WHITEPAPER

PRECISION DECEPTION: RISE OF AI-POWERED SOCIAL ENGINEERING

TLP:CLEAR

Key Findings

- Artificial intelligence (AI)-generated lures are employed to impersonate executives, disseminate misleading announcements, influence public opinion and support coordinated disinformation campaigns.
- Although AI often is touted as a game-changer for the social-engineering landscape, in the context of phishing, most threat actors still lean on phishing-as-a-service (PhaaS) platforms and off-the-shelf kits and use AI primarily for content drafting and localization.
- AI-assisted voice phishing (vishing) and voice deepfakes are expected to increase as tools become more accessible and threat actors gain expertise.
- Generative AI tools have improved the quality of phishing schemes, business email compromise (BEC) lures and vishing scripts. However, they function more as an efficiency upgrade rather than a fundamental strategy for profit-driven criminal groups.
- There is limited evidence of AI-driven tools circulating in underground markets at the time of this report and discussions among threat actors rarely reference the operational use of generative AI. This absence suggests that while interest may be growing, practical adoption still is in its infancy.
- Widespread use of AI in everyday cybercrime will depend on a decrease in the costs of model hosting and the emergence of “state-of-the-art” AI kits comparable to today’s popular PhaaS offers.

*** Actor handles marked with an asterisk have been redacted for reasons of operational security.**

Overview

Social engineering presents a significant and evolving threat within the cybersecurity landscape that continuously adapts to technological advancements. The emergence of Artificial Intelligence — particularly generative AI — fundamentally transformed the nature of these attacks by enhancing their scale, precision and psychological impact. Where once there were broad, opportunistic scams, we now face highly sophisticated, personalized campaigns that exploit human characteristics with alarming accuracy. As AI tools become increasingly accessible and widely available, malicious actors are leveraging them to develop compelling phishing content, impersonate trusted individuals using deepfake technologies and produce images capable of influencing financial markets. This marks a paradigm shift from volume-based social engineering to a focus on quality and adaptive threat models.

This report offers a comprehensive analysis of the rise of AI-powered social engineering. It investigates the motivations driving modern attacker behavior, profiles key techniques — including spear-phishing, vishing and deepfakes — and evaluates the tool sets that facilitate these campaigns. We provide insights into AI-driven tools our intelligence experts have observed for sale on underground cybercrime forums to illustrate the tangible impacts across various sectors. We also assess the adoption of AI-driven tools in cybercrime operations compared to traditional tools.

Threat Actor Motivation and Objectives

As AI continues to gain prominence, cybercriminals are increasingly adopting the technology to enhance their operations, improve productivity and reduce costs. The KnowBe4 2025 Phishing Threat Trends Report indicates 82.6% of phishing emails identified between September 2024 and February 2025 used AI.[1] Previously, mass phishing campaigns were often recognizable due to awkward phrasing, spelling errors and obvious copy-and-paste templates, which highlighted the limitations of manual writing techniques. However, with the advancement of large language models (LLMs), AI now can generate fluent, multilingual and contextually relevant content across various formats including text, voice, video and images within seconds. This enables a single operator to produce hundreds of professional-looking emails or messages that effectively mimic a company's tone and branding.

The objectives of contemporary cyberattacks have evolved significantly, extending far beyond mere credential harvesting. Today, AI-generated lures are employed to impersonate executives in schemes such as "CEO fraud" wire transfers, disseminate misleading announcements, influence public opinion and support coordinated disinformation campaigns. These tactics not only facilitate BEC and financial fraud, but also pose considerable risks to psychological well-being, organizational reputation and societal integrity.

AI has empowered even novice individuals to execute sophisticated attacks by lowering the technical skill barrier and providing scalable, cost-effective tools. Using either custom-built or off-the-shelf AI agents — whether text-based, voice-enabled or video-driven — malicious campaigns can iterate rapidly, broaden their reach and fluidly transition across various channels, including email, short message service (SMS), voice calls, social media and visual content. As a result, we are witnessing an evolving threat landscape characterized by increasingly convincing, highly personalized social-engineering tactics deployed at a scale and velocity that traditional detection systems and user awareness training are ill-equipped to counter.

AI-powered Techniques for Social Engineering

Note: *We classify AI-driven social engineering into three distinct channels: text, voice and visual/multimedia. Each channel leverages unique human trust signals, employs different AI capabilities and targets victims in specific ways. This structured approach enables us to thoroughly analyze the tools, tactics and targeting strategies threat actors use within each channel.*

- ***Text-based attacks*** *are primarily centered on language and context, often seeking broad outreach or personalized deceptions through emails, chats or messaging platforms.*
- ***Voice-based attacks*** *capitalize on tone, urgency and real-time interaction — specifically targeting high-value individuals via impersonated phone calls or voice messages.*
- ***Visual and multimedia deception*** *involves the use of synthetic images and deepfake videos to impersonate individuals during virtual meetings, and to spread disinformation across social media platforms.*

Text-based Attacks

AI Drives Polymorphic Phishing for Evasion and Spear-phishing

AI is driving a significant shift from traditional phishing attacks to high-volume, polymorphic campaigns characterized by increased personalization, evasion tactics and rapid adaptability. While the adoption of these methods remains gradual and they often are used alongside conventional phishing tools, the trend is becoming increasingly evident. According to KnowBe4, 76.4% of phishing attacks in 2024 incorporated at least one polymorphic element.This figure increased to 82.6% between September 2024 and February 2025. This uptick reflects a broader integration of polymorphism and AI-assisted content generation rather than a complete shift to fully autonomous AI-driven campaigns.

Polymorphic phishing entails systematically modifying observable characteristics, such as sender addresses, subject lines, email content, URLs, file names, headers and even the hypertext markup language (HTML)/cascading style sheets (CSS) structure, so that each message presents itself as unique. These subtle variations enable phishing emails to bypass systems designed to detect known malicious indicators, including Microsoft Defender, native cloud filters and traditional secure email gateways (SEGs). To circumvent domain-authentication checks, the majority of polymorphic phishing emails are dispatched from compromised accounts (52%), followed by phishing domains (25%) and web mail services (20%). KnowBe4 also notes a 47.3% year-over-year increase in the proportion of attacks successfully evading Microsoft's native security measures and SEGs. The report cautions that by 2027, methodologies such as grouping similar emails for phishing detection may become ineffective as the depth of mutation continues to advance. While high-volume polymorphic phishing campaigns remain widespread, there is a growing shift toward more targeted and hyper-personalized attacks — commonly referred to as spear-phishing. According to the Darktrace cybersecurity platform, spear-phishing currently accounts for 38% of all phishing attempts and is expected to increase in the coming years.[2]

Academic and industry studies

A 2025 study on phishing email click-through rates found messages generated by LLMs achieved a markedly higher engagement rate — 57.4% compared with just 13.4% for emails authored by humans.[3] Further research showed modern LLMs can autonomously identify potential targets, gather publicly available information, craft customized lure emails, optimize delivery strategies and adapt based on performance data.[4] While this underscores AI's malicious potential, it remains a theoretical capability rather than a broadly adopted tactic among today's threat actors.

Messages generated by LLMs achieved a markedly higher engagement rate.

To assess how much potential generative AI translated into practical use, Google's Threat Intelligence Group (GTIG) studied real-world interactions between threat actors and Gemini — Google's AI-powered assistant.[5] The findings indicate that although cybercriminals are actively exploring generative AI, it has not yet emerged as a transformative tool in cyber operations. GTIG also observed attackers mainly used AI for support tasks such as research, code troubleshooting, translation and content creation. For advanced threat actors, these tools provide useful frameworks to streamline activities, while less-experienced individuals use generative AI to enhance their learning and execution of established techniques.

On May 30, 2024, the OpenAI research organization disclosed it identified and disrupted five online covert influence operations that exploited its generative AI technologies.[6] The operations sought to manipulate public opinion and influence geopolitics and originated from China, Iran, Israel and Russia. OpenAI's technology was used to generate convincing social media posts, translate and edit articles, write headlines and debug computer programs to promote certain political figures or influence public discourse on geopolitical conflicts. The company ascertained the use of its AI did not appear to have expanded the reach or impact of such efforts.

Our observations

These open source findings closely mirror our observations. Although AI is often touted as a game-changer for the social-engineering landscape, in the context of phishing, most threat actors still lean on PhaaS platforms and off-the-shelf kits and use AI primarily for content drafting and localization — not for true automation or innovation.

Several factors contribute to this ongoing trend, which include:

1. **Complexity of AI integration** — The effective use of AI in phishing campaigns requires more than simple content generation. It involves training or configuring models, automating them within an attack infrastructure, integrating them with delivery systems and devising methods to evade detection. Consequently, many cybercriminals favor plug-and-play phishing kits and PhaaS platforms as these options are easier to implement, faster to deploy and have a proven track record of success.
2. **Resource limitations** — While open source LLMs exist, powerful AI models demand significant computing resources, which are not always accessible to lower-tier or mid-tier cybercriminals. Although jailbroken or underground market models are available, they often come with limitations in functionality, impose access fees or raise concerns regarding reliability.
3. **Effectiveness of traditional tactics** — Mass phishing campaigns, BEC and social engineering via social media remain highly effective strategies, particularly when targeting unsuspecting or inadequately trained users.

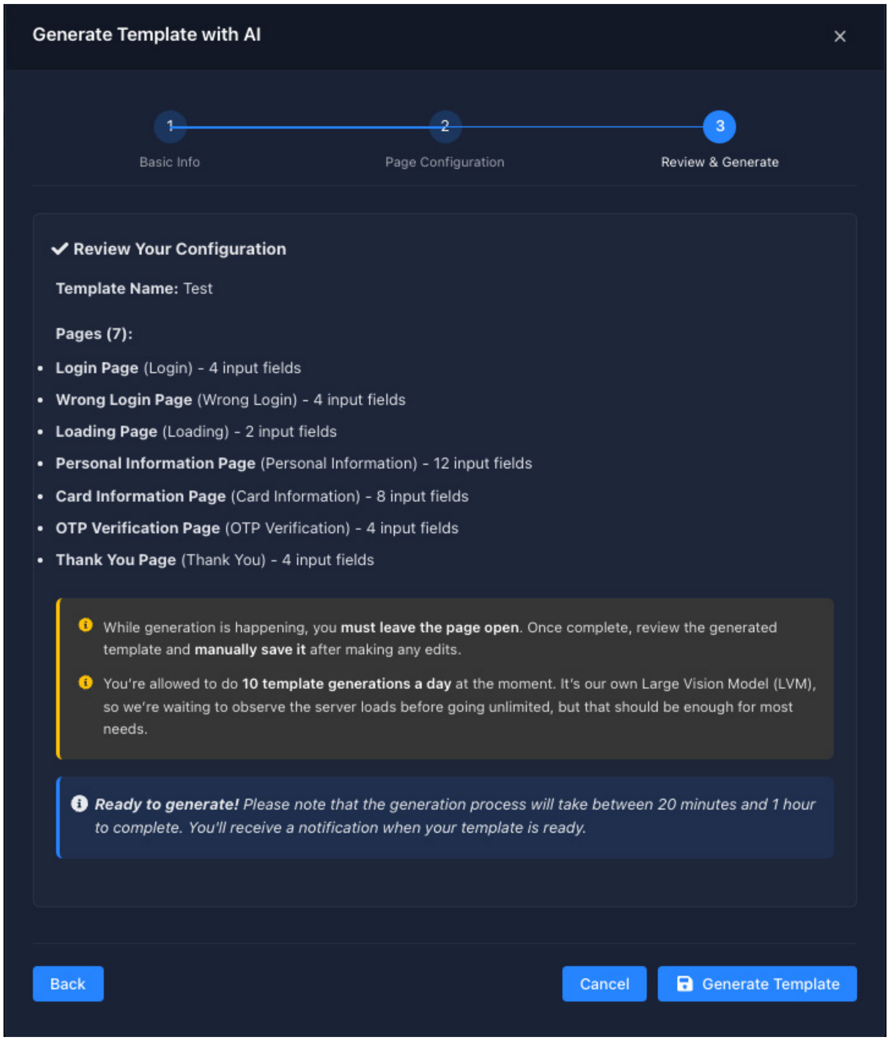
Generative AI platforms and AI-driven underground phishing tools

When selecting a generative AI platform, threat actors have a variety of options that cater to their specific goals, risk tolerance, and technical expertise. One commonly misused platform is ChatGPT, which can be manipulated through jailbroken prompts to circumvent safety mechanisms and generate phishing-like content under seemingly innocuous pretenses. A variety of online resources exist that provide guidance on how to jailbreak the platform (see Figure 1).

Figure 1: The image depicts a screenshot of a YouTube video cover highlighting the top ChatGPT jailbreak prompts captured July 28, 2025

More technically proficient actors may opt for open source LLMs, such as LLaMA or GPT-J, which can be fine-tuned or modified to eliminate restrictions for unrestricted applications. Additionally, some individuals may resort to underground tools such as WormGPT or Xanthorox AI.[7]

Similar to legitimate industries that are leveraging AI to maintain a competitive advantage, actors seeking a PhaaS or phishing kit are looking to integrate AI into their offers to stay relevant in the market. A prominent example is the SheByte PhaaS, which is believed to be the successor to the LabHost PhaaS. In June 2025, we reported the SheByte PhaaS allows users to create their own templates, generate templates using AI or select from a variety of community templates.[8] When generating a template with AI, users can upload a reference image and are encouraged to provide specific instructions to ensure the final page closely resembles the original (see Figure 2). The platform allegedly uses its own large vision model (LVM) and template generation takes 20 minutes to an hour.



✓ Review Your Configuration

Template Name: Test

Pages (7):

- **Login Page** (Login) - 4 input fields
- **Wrong Login Page** (Wrong Login) - 4 input fields
- **Loading Page** (Loading) - 2 input fields
- **Personal Information Page** (Personal Information) - 12 input fields
- **Card Information Page** (Card Information) - 8 input fields
- **OTP Verification Page** (OTP Verification) - 4 input fields
- **Thank You Page** (Thank You) - 4 input fields

While generation is happening, you **must leave the page open**. Once complete, review the generated template and **manually save it** after making any edits.

You're allowed to do **10 template generations a day** at the moment. It's our own Large Vision Model (LVM), so we're waiting to observe the server loads before going unlimited, but that should be enough for most needs.

Ready to generate! *Please note that the generation process will take between 20 minutes and 1 hour to complete. You'll receive a notification when your template is ready.*

Figure 2: The image depicts a screenshot of the SheByte "Generate Template with AI" page captured June 12, 2025.

Voice-based Attacks

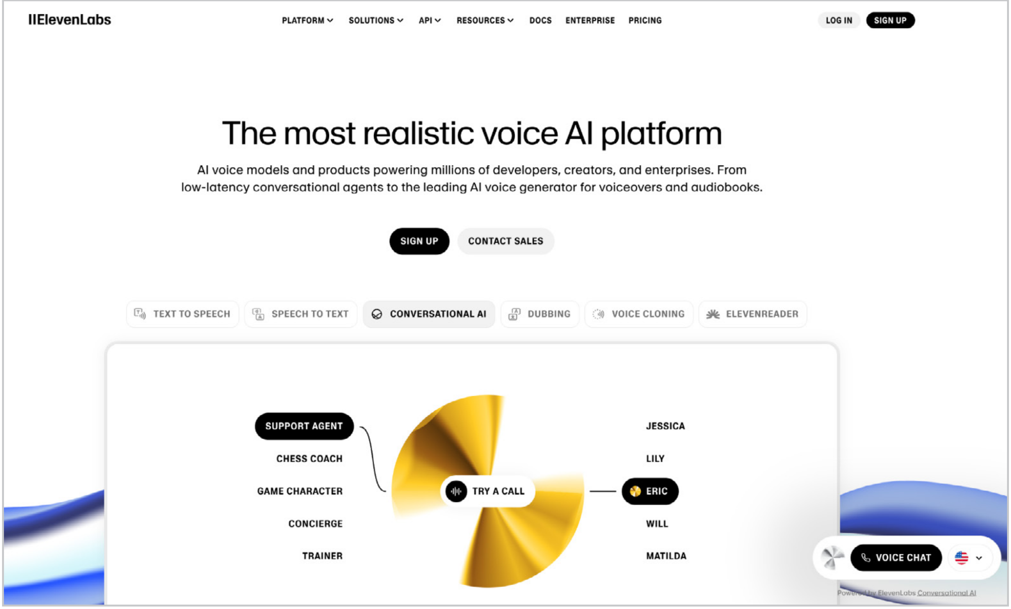
Audio Deepfakes with Little Training Data Drive Vishing and CEO Impersonation Fraud

AI is playing an increasingly significant role in vishing attacks, enhancing their convincibility, scalability and targeting capabilities. According to the CrowdStrike cybersecurity company, vishing incidents surged by 442% between the first and second halves of 2024.[9] Although vishing remains less prevalent than traditional email phishing, the integration of AI in these attacks is gaining traction as threat actors seek to exploit advancements in voice synthesis, natural language processing and deepfake audio technologies.

The primary enabler of AI-driven vishing is neural voice cloning, which is a type of audio deepfake that uses deep learning models to accurately replicate an individual's voice. By training on just a few minutes of audio, malicious actors can generate synthetic speech that convincingly mimics specific individuals. These voice clones are employed to manipulate victims into transferring funds, divulging credentials or granting access to sensitive information — often under the guise of urgency or authority.

Several technologies contribute to this trend, which include:

- **Text-to-speech (TTS) synthesis models:** Microsoft's VALL-E and Google's Tacotron 2, along with commercial platforms such as ElevenLabs and iSpeech, can generate highly realistic synthetic voices that convey emotional nuances (see Figure 3). These tools often require minimal training data — sometimes just a few seconds of audio — to accurately replicate a target voice.

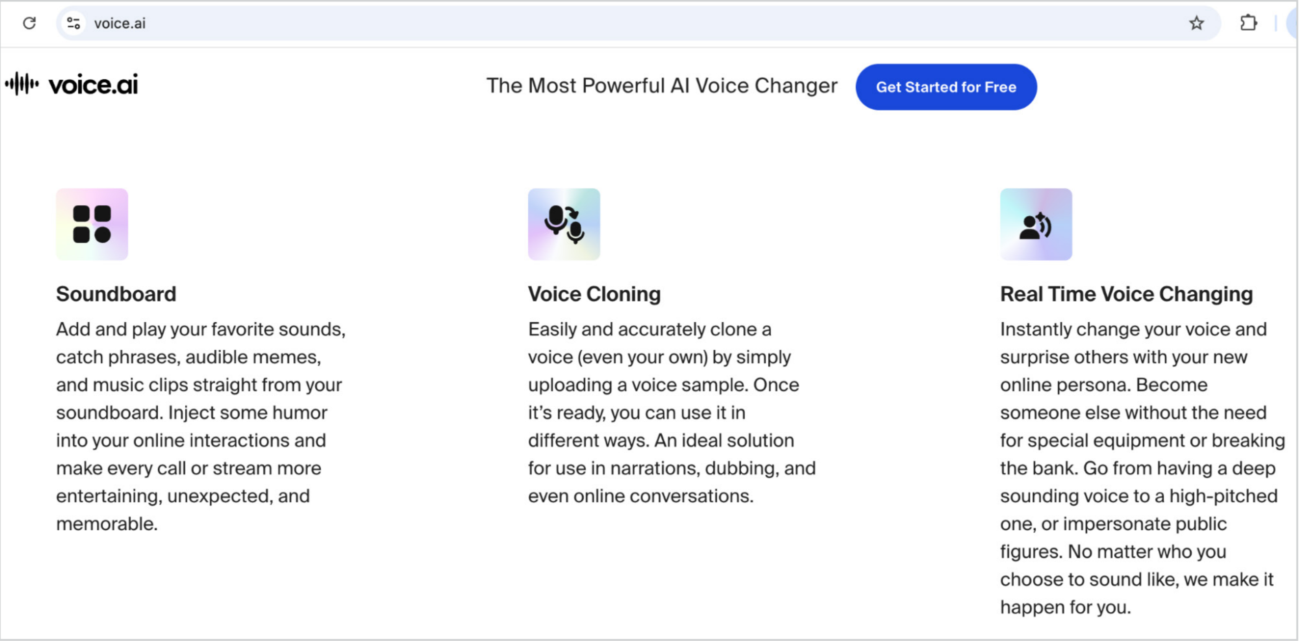


The most realistic voice AI platform

AI voice models and products powering millions of developers, creators, and enterprises. From low-latency conversational agents to the leading AI voice generator for voiceovers and audiobooks.

Figure 3: The image depicts a screenshot of the homepage of the ElevenLabs platform captured July 29, 2025.

- **Real-time voice conversion tools:** Voicemod, iMyFone MagicMic and Voice.ai enable users to impersonate another person's voice during live phone calls (see Figure 4). These tools allow for real-time manipulation of pitch, tone and speech patterns.


voice.ai

voice.ai

The Most Powerful AI Voice Changer

Get Started for Free

Soundboard

Add and play your favorite sounds, catch phrases, audible memes, and music clips straight from your soundboard. Inject some humor into your online interactions and make every call or stream more entertaining, unexpected, and memorable.

Voice Cloning

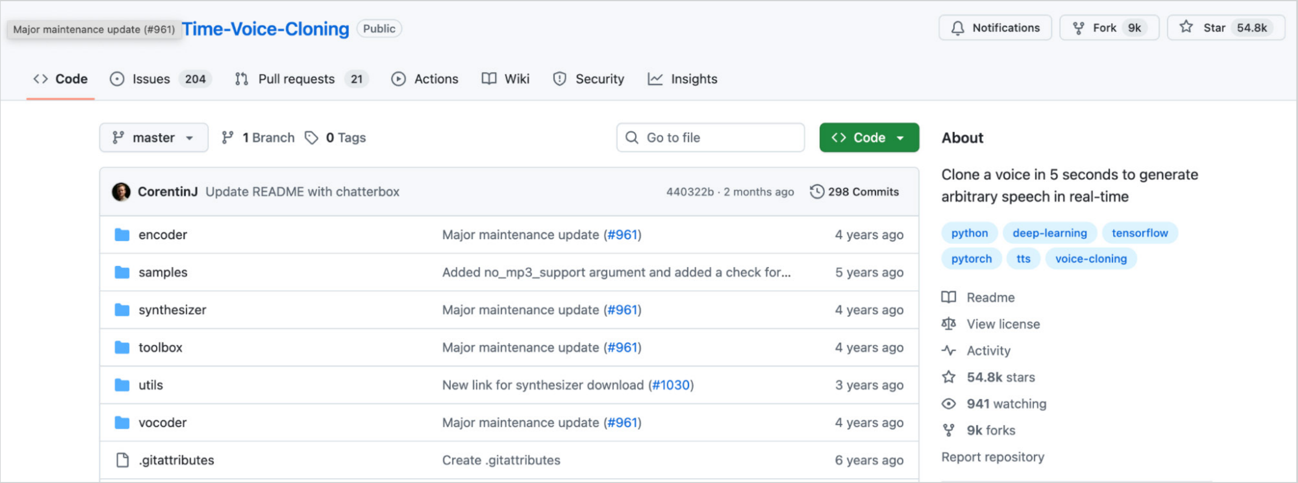
Easily and accurately clone a voice (even your own) by simply uploading a voice sample. Once it's ready, you can use it in different ways. An ideal solution for use in narrations, dubbing, and even online conversations.

Real Time Voice Changing

Instantly change your voice and surprise others with your new online persona. Become someone else without the need for special equipment or breaking the bank. Go from having a deep sounding voice to a high-pitched one, or impersonate public figures. No matter who you choose to sound like, we make it happen for you.

Figure 4: The image depicts a screenshot of the homepage of the Voice.ai platform captured July 29, 2025.

- **Open source audio manipulation frameworks:** Coqui TTS, ESPnet, Descript's Overdub and CorentinJ's Real-Time-Voice-Cloning provide malicious actors with the ability to train, fine-tune and deploy custom voice models using minimal resources (see Figure 5). These platforms are readily available on GitHub and other public repositories, lowering the barrier to entry for sophisticated vishing campaigns.


Major maintenance update (#961) Time-Voice-Cloning Public

Notifications | Fork 9k | Star 54.8k

Code | Issues 204 | Pull requests 21 | Actions | Wiki | Security | Insights

master | 1 Branch | 0 Tags | Go to file | Code

CorentinJ Update README with chatterbox 440322b · 2 months ago 298 Commits

encoder	Major maintenance update (#961)	4 years ago
samples	Added no_mp3_support argument and added a check for...	5 years ago
synthesizer	Major maintenance update (#961)	4 years ago
toolbox	Major maintenance update (#961)	4 years ago
utils	New link for synthesizer download (#1030)	3 years ago
vocoder	Major maintenance update (#961)	4 years ago
.gitattributes	Create .gitattributes	6 years ago

About

Clone a voice in 5 seconds to generate arbitrary speech in real-time

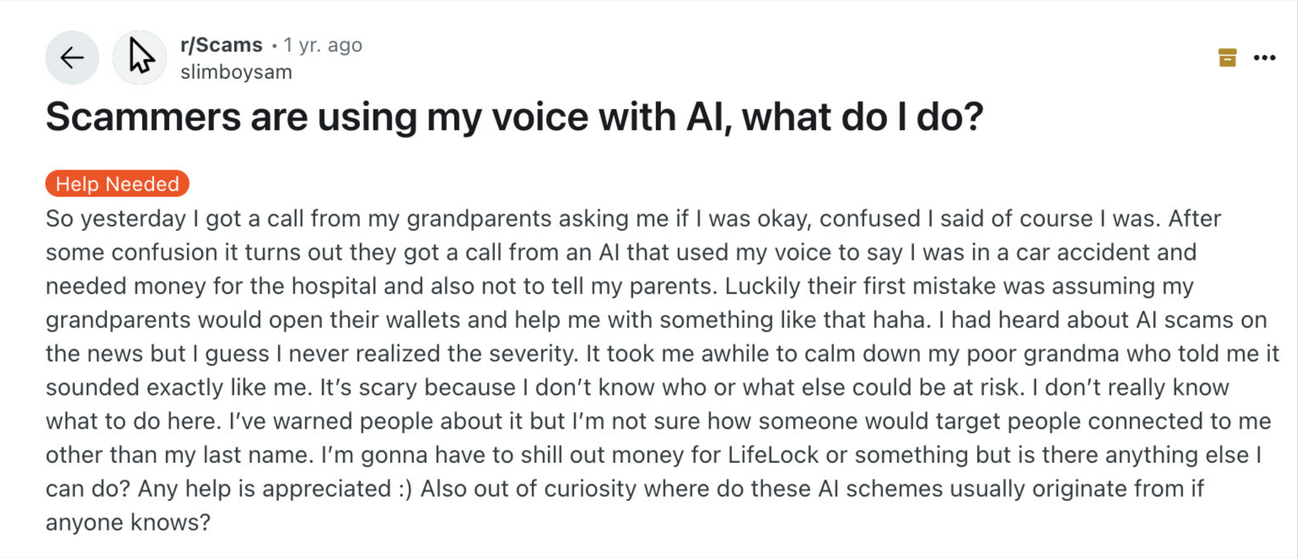
python deep-learning tensorflow pytorch tts voice-cloning

Readme
View license
Activity
54.8k stars
941 watching
9k forks
Report repository

Figure 5: The image depicts a screenshot of the GitHub repository for the Real-Time-Voice-Cloning framework captured July 29, 2025.

Cases of executive impersonation using AI voice cloning have been reported in financial fraud scenarios where attackers successfully tricked employees into authorizing large transfers or disclosing sensitive information. One prominent example involved a deepfake scam targeting WPP CEO Mark Read.[10] Fraudsters cloned his voice and used publicly available images to impersonate him during a fake Microsoft Teams meeting. In another incident, employees at the cybersecurity firm Wiz received voice messages that mimicked their CEO in an attempt to steal credentials using audio extracted from a previous conference appearance.[11]

AI-driven vishing is also affecting everyday individuals. Numerous reports on platforms such as Reddit describe people receiving urgent phone calls from voices mimicking their loved ones, often claiming to be in distress and requesting emergency money transfers (see Figure 6). These emotionally manipulative attacks exploit the victim's trust and sense of urgency, making them particularly effective against unsuspecting individuals who have no reason to question the authenticity of a familiar voice.



r/Scams • 1 yr. ago
slimboysam

Scammers are using my voice with AI, what do I do?

Help Needed

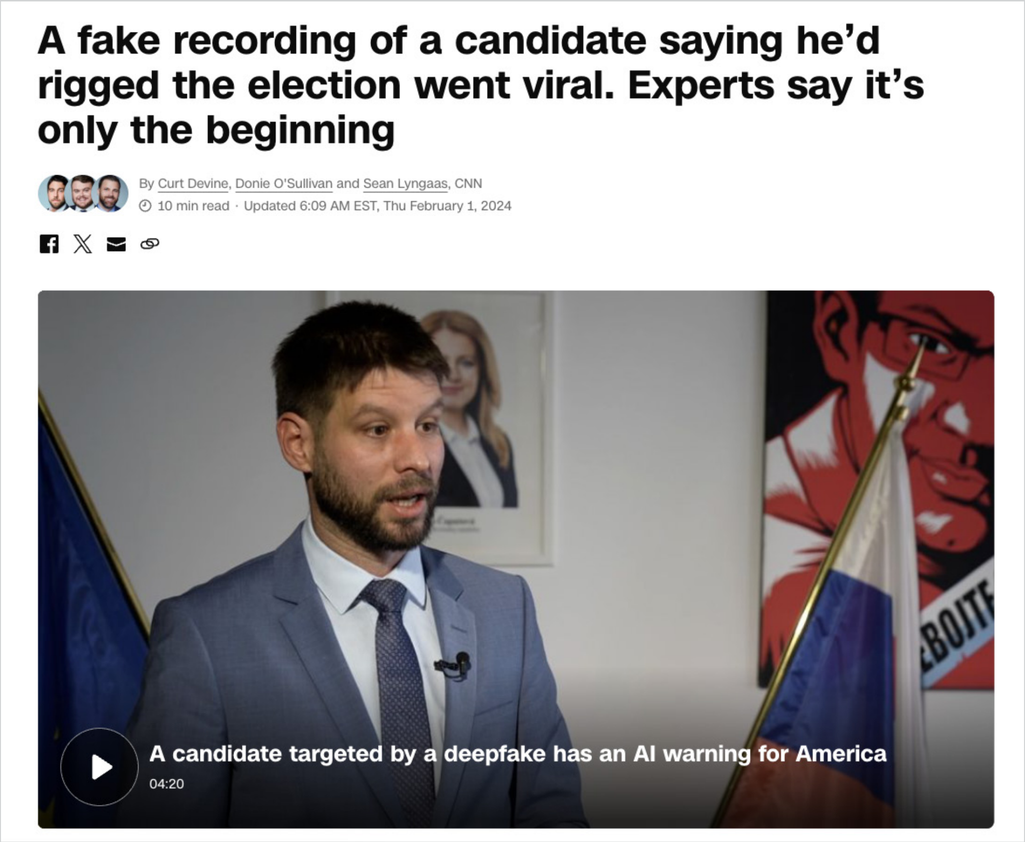
So yesterday I got a call from my grandparents asking me if I was okay, confused I said of course I was. After some confusion it turns out they got a call from an AI that used my voice to say I was in a car accident and needed money for the hospital and also not to tell my parents. Luckily their first mistake was assuming my grandparents would open their wallets and help me with something like that haha. I had heard about AI scams on the news but I guess I never realized the severity. It took me awhile to calm down my poor grandma who told me it sounded exactly like me. It's scary because I don't know who or what else could be at risk. I don't really know what to do here. I've warned people about it but I'm not sure how someone would target people connected to me other than my last name. I'm gonna have to shill out money for LifeLock or something but is there anything else I can do? Any help is appreciated :) Also out of curiosity where do these AI schemes usually originate from if anyone knows?

Figure 6: The image depicts a screenshot of a Reddit post where a user explains how threat actors use AI-generated voices captured July 28, 2025.

On July 25, 2025, The Washington Post published an article about "AI-faked voices" featuring a warning from OpenAI CEO Sam Altman.[12] He specifically highlighted the threat posed by voice impostors targeting users' financial accounts and emphasized the need for financial institutions to move away from voice-based authentication and adopt more secure alternatives.

In addition to financial fraud, voice deepfakes are emerging as tools for political manipulation. Ahead of the Slovakian parliamentary elections in 2023, a two-minute audio clip surfaced that seemed to feature candidate Michal Simečka in a conversation with journalist Monika Tódová (see Figure 7).[13] The recording included fabricated dialogue discussing vote-buying and raising beer prices. Although the clip later was confirmed to be a deepfake — generated by AI trained on the individuals' voices — it circulated widely online during the final days before the election.

This timing raised concerns the deepfake may have influenced public opinion and contributed to Simečka's Progressive Slovakia party finishing in second place.



A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning

By Curt Devine, Donie O'Sullivan and Sean Lyngaas, CNN

10 min read · Updated 6:09 AM EST, Thu February 1, 2024

A candidate targeted by a deepfake has an AI warning for America

04:20

Figure 7: The image depicts a screenshot of a CNN article detailing an incident with Michal Simečka captured July 28, 2025.

While still relatively rare compared to traditional phishing, AI-assisted vishing and voice deepfakes are expected to increase as tools become more accessible and threat actors gain expertise. Unlike email-based phishing, voice impersonation bypasses many digital safeguards, relying entirely on social cues and psychological pressure. As voice synthesis technologies improve and language models are integrated into audio interfaces and call workflows, the threat of vishing is likely to expand in both corporate and consumer environments.

AI-driven Voice Phishing Tools in Underground

Despite the widespread availability of AI-powered vishing tools in the market, some threat actors still prefer those found on underground forums. These illicit tools are typically preconfigured for malicious purposes and often include built-in anonymization features that may not be present in legitimate commercial platforms. Additionally, they frequently offer integration with other illegal services, such as spoofed caller ID modules or credential-harvesting scripts, making them more attractive for orchestrating sophisticated end-to-end attacks. The following section highlights some of the newest tools observed in underground markets.

An AI-powered multipurpose call center platform

On July 14, 2025, we reported the actor ***Gold** allegedly developed and offered to lease a vishing platform that allegedly is an “all-in-one call center” to target Coinbase users.[14] The call center does not require human operators because it is a comprehensive platform built to use AI agents and systems to run custom phishing campaigns. The actor allegedly integrated three AI models into the project: the post-trained Gemma 3 open source LLM from Google, the Whisper V3-large powerful automatic speech recognition and translation model from OpenAI and the Sesame CSM-1B open source conversational speech model.

The platform can run autonomously once a user sets up the phishing campaign. All calls are made by **the actor’s** AI agents who act as callers. They are capable of handling real-time communications and have a voice activity detection (VAD) signal-processing technique implemented. This allegedly allows agents to recognize subtle vocal cues and maintain an apparently humanlike conversation flow. The system allegedly can preserve the context of any conversation for future use in subsequent calls to the same victim.

The actor offered to lease access to the AI-powered call center as a subscription for several thousand US dollars plus a commission on profits. The actor also offered to sell the source code for the project.

AI-driven voice bot for eliciting OTP codes, PINs, and payment card data

On July 22, 2025, we reported the actor ***Bronze** offered to lease AI-powered software for payment card fraud.[15] The actor’s service allegedly includes AI-driven voice over IP (VoIP) software specifically designed to obtain one-time password (OTP) codes. The software aka voice bot uses either prerecorded or AI-generated voice responses to deceive victims into revealing OTP codes, payment card information and personal identification numbers (PINs). The information obtained automatically is sent to the individual’s Telegram bot.

The actor claimed to have received 35 calls from potential victims out of 100 phishing SMS texts sent and allegedly acquired OTPs and payment card data from an average of 10 individuals. The actor sought several thousand US dollars for ready-to-use phishing software, which included prerecorded social-engineering voice scripts and Telegram integration.

Remote job interviews & CEO: Visual and multimedia deception

As generative AI capabilities continue to advance, attackers increasingly are leveraging synthetic media — particularly deepfake videos and AI-generated images — to conduct highly convincing social-engineering campaigns. This form of visual and multimedia deception targets one of the most fundamental human trust signals — appearance. When combined with contextual cues such as names, titles or prior communications, these synthetic assets can significantly enhance the believability of fraudulent interactions.

The following sections analyze the trends related to the use of deepfake videos and AI-generated images, drawing on real-world cases reported in the media and observed on popular public web forums. These cases demonstrate how adversaries exploit these technologies and highlight the tangible consequences of their actions.

Analyst Comment: *We published a series of Finished Intelligence (FINTEL) reports within our cyber intelligence platform, Verity471, that explore the complexities of deepfakes. These reports examine how threat actors are using deepfake technology and offer key recommendations for effective detection. Additionally, they provide comprehensive insight into the creation of deepfakes and the tactics malicious actors employ.*

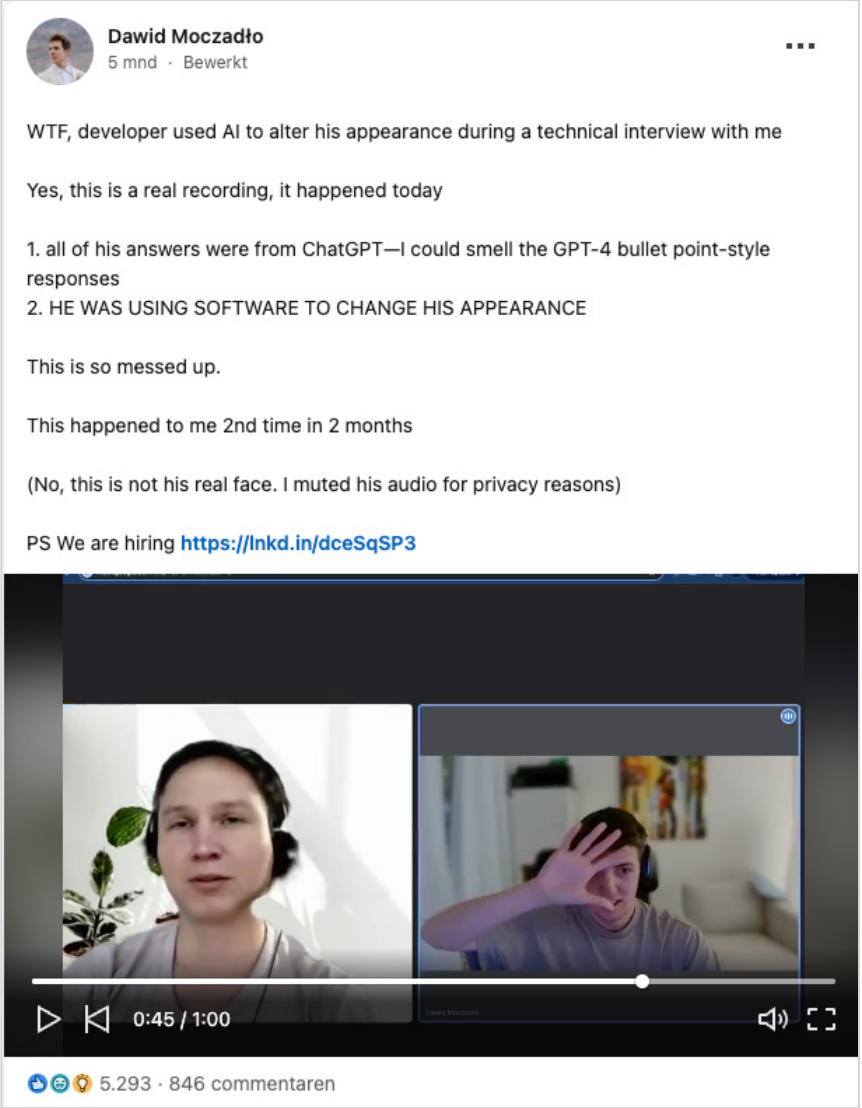
Date	Title	Link
Jan. 23, 2025	**Intelligence Bulletin** — Know-your-customer fraud report series: Underground manuals, common tools, resources	https://verity.intel471.com/intelligence/fintelReportView/report--840498e1-b6b8-5c9a-a694-5c13d2a835f8
March 24, 2025	**Whitepaper** — Know-your-customer fraud report series: Identity fraud techniques	https://verity.intel471.com/intelligence/fintelReportView/report--840498e1-b6b8-5c9a-a694-5c13d2a835f8
May 22, 2025	**Intelligence Bulletin** — Know-your-customer fraud report series: Deepfake technology overview, DeepLiveCam tool test	https://verity.intel471.com/intelligence/fintelReportView/report--fc3c854a-0a23-5f6e-872c-ac1ea623620f
June 13, 2025	**Intelligence Bulletin** — Artificial intelligence capabilities advance, enable cybercrime operations	https://verity.intel471.com/intelligence/fintelReportView/report--0e62c7d9-558c-5985-9281-debff044d1ed

These reports present detailed insights into the mechanics behind deepfakes and their potential for exploitation. To access these and expand your knowledge further, please reach out to our sales team: sales@intel471.com.

Deepfake Videos

Deepfake video technology enables malicious actors to create highly realistic video content that mimics a person's facial expressions, speech patterns and visual mannerisms. This capability can be exploited to impersonate executives in fraudulent video calls, disseminate false announcements or generate misleading content for disinformation campaigns. A prominent instance of deepfake misuse occurred in early 2024 when a finance employee at a multinational corporation was duped into transferring tens of millions of US dollars after participating in a video call composed entirely of AI-generated deepfakes, including a convincing impersonation of the company's chief financial officer (CFO). The employee believed he was engaging with genuine colleagues and authorized the transaction.[16] There are several other examples in open source where regular citizens were targeted with deepfake videos, often for the purposes of fraud, extortion or manipulation.

Another trend we observed is the use of deepfake videos during remote job interviews, enabling malicious actors to pose as legitimate candidates. According to a report by CNBC, 1 in 4 job candidates could be fake by 2028.[17] In a widely shared LinkedIn post, a user described interviewing a candidate for a development role via video.[18,19] When asked to hold his hand in front of his face — a basic liveness check — the candidate ignored the request twice and continued answering questions as if nothing had happened, prompting the interviewer to end the call (see Figure 8).



Dawid Moczadło
5 mnd · Bewerkt

WTF, developer used AI to alter his appearance during a technical interview with me

Yes, this is a real recording, it happened today

1. all of his answers were from ChatGPT—I could smell the GPT-4 bullet point-style responses
2. HE WAS USING SOFTWARE TO CHANGE HIS APPEARANCE

This is so messed up.

This happened to me 2nd time in 2 months

(No, this is not his real face. I muted his audio for privacy reasons)

PS We are hiring https://lnkd.in/dceSqSP3

Figure 8: The image depicts a screenshot of a LinkedIn post highlighting an interview candidate using AI-generated video captured July 29, 2025.

Posts on Reddit and LinkedIn increasingly point to similar encounters where interviewees appear to be using software to manipulate their appearance or mask their true identity in real time. According to the Palo Alto Networks cybersecurity firm, AI has made it possible to construct a convincing synthetic job applicant in as little as 70 minutes.[20] The firm has linked several such incidents to North Korean information technology (IT) workers who use deepfakes and synthetic identities to secure remote jobs at Western technology companies, often as a way to generate foreign income while concealing national affiliations (see Figure 9).

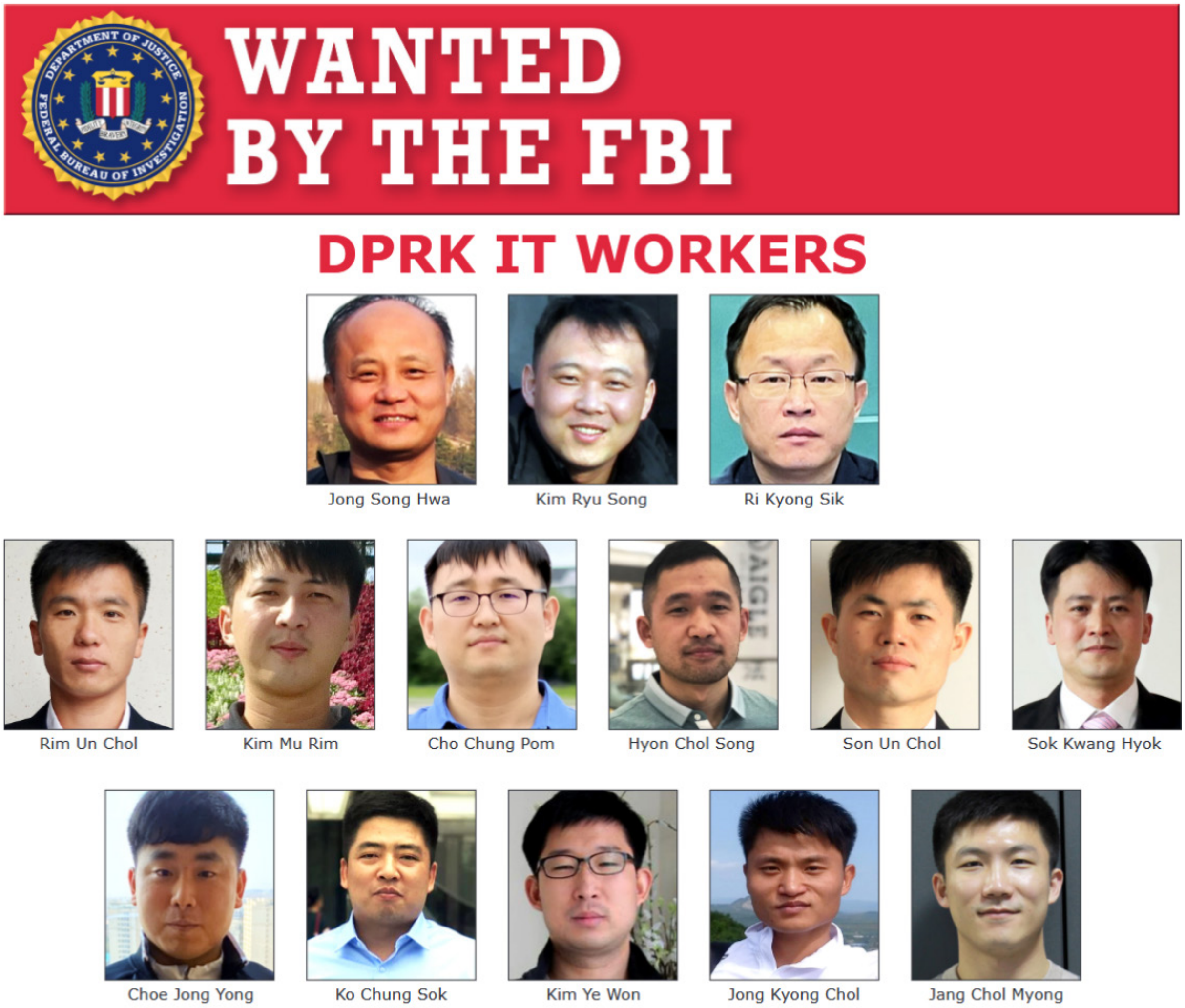


Figure 9: The image depicts a screenshot of a wanted poster highlighted in a Palo Alto report March 20, 2025.

In May 2024, Intel 471 geopolitical analysts reported China is using AI to create news anchor personas to spread disinformation and propaganda.[21] The content of the AI-generated videos varies — from U.S. fuel prices, food costs and gun control, to the shortfalls of the Taiwanese president — but they all focus on issues or individuals perceived to be a threat to China. The clips typically look and sound amateurish, therefore the near-term effectiveness of such online campaigns very likely will be minimal.

Although video deepfakes remain relatively uncommon in day-to-day cybercrime, their misuse is no longer theoretical. High-impact cases demonstrate threat actors already are experimenting with deepfake video in targeted, high-leverage scenarios. As the technology becomes more accessible and faster to deploy, the likelihood of deepfake-enabled deception is expected to rise across a range of threat vectors, particularly in executive impersonation, BEC, financial fraud and remote workforce manipulation. In parallel, disinformation campaigns, including politically motivated video forgeries, are poised to become a growing vector of concern in the future.

Artificial Intelligence-generated Images

AI-generated images, including synthetic profile photos, forged documents and fabricated scenes, also increasingly are being exploited in social-engineering attacks aimed at manipulating trust and deceiving targets. Especially prevalent is the use of hyper-realistic profile pictures on fake LinkedIn or social media accounts and phishing websites to impersonate recruiters, vendors or internal employees. These visuals help attackers build credibility and lower suspicion in schemes such as BEC, credential harvesting and fraudulent onboarding. Because AI-generated faces often appear flawless and authentic, especially in an era of normalized image filters, they can bypass both human judgment and basic verification tools (see Figure 10).

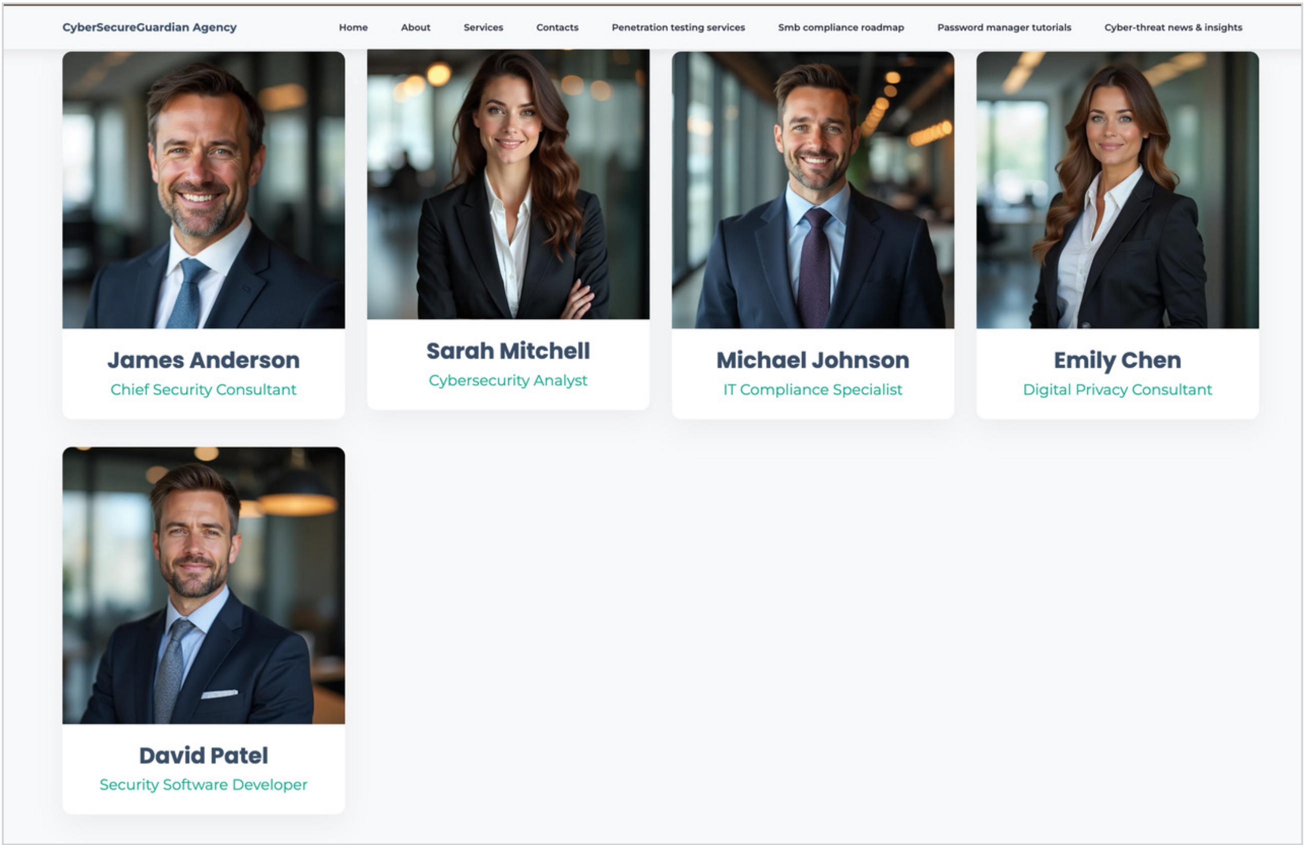

CyberSecureGuardian Agency
Home
About
Services
Contacts
Penetration testing services
Smb compliance roadmap
Password manager tutorials
Cyber-threat news & insights
James Anderson
Chief Security Consultant
Sarah Mitchell
Cybersecurity Analyst
Michael Johnson
IT Compliance Specialist
Emily Chen
Digital Privacy Consultant
David Patel
Security Software Developer

Figure 10: This image depicts an example of AI-generated employee images on the fraudulent website keepass-security.com, which the security researcher Jerome Segura [shared on LinkedIn](#).

Another concerning use for AI-generated images is face-swap sextortion where attackers use generative tools to graft a victim's face onto explicit imagery and then threaten to publish it unless paid, or simply deploy it as a form of harassment.[22] These attacks overwhelmingly target women and girls — including public figures and, increasingly, middle and high school students — causing severe psychological, reputational and social harm. Academic surveys across 10 countries showed roughly 2.2% of respondents have been victimized by non-consensual synthetic intimate imagery.[23]

AI-generated images also are leveraged to generate fake news. Fabricated images of disasters, violence or political unrest are circulated to create panic, sway public opinion or manipulate media narratives — often part of influence or disinformation operations. A notable example of this is a fake AI-generated image of an explosion near the Pentagon May 23, 2025, that spread on X aka Twitter via a verified-looking account and briefly caused a dip in the stock market.[24]

Tools for Deepfake Videos, Artificial Intelligence-generated Images

Multiple actors offer to create deepfake synthetic videos

On July 29, 2025, we reported the actor **Silver***allegedly developed an AI project capable of generating a short deepfake video in about 30 to 40 minutes when installed on a high-performance computer equipped with a graphics card, such as an Nvidia GeForce RTX 4090 model.[25] An additional 20 minutes is required for voice-cloning procedures and the software allows users to produce deepfakes in any language. The actor sought to sell the project for thousands of US dollars and would provide training on its deployment and optimal use.

The actor presented two samples of deepfake videos. One featured U.S. President Donald Trump "promoting" the deepfake service. The quality of the deepfake is relatively high, demonstrating the actor's technical capabilities (see Figure 11).

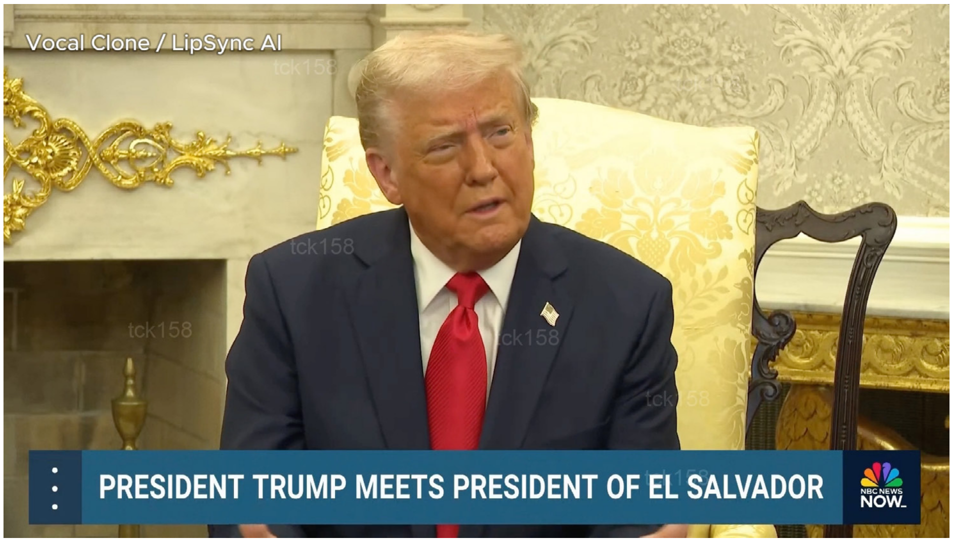


Figure 11: The image depicts a screenshot of the deepfake video **Silver*** *created captured July 29, 2025.*

Assessment

Generative AI tools have improved the quality of phishing schemes, BEC lures and vishing scripts. However, they function more as an efficiency upgrade rather than a fundamental strategy for profit-driven criminal groups. The economics are still not favorable — a ready-to-use PhaaS kit costs significantly less than US $200 and ensures full anonymity. In contrast, training or fine-tuning a private AI model, sourcing clean audio for voice cloning or creating realistic face swaps requires considerable time, expertise and resources, which can diminish profit margins. Latency and error rates also reduce the appeal — one mistake in a deepfake video call or synthetic voice can derail an otherwise flawless scam.

There is limited evidence of AI-driven tools circulating in underground markets.

Moreover, there is limited evidence of AI-driven tools circulating in underground markets at the time of this report and discussions among threat actors rarely reference the operational use of generative AI. This absence suggests that while interest may be growing, practical adoption is still in its infancy. Additionally, public application programming interfaces (APIs) for LLMs create audit trails that many criminals would prefer to avoid. In short, for most threat actors, traditional tactics still offer the best risk-to-reward ratio.

AI is proving to be beneficial for disinformation and influence operations.

On the other hand, AI is proving to be beneficial for disinformation and influence operations. Actors involved in propaganda — whether state-sponsored or ideologically driven — do not require real-time perfection. Instead, they need scalability, linguistic reach and a sense of authenticity. Generative text, voice-overs and imagery meet these requirements at a minimal cost, allowing small teams to flood channels with culturally tailored narratives, deepfake commentary and convincing videos that previously required professional-grade studios. Since the measure of impact is resonance rather than direct monetization, minor inaccuracies or the risk of detection are acceptable trade-offs, making AI a natural accelerator for misinformation campaigns.

Looking ahead, we are likely to see selective escalation rather than mass adoption. We can expect more deepfake-enabled impersonation calls targeting executives and AI-voiced fraud against high-value targets, alongside a surge in synthetic media during elections, geopolitical flash points and social justice debates. Widespread use in everyday cybercrime will depend on a decrease in the costs of model hosting and the emergence of "state-of-the-art" AI kits comparable to today's popular PhaaS offers. Until those challenges are addressed, generative AI will continue to refine old tactics for financially motivated actors while driving the next wave of influence operations for those who trade in opinions rather than cash.

Recommendations

AI-driven social-engineering tactics such as hyper-personalized phishing messages, deepfake voice calls and synthetic media are rapidly escalating in scale and sophistication. To address these growing threats, we recommend:

Harden

Caller verification protocols: Require dual-channel verification for sensitive requests, such as payment approvals, wire transfers or credential resets, received via phone or voicemail.

- **Benefit:** Blocks real-time deepfake voice scams by inserting a human or process checkpoint, denying attackers a single point of failure.

Retire insecure identity verification processes: Eliminate “voice ID,” caller name recognition and unverified video submissions as authentication factors for high-risk workflows.

- **Benefit:** Removes outdated trust mechanisms that AI tools now easily spoof with synthetic voice and video.

Protect executive media assets: Watermark and digitally sign executive videos and public communications using verifiable content authenticity standards such as the Coalition for Content Provenance and Authenticity (C2PA).

- **Benefit:** Makes it harder for attackers to repurpose existing media into convincing impersonation deepfakes.

Detect

Synthetic media anomaly detection: Deploy deepfake and cloned voice detection tools in inbound communication channels, i.e., help desk, talent acquisition, etc.

- **Benefit:** Identifies manipulated media assets and voice attacks before they influence business workflows or personnel.

Monitor for AI-generated text patterns: Use content inspection and anomaly detection in email and chat to flag AI-style language such as an overly formal tone, synthetic phrasing or generic urgency.

- **Benefit:** Detects LLM-generated phishing or impersonation attempts that bypass traditional spam.

Threat intelligence ingestion for AI tool abuse: Subscribe to threat intel feeds that track AI toolkits and impersonation campaigns, such as deepfake marketplaces.

- **Benefit:** Keeps detection rules and awareness aligned with how attackers currently are using generative AI in the wild.

Prevent

Executive-targeted impersonation drills: Train executive teams and their staff using real examples of deepfake calls, cloned voicemails and urgent AI-generated messages.

- **Benefit:** Builds muscle memory for recognizing red flags and following fallback verification procedures under pressure.

Media literacy training for all staff: Incorporate short, recurring training modules to help employees recognize synthetic media cues such as lip-synch errors, visual glitches or unnatural pauses.

- **Benefit:** Empowers employees to visually and audibly detect deepfakes before responding or escalating incidents.

Runbooks for synthetic impersonation events: Develop and rehearse playbooks for suspected AI impersonation events — whether a deepfake video, fraudulent audio message or suspicious internal directive.

- **Benefit:** Ensures rapid, confident response to incidents that otherwise may create doubt or panic.

Runbooks for synthetic impersonation scenarios: Establish and rehearse specific response plans for suspected impersonation via AI, including executive spoofing, fake video directives and audio scams.

- **Benefit:** Reduces confusion and time to contain during fast-moving AI-enabled fraud attempts.

Sources

To access Intel 471 reports referenced in this paper, please reach out to our sales team: sales@intel471.com.

1. 20 March 2025. KnowBe4. Phishing Threat Trends Report

2. 19 February 2025. Darktrace. Darktrace's 2024 Annual Threat Report Reveals Continued Rise in MaaS Threats and Growing Use of Evasion Tactics

3. 11 March 2025. Utrecht University. Impact of AI Personalization on Email Clicks and Conversions: Insights from a Real-World AI-Personalized Phishing Simulation

4. 30 Nov 2024. Arxiv — Heiding et al. Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects

5. 29Jan2025. GTIG. Adversarial Misuse of Generative AI

6. OpenAI statement

7. Xanthorox AI

8. Intel 471 Report: SheByte

9. 2025. Crowdstrike. Global Threat Report

10. 10 May 2024. Guardian. CEO of world's biggest ad firm targeted by deepfake scam

11. 28 Oct 2024. TechCrunch. Wiz CEO says company was targeted with deepfake attack that used his voice

12. 25 July 2025. The Washington Post. OpenAI CEO Sam Altman is right and very wrong about AI-faked voices

13. 1 Feb 2024. CNN. A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning

14. Intel 471 Info Report: AI-powered multipurpose call center platform by the actor Gold

15. Intel 471 Info Report: AI-powered card fraud tool

16. 4 Feb 2024. CNN. Finance worker pays out $25 million after video call with deepfake 'chief financial officer'

17. 11 Jul 2025. CNBC. How deepfake AI job applicants are stealing remote work

18. 11 Feb 2025. The Register. I'm a security expert, and I almost fell for a North Korea-style deepfake job applicant

19. 09 Feb 2025. LinkedIn. Dawid Moczadło's post

20. 21 Apr 2025. Palo Alto. False Face: Unit 42 Demonstrates the Alarming Ease of Synthetic Identity Creation

21. Intel 471 Geopolitical Sport Report

22. 10 June 2024. ICCT. Exploitation of Generative AI by Terrorist Groups

23. 13 Feb 2024. Arxiv — Umbach et al. Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries

24. 23 May 2023. Guardian. Fake AI-generated image of explosion near Pentagon spreads on social media

25. Intel 471 Actors Silver offer to create deepfake synthetic videos

About Intel 471

Intel 471 equips enterprises and government agencies with intelligence-driven security offerings powered by real-time insights into cyber adversaries, threat patterns, and potential attacks relevant to their operations. By integrating human-sourced intelligence with advanced automation and curation, the company's platform enhances security measures and enables teams to bolster their security posture by prioritizing controls and detections based on real-time cyber threats. Organizations are empowered to neutralize and mitigate digital risks across dozens of use cases across our solution portfolios: Cyber Threat Exposure, Cyber Threat Intelligence, and Cyber Threat Hunting. Learn more at intel471.com.

Our customers' eyes and ears outside the wire.