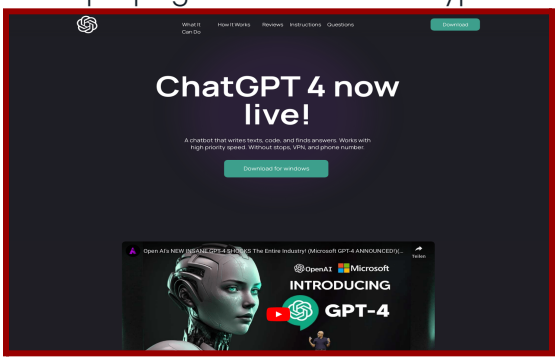
Threat actors continued to capitalize on the widespread interest in synthetic media tools to commit fraud and deliver malware during Q2 2023. ZeroFox Intelligence observed an increase in campaigns leveraging popular tools as lures within social engineering campaigns, including ChatGPT and Midjourney—a generative text-to-image model. This trend is very likely to persist for the foreseeable future; threat actors invariably seek to exploit trending topics to entice and trick victims, and AI-related developments will almost certainly remain high-profile over the coming quarters.

> Increase in ChatGPT-Related Domains

Monthly ChatGPT-related domain registrations have reportedly increased by 910 percent from November 2022 to early April 2023, with one in every 25 domains identified as malicious.[8] [9] In May 2023, Meta took down more than 1,000 malicious URLs shared across its platforms that leveraged ChatGPT as a lure to propagate around 10 malware strains since March 2023.[10]

> Malware Delivered via Malicious AI Google Search Ads

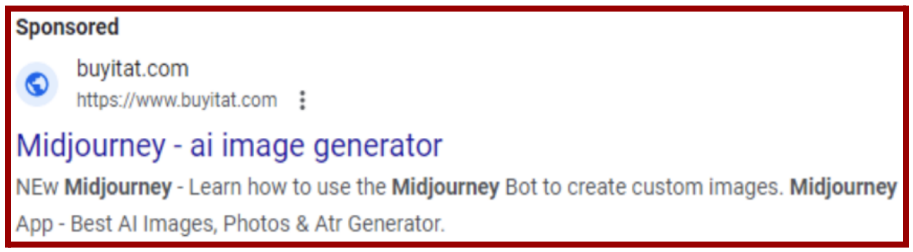
In May 2023, RedLine Stealer was delivered via a BatLoader campaign that leveraged malicious Google Search Ads posing as legitimate ChatGPT and Midjourney applications.[11] BatLoader is a dropper malware that is known for employing malvertising and social engineering tactics to propagate other various types of malware.



[8] hXXps://www.infosecurity-magazine[.]com/news/chatgpt-related-malicious-urls-rise/
[9] hXXps://www.helpnetsecurity[.]com/2023/05/04/malicious-chatgpt/
[10] hXXps://www.reuters[.]com/technology/meta-says-chatgpt-related-malware-is-rise-2023-05-03/
[11] hXXps://www.esentire[.]com/blog/batloader-impersonates-midjourney-chatgpt-in-drive-by-cyberattacks

Malicious ChatGPT Landing Site

Source:
hXXps://www.esentire[.]com/blog/batloader-impersonates-midjourney-chatgpt-in-drive-by-cyberattacks

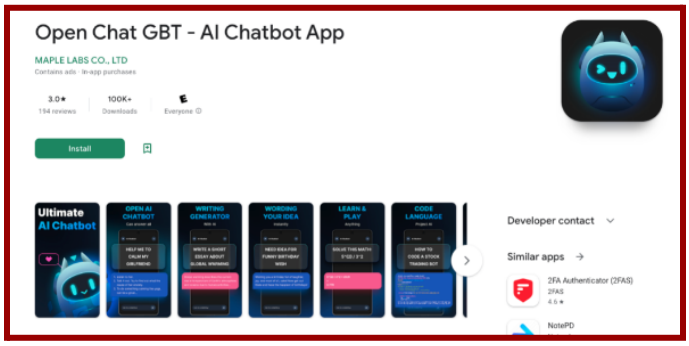


Malicious Midjourney Google Search Ad

Source:
hXXps://www.trendmicro[.]com/en_gb/research/23/e/malicious-ai-tool-ads-used-to-deliver-redline-stealer.html

› Fleeceware Apps Mimic ChatGPT

Fleeceware apps designed to resemble ChatGPT on both Google's Play and Apple's App Store continued to proliferate in Q2 2023. Fleeceware apps often contain hidden fees designed to overcharge users while providing limited functionality.[12]



"Open Chat GBT" app on Google Play Store

Source:
hXXps://news.sophos[.]com/en-us/2023/05/17/fleecegpt-mobile-apps-target-ai-curious-to-rake-in-cash/

Deep and Dark Web Findings

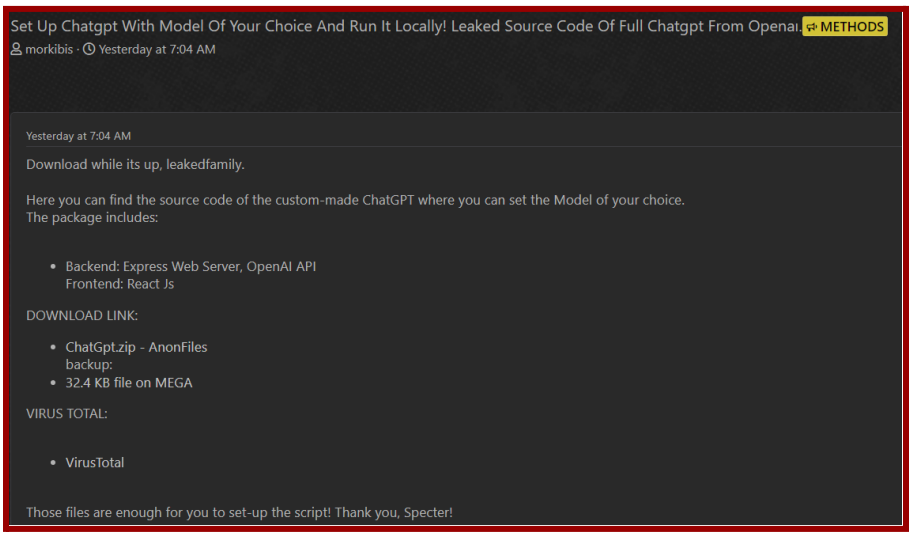
ZeroFox Intelligence continued to observe synthetic media-related chatter in closed forums and Deep and Dark Web (DDW) marketplaces during Q2 2023. DDW actors remained motivated to leverage synthetic media tools to bypass security measures and

[12] hXXps://news.sophos[.]com/en-us/2023/05/17/fleecegpt-mobile-apps-target-ai-curious-to-rake-in-cash/

enhance existing operations while continuing to experiment with and exchange expertise on various tools, including deepfakes and ChatGPT. However, ZeroFox Intelligence did not observe any references to specialized tools beyond those available on OS.

> Alleged ChatGPT Source Code Leak Disclosed on DDW Forum

On June 5, 2023, well-regarded threat actor "morkibis" disclosed an alleged source code leak of ChatGPT on the vetted leak-sharing community LeakBase that would enable threat actors to instruct ChatGPT to write malicious code. According to the threat actor, the source of the leak was the OpenAI laboratory, though ZeroFox Intelligence has been unable to verify this claim as of this writing. Threat actors have persistently attempted to exploit and manipulate ChatGPT since its release in November 2022, with most seeking to create or discover methods that would allow users to write malicious code effectively.



LeakBase Post Sharing an Alleged Source Code Leak of ChatGPT

Source: ZeroFox Intelligence

Outlook

The threat from synthetic media will likely remain broadly consistent over the next quarter in spite of the rapidly evolving AI landscape. While the frequency of publicly reported attacks is likely to remain sporadic in the short-term, threat actors—ranging from cybercriminals to nation-state actors—will almost certainly continue to seek ways in which to implement synthetic media tools more effectively within both financially and politically motivated campaigns. Specifically, ZeroFox Intelligence anticipates that synthetic

media-driven disinformation campaigns will increase towards the end of 2023, particularly leading up to the 2024 U.S. presidential election. ZeroFox Intelligence assesses that the main focus for the deployment of synthetic media in threat actor toolkits in the near term will be toward social engineering lures and extortion and disinformation campaigns.

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 Synthetic Media Intelligence for ongoing awareness and recommendations for defending against synthetic media-enabled threats.

› 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, disinformation, or malinformation (MDM) targeting the organization.

› Protect

- Apply watermarks and/or use blockchain technology to verify media content.[13]
- Explore digital provenance solutions through the Coalition for Content Provenance and Authenticity (CP2A) or the Content Authenticity Initiative (CAI).

[13] hXXps://c2pa[.]org/

Appendix: 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.