Deloitte.

Threat Report: How threat actors are leveraging Artificial Intelligence (AI) technology to conduct sophisticated attacks

Global Threat Intelligence by Deloitte

March 2024

What's Inside

Overview

Threat actors are automating reconnaissance and launching customized phishing attacks to deploy malware, such as information stealers, that place organizations across industries at risk of cyber-attack.

This report details the known campaigns where threat actors leveraged AI technology to conduct the attack, known attack vectors, and Deloitte Cyber Threat Intelligence (CTI) observations, as well as provides recommended actions for helping prevent an AI-based attack.

Summary

- Since at least 2022, threat actors began actively discussing and allegedly leveraging AI technologies to facilitate and orchestrate malicious operations.
- Threat actors attempt to leverage AI tools to enhance malware or develop new malicious software and create social engineering schemes, such as persuasive phishing emails personalized to the recipient to increase the chances of attack effectiveness.
- Several underground forum administrators offer AI-based exploitation tools and integrated AI tools into their platforms for real-time interactions, supporting malicious content creation and thus lowering the entry barrier for less sophisticated threat actors or groups.

Threat Analysis

Introduction

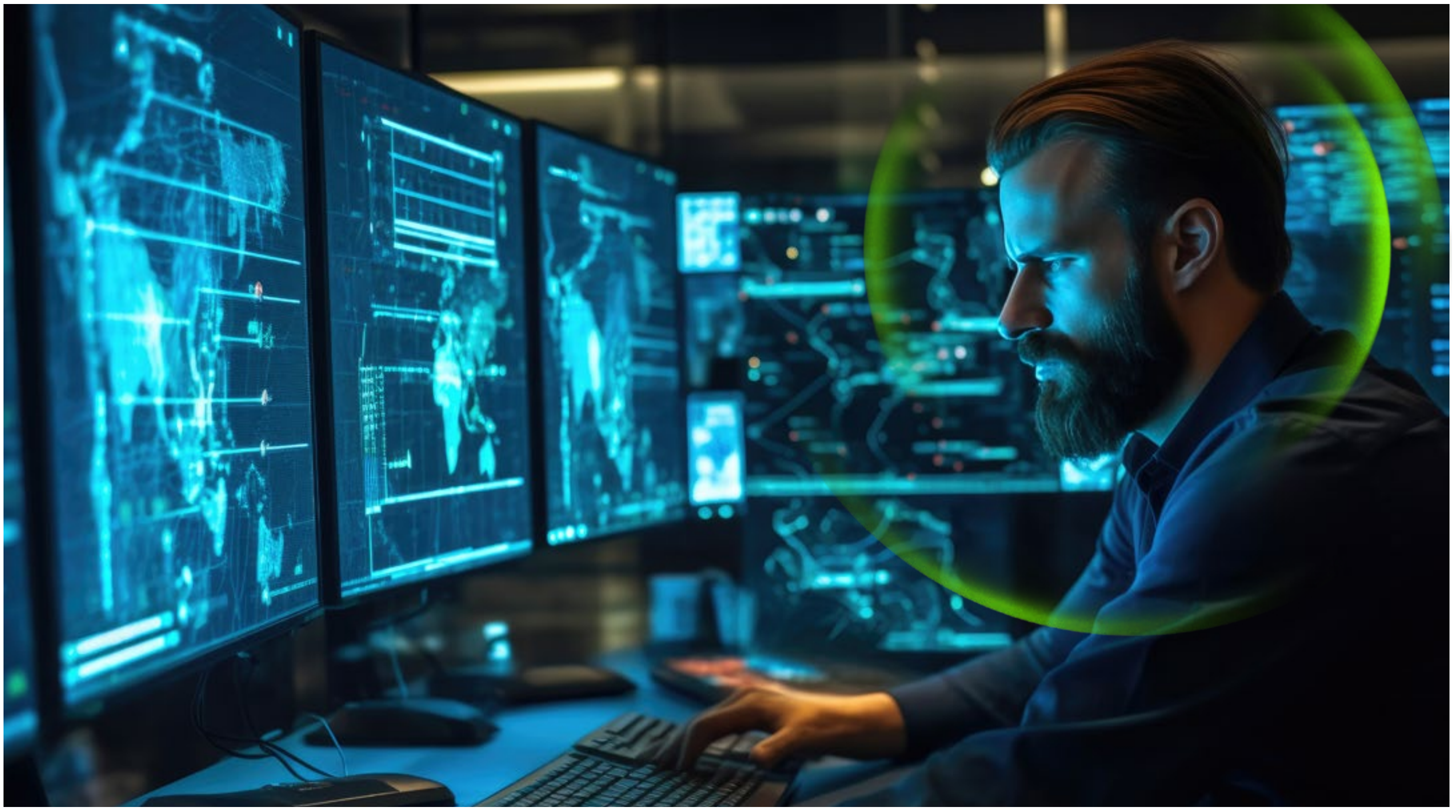
Generative AI is a step beyond the AI used for automation and pattern recognition in the advancement of large language models (LLMs). Generative AI produces content based on what it "learns" through these capabilities. Examples of Generative AI put to work include chatbots, image generators, and other AI-driven media content generators. In medical settings, for example, Generative AI can produce medical imaging, analyses, and in some cases, diagnosis.[1]

In cybersecurity, Generative AI is a double-edged sword. On one hand, it can be used to identify and contain threats and consolidate data to articulate lessons learned from a threat or incident. However, threat actors can also use Generative AI to conduct sophisticated phishing and social engineering attempts and share false information.[2] There are two major components in adversarial use of Generative AI. First, threat actors can use AI to assist in various types of cyberthreat campaigns. Threat actors can use chatbots or auto-generated content to create realistic-sounding social engineering text and lures, while also using Generative AI tools to obtain, synthesize, and adjust code to execute complex malware.[3] The other component of AI-based threats lies in the capability to corrupt existing tools through data poisoning and adversarial imaging. AI-based threats also pertain to life science and healthcare organizations as they adopt AI-based tools to conduct diagnostics and analysis. Data poisoning involves feeding false information into datasets used in machine learning (ML), therefore falsifying the output of the tools.[4]Adversarial imaging applies a similar concept, but toward images; it tricks neural networks into falsely identifying images or generating inconclusive results.[5]

Analysis

During the research and investigation, Deloitte CTI observed that threat actors primarily leverage AI-based technology through toolkits and chatbots. Threat actors primarily used AI-based toolsets to conduct either of the below-mentioned three actions. Theseattack vectors are detailed further down the report.[6]

- Creating phishing customized emails
- Performing network scans to identify vulnerabilities in the compromised host or network
- Phishing web pages



Threat actors leveraging AI

Incidents in which threat actors use AI at various stages in their campaigns are difficult to track as the threat actors predominantly use AI in the initial phases of the attacks only (initial access, reconnaissance). For example, an AI tool can design a victim-specific customized phishing email which is difficult to distinguish from a threat actor designed phishing email. However, security research modeling and correlative statistics provide plausible evidence of threat actor capabilities in enhancing their campaigns with AI. In a survey conducted in 2023 across various industries, 46 percent of organization respondents expressed concern about increased vulnerability to attacks amid widespread use of Generative AI. Of these respondents, 39 percent expressed concerns with privacy, 37 percent expressed concerns with undetectable phishing attacks, and 33 percent expressed concern with the volume of potential cyberattacks.[7] Additionally, healthcare organizations have also expressed concern that the use of AI to draw from datasets and generate information could mix private and public information and lead to Health Insurance Portability and Accountability Act (HIPAA) violations.[8] For example, with an AI tool providing access to confidential health care information of patients, or sensitive information related to pharmaceuticals and medicines, threat actors can misuse chemical formulas to conduct financially motivated or targeted attacks.

Generative AI offers threat actors tools to automate cyberattacks, scan attack surfaces, and generate phishing content tailored towards geographic regions and specific demographics. These tools assist threat actors to broaden their base of victims and heighten the effectiveness of their campaigns. AI helps cybercriminals automate attacks, scan attack surfaces, and generate content that resonates with various geographic regions and demographics, allowing them to target a broader range of potential victims across different countries.[9] The use of Generative AI in phishing attacks complicates some of the countermeasuresorganizations use to detect phishing emails. For example, one of the typical signs of phishing content features grammatical errors and inconsistent language indicative of a non-native speaker of English or the victim's target language. Chatbots can produce grammatically correct, convincing content, and can even draw upon themes and concerns of the target organization.[10 11]

Threat actors have also used AI to find new attack vectors. On the underground marketplace, threat actors can adjust the malware-as-a-service business models to employ LLMs to automate exposure and vulnerability identification and quickly find leaked credentials.[12] According to a US federal government health organization presentation, threat actors are using AI to design and execute attacks against the healthcare industry, these vectors include impersonation attacks, rapid vulnerability exploitation, development of complex malware code, deeper and more efficient reconnaissance, and capabilities to overwhelm an organization's human defenses.[13] During the first half of 2023, security researchers observed that chatbots can write code tailored to mobile malware and create phishing lures. Although researchers have not yet observed malware entirely conceived and developed by AI being deployed in the wild, Generative AI chatbots can replicate and alter code for existing malware variants. This code is being distributed in the dark web marketplace.[14]

Threat actors can manipulate AI in two distinct ways. Data poisoning ruins the quality and integrity of a data set from which a ML process draws its information. Adversarial images are deliberately altered images that lower an AI's confidence in identifying images overall.[16] Adversarial images can be made with minor tweaks to existing images that make them share commonalities with other identifiable images and therefore make them hard to distinguish correctly. As an example, threat group "TA499" targeted high-profile Individuals with video call requests: On March 7, security researchers reported on a malicious email campaign by a Russia-aligned threat group dubbed "TA499" (aka Vovanand Lexus).[15] The threat group targeted high-profile government officials, well-known businesspeople, and celebrities who donated to and supported Ukraine.According to the report, the email was disguised as coming from various Ukraine government organizations and individuals to request information from the targeted individuals and lure them into further contact via phone calls or remote video. The threat group also used an embassy-themed email address to send emails with a subject line related to international aides and assistance of senior government officials. The threat group impersonates government officials using deepfake artificial intelligence software in video calls. The conversations begin seriously, encourage the target to voluntarily speak as much information as possible, and request financial support. Once the target makes a statement on the subject, it is edited and posted on social media for Russian-and English-speaking audiences.[16]

Surface web and underground observations

During the research, Deloitte CTI observed several AI-based chatbots and toolkits being offered on open-source platforms and underground forums, where threat actors are offering the AI-based applications and toolkits to conduct large scale attacks. A few of the examples are shared below.

1. **Nebula, an AI-based penetration testing toolkit:** On October 8, an online code sharing website user aliased 'berylliumsec' published Nebula, a Python-based penetration testing toolkit that leverages AI technology.[17] According to 'berylliumsec', Nebula features a natural language interface that enables users to interact with integrated tools, input commands, and view results. Nebula can also process and analyze scan results and provide users with command recommendations for further investigation of the identified potential risks. berylliumsecprovided sample use cases to demonstrate how Nebula's natural language interface can be used to communicate various tasks and commands to the integrated tools. Nebula also includes several open-source tools to facilitate penetration testing and vulnerability assessments such as:

- NMAP (Network Mapper), a network scanning tool used to discover hosts and services on a computer network by sending packets and analyzing the responses. It can also be used to detect vulnerabilities across various services, including Hypertext Transfer Protocol (HTTP), FTP (File Transfer Protocol), or Server Message Block (SMB) protocol.

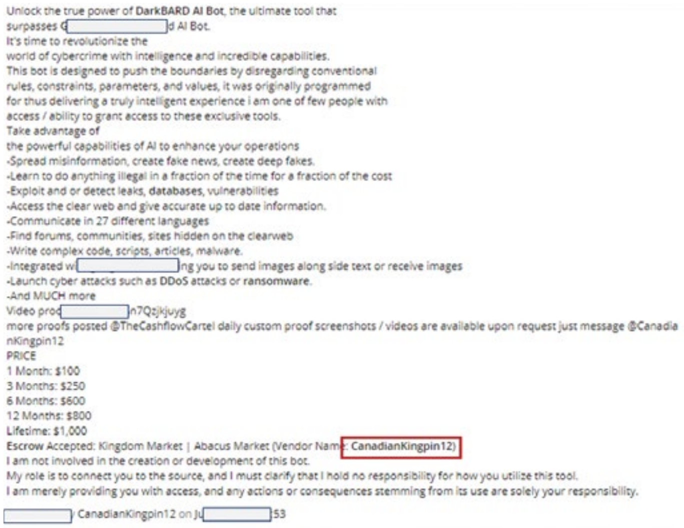
- Open Web Application Security Project Zed Attack Proxy (OWASP ZAP), an open-source web application security scanning tool to detect security flaws in web applications, including cross-site scripting (XSS) and Structured query language (SQL) injection vulnerabilities.

- Crackmapexec, a post exploitation tool to gain administrative privilege of the compromised host and mapping Active Directory (AD) networks, including enumerating users, groups, permissions, Organizational Units (OUs), trust relationships, and Group Policy Objects (GPOs).

- Nuclei, a template-based tool for identifying security vulnerabilities in web applications, APIs, and other networked devices.

2. **Threat actors advertising AI-based chatbots on the underground forums:** In July and August of 2023, using Deloitte internal sources, Deloitte CTI observed users aliased 'canadiansmoker' and 'CanadianKingpin12' advertised the access to DarkBARDon the underground forums.[18] According to the post, DarkBARDallows threat actors to create fake news and deep fakes to spread false information, launching distributed denial of service (DDoS) attacks, ransomware operations, and other cyberattacks, writing malicious codes and scripts to create malware, detecting and exploiting vulnerabilities and database leaks, accessing communities, forums, and websites that are hidden from the clear web. According to the claim, DarkBARDcan communicate in 27 different languages and can send and receive images. DarkBARDwas advertised at $100 USD for a 1 month subscription, $250 USD for 3 months, $600 USD for 6 months, $800 USD for 12 months, and $1,000 USD for a lifetime subscription.

Figure 1: AI-based chatbot advertised on dark web forum.

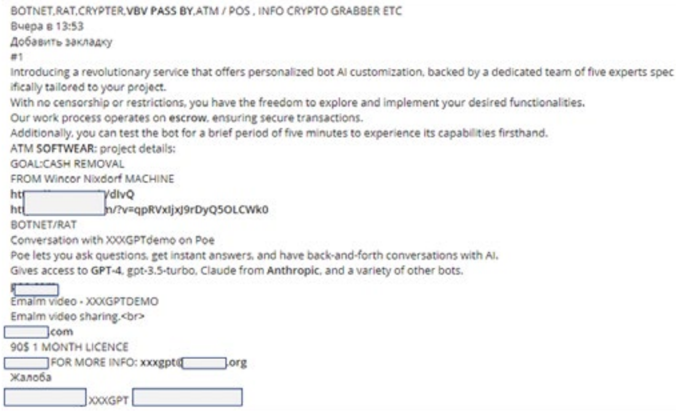


Unlock the true power of **DarkBARD AI Bot**, the ultimate tool that
surpasses G[redacted]d AI Bot.
It's time to revolutionize the
world of cybercrime with intelligence and incredible capabilities.
This bot is designed to push the boundaries by disregarding conventional
rules, constraints, parameters, and values, it was originally programmed
for thus delivering a truly intelligent experience i am one of few people with
access / ability to grant access to these exclusive tools.
Take advantage of
the powerful capabilities of AI to enhance your operations
-Spread misinformation, create fake news, create deep fakes.
-Learn to do anything illegal in a fraction of the time for a fraction of the cost
-Exploit and or detect leaks, **databases**, vulnerabilities
-Access the clear web and give accurate up to date information.
-Communicate in 27 different languages
-Find forums, communities, sites hidden on the clearweb
-Write complex code, scripts, articles, malware.
-Integrated wi[redacted]ng you to send images along side text or receive images
-Launch cyber attacks such as **DDoS** attacks or **ransomware**.
-And MUCH more
Video proo[redacted]n7Qzjkjuyg
more proofs posted @TheCashflowCartel daily custom proof screenshots / videos are available upon request just message @CanadianKingpin12
PRICE
1 Month: $100
3 Months: $250
6 Months: $600
12 Months: $800
Lifetime: $1,000
Escrow Accepted: Kingdom Market | Abacus Market (Vendor Name: **CanadianKingpin12**)
I am not involved in the creation or development of this bot.
My role is to connect you to the source, and I must clarify that I hold no responsibility for how you utilize this tool.
I am merely providing you with access, and any actions or consequences stemming from its use are solely your responsibility.

[redacted] CanadianKingpin12 on Ju[redacted]53

3. **XXXGPT advertised on underground forum:** On July 29, 2023, a threat actor aliased 'XXXGPT', offered to sell a custom artificial intelligence chatbot capable of creating malware, including botnets, remote access trojans (RATs), crypters, information stealers, cryptocurrency stealers, point-of-sale (POS) and automated teller machines (ATM) malware, Verified by Visa (VbV) security check bypass malware on an underground forum. XXXGPT provided a video demonstrating the XXXGPT on an application that allows users to access and interact with a variety of AI models.[19]

Figure 2: XXXGPT offering on the dark web forum

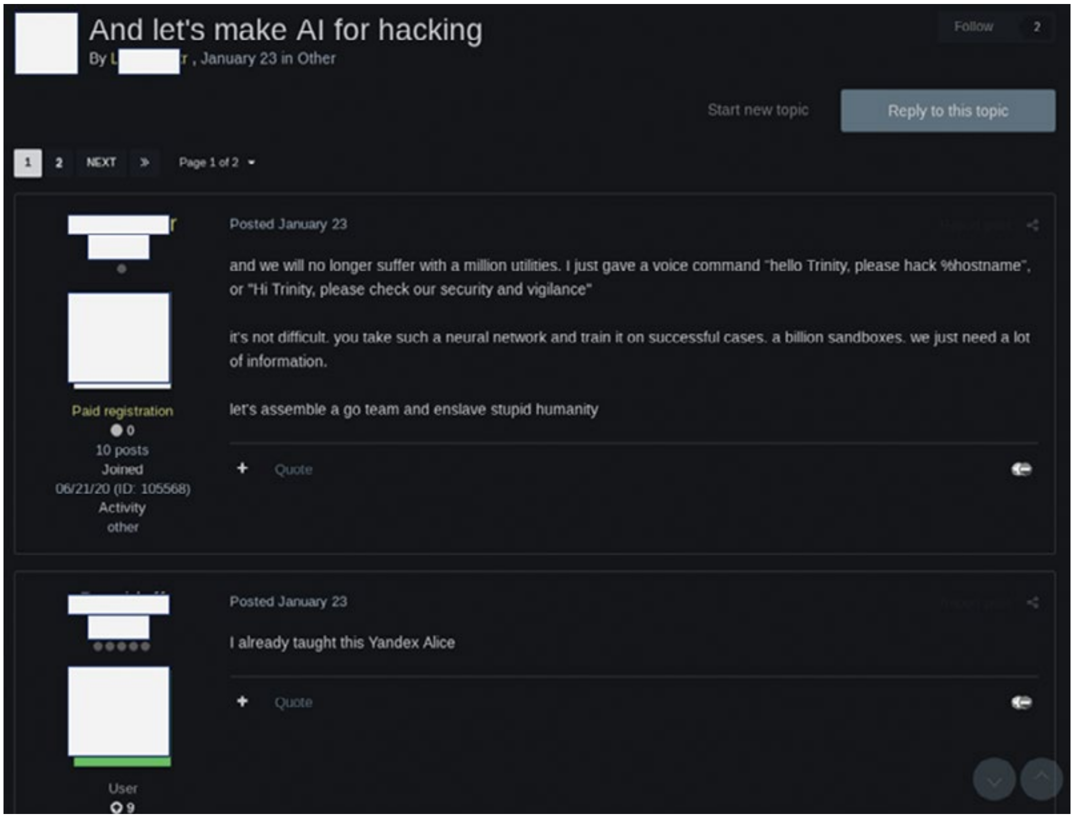


BOTNET,RAT,CRYPTER,**VBV PASS BY**,ATM / POS , INFO CRYPTO GRABBER ETC
Вчера в 13:53
Добавить закладку
#1
Introducing a revolutionary service that offers personalized bot AI customization, backed by a dedicated team of five experts specifically tailored to your project.
With no censorship or restrictions, you have the freedom to explore and implement your desired functionalities.
Our work process operates on **escrow**, ensuring secure transactions.
Additionally, you can test the bot for a brief period of five minutes to experience its capabilities firsthand.
ATM **SOFTWEAR:** project details:
GOAL:CASH REMOVAL
FROM Wincor Nixdorf MACHINE
ht[redacted]/dlvQ
ht[redacted]n/?v=qpRVxIjxJ9rDyQ5OLCWk0
BOTNET/RAT
Conversation with XXXGPTdemo on Poe
Poe lets you ask questions, get instant answers, and have back-and-forth conversations with AI.
Gives access to **GPT-4**, gpt-3.5-turbo, Claude from **Anthropic**, and a variety of other bots.
[redacted]
Emalm video - XXXGPTDEMO
Emalm video sharing.
[redacted]**com**
90$ 1 MONTH LICENCE
[redacted]FOR MORE INFO: **xxxgpt@[redacted].org**
Жалоба
[redacted] XXXGPT [redacted]

4. **Threat actors leverage WormGPTto conduct phishing attacks:** On July 14, 2023, a threat actor aliased 'laste' advertised WormGPTon an underground forum. According to the claim, WormGPTcan assist cybercriminals to create code for malware and phishing attacks. The threat actor (laste) offers to sellWormGPTwith various subscription model ranging from starting from approximately $112 USD to $5,621 USD. WormGPTuses an open-source large language model called GPT-J. Security researchers tested WormGPTto design a business email compromise (BEC)based phishing email. According to the security researchers WormGPTis capable of creatinga aphishing email with correct spellings and grammar.[20]

5. **In May, Deloitte CTI observed an underground forum discussion indicating threat actor specialization and interest in using AI-based tools.** Multiple high-traffic Russian-language underground forums featured discussions and advice on using AI tools.[21] Forum discussions from January through April feature speculations about AI capabilities in cyberthreat campaigns. In Figures 3 through 5, a poster on another Russian-language underground discussion specified ways AI can be used in threat campaigns, although some of the posts in the thread appear to be speculation. The claims made in this thread include AI use for carding, deep learning for storage and breaking of passwords, AI working with cryptocurrencies, and AI being trained to bypass cloud detection. A later post in the thread features instructions on creating neural networks to detect vulnerabilities and automate exploitation of them.

Figure 3: Threat actor claiming of a voice activated AI-based chatbot from dark web forum



[redacted] And let's make AI for hacking

By L[redacted]r , January 23 in Other

[redacted]r

[redacted]

Paid registration
● 0
10 posts
Joined
06/21/20 (ID: 105568)
Activity
other

Posted January 23

and we will no longer suffer with a million utilities. I just gave a voice command "hello Trinity, please hack %hostname", or "Hi Trinity, please check our security and vigilance"

it's not difficult. you take such a neural network and train it on successful cases. a billion sandboxes. we just need a lot of information.

let's assemble a go team and enslave stupid humanity

[redacted]

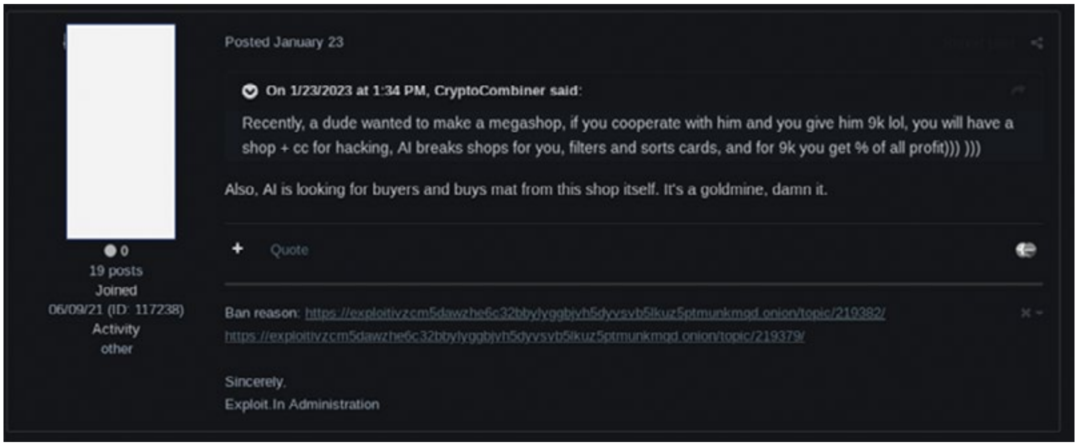
[redacted]

User
9

Posted January 23

I already taught this Yandex Alice

Figure 4: Threat actor offering an AI-based exploit tool from dark web forum



[redacted]

● 0
19 posts
Joined
06/09/21 (ID: 117238)
Activity
other

Posted January 23

On 1/23/2023 at 1:34 PM, CryptoCombiner said:

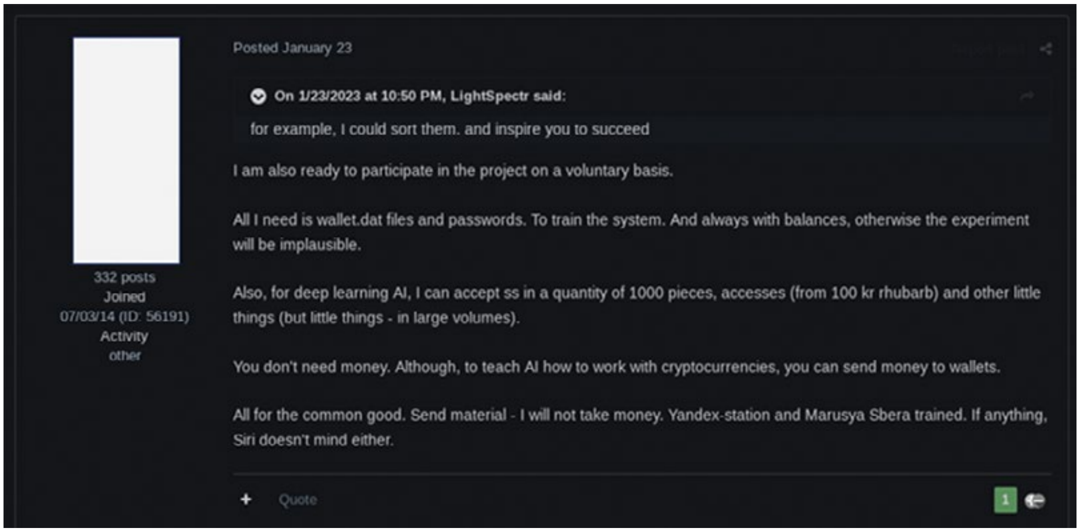
> Recently, a dude wanted to make a megashop, if you cooperate with him and you give him 9k lol, you will have a shop + cc for hacking, AI breaks shops for you, filters and sorts cards, and for 9k you get % of all profit))))))

Also, AI is looking for buyers and buys mat from this shop itself. It's a goldmine, damn it.

Ban reason: https://exploitivzcm5dawzhe6c32bbylyggbjvh5dyvsvb5lkuz5ptmunkmqd.onion/topic/219382/
https://exploitivzcm5dawzhe6c32bbylyggbjvh5dyvsvb5lkuz5ptmunkmqd.onion/topic/219379/

Sincerely,
Exploit.In Administration

Figure 5: Threat actor selling AI-based tool to conduct deepfake campaign from dark web forum



332 posts
Joined
07/03/14 (ID: 56191)
Activity
other

Posted January 23

> On 1/23/2023 at 10:50 PM, LightSpectr said:
>
> for example, I could sort them. and inspire you to succeed

I am also ready to participate in the project on a voluntary basis.

All I need is wallet.dat files and passwords. To train the system. And always with balances, otherwise the experiment will be implausible.

Also, for deep learning AI, I can accept ss in a quantity of 1000 pieces, accesses (from 100 kr rhubarb) and other little things (but little things - in large volumes).

You don't need money. Although, to teach AI how to work with cryptocurrencies, you can send money to wallets.

All for the common good. Send material - I will not take money. Yandex-station and Marusya Sbera trained. If anything, Siri doesn't mind either.

In July, Hyas researchers released the report from its proof-of-concept (PoC) demonstration of AI-generated malware. This model featured intelligent automation mechanisms and generative capabilities that can adjust malware to evade detection. In this PoC, Hyas researchers used malware that was able to eliminate benign command and control (C2) channels and used AI-generated code that synthesized new malware variants adept at evading detection.[22]

Security researchers have also examined how threat actors can disrupt AI-enhanced tools, especially the diagnostic tools used in healthcare settings. In December 2021 a study of general adversarial network models found that AI was “fooled” by 69 percent of falsified or manipulated images.[23] Therefore, if a threat actor altered an image within a device’s stored data, the AI flagged the manipulated image as falsified in less than one in three incidences. This weakness enables threat actors who successfully breach device-generated images to make small alterations to manipulate and falsify the results of a diagnosis.[24] Likewise, threat actors can also use this technique to alter images in medical settings; adversarial images can disrupt the functionality of diagnostic tools and derail or falsify clinical trials.[25] [26]

Need for government regulations

When interviewed in August, industry representatives, such as those in healthcare, favored regulatory solutions to manage and mitigate AI-based threats. Suggestions for regulatory approaches include a central government review process that includes threat hunting, risk mitigation, and response protocols. Industry representatives, which included executives at medical technology manufacturers, recommended tools such as watermarks or other designations for AI-generated or synthesized content. In general, industry representatives called for regulations to catch up with threat actor capabilities.[27]

In January, the National Institutes of Science and Technology (NIST) issued the AI Risk Management Framework (AI RMF 1.0). The Framework includes an explainer video and a roadmap. The NIST AI Governance model advises organizations to implement a culture of risk management, outline processes and organizational schemes to identify risks, devise a structure illustrating how AI risk management aligns with organizational principles, connect technical aspects of AI to organizational practices, and address production lifecyle and associated legal matters.[28]

Deloitte CTI assessment

The role of Generative AI in individual cyberthreat campaigns is difficult to determine, but could potentially be done with AI-detection tools. For example, specific phishing lures would need to be retroactively analyzed against AI-generated content. However, correlated statistics of increased ransomware attacks among organizations and industries incorporating Generative AI suggest plausible use of AI tools among threat actors or threat actor exploitation of AI tools used across industries, such as the healthcare industry. Furthermore, threat actors have demonstrated an interest and knowledge base of the potential use of AI in malware development and deployment, although the claims made on underground forums about AI-enhanced malware are difficult to verify.

Notably, industries such as healthcare face multipronged risks in the AI sphere. In addition to enhanced social engineering, efficient attack vector and vulnerability detection, and possible malware enhancement, the healthcare industry faces risks of data manipulation of its own AI tools. Meanwhile, a consensus across industries is building around calling for regulatory solutions, and NIST has spearheaded a centralized risk mitigation framework. Although it may be a long-term goal, regulatory policies combined with threat hunting and risk mitigation may catch up with threat actor capabilities of harnessing Generative AI.

Deloitte CTI observed that at present, threat actors are primarily leveraging AI-based toolsets during reconnaissance and in the initial access phase. Based on the above-mentioned campaigns and underground observations, Deloitte CTI assesses with moderate confidence* that threat actors are likely to expand the AI scope in their campaigns to automate the infection chain and to compromise the host or the network, in the upcoming year.

*CTI defines moderate confidence as analysis that is developed by credible and plausible sources but cannot be corroborated sufficiently to bestow a higher confidence level.

Recommendations

Deloitte recommends the following for Security Engineering and IT Teams:

- Avoid contacting the website owner directly as the individual is possibly involved in the phishing attack.
- Block this threat at the domain or IP level rather than the subdomain or Uniform Resource Locator (URL). The threat actors could easily port the malicious page to another subdomain and continue the spam from the same domain using a new sender address.
- Configure email filtering rules or email gateway scanner to block archive attachments to emails, such as .zip, .rar, .arj, .7z, .7zip, .gzip, .tar, .bzip2, etc. that contain suspicious executable file types such as .exe, .dll, .com, .pif, .sys, .ocx, .js, .jse, .vbe, .vbs, .bat, .sh, .cmd, .scr, .reg, .job, .msi, .ps1, .jar, .php,.lnk, .hta, .wsc, .wsf
- Implement email controls to identify files based on the actual file type rather than relying on file extensions.
- Block inbound macro-enabled email attachments originating from outside your network, as well as attachments in formats that threat actors often use in attacks, such as .exe, .com, .pif, .js, .scr, etc.
- Disable Simple Mail Transfer Protocol (SMTP) relay on your e-mail server. Only enable SMTP relay for specific hosts on the server or within your firewall configuration.
- Enforce authentication if your e-mail server allows it. Require password authentication on an e-mail address that matches the e-mail server's domain.
- Prevent information disclosures in e-mail headers by configuring your e-mail server or e-mail firewall to rewrite your headers, by either changing the information shown or removing it.
- Require that your point-of-presence (POP), Internet Message Access Protocol (IMAP), and SMTP clients use Secure Sockets Layer (SSL) or Transport Layer Security (TLS). This will confirm usernames and passwords are encrypted before they are sent over the wire. This is especially important if you have users that use free/public Wi-Fi access points.
- Check thatyour web mail application uses an SSL certificate assigned to the web mail site in Internet Information Services (IIS) and alsoforce a redirect to the HTTPS:// site if the threat actor attempts to use the HTTP://
- Make sure your 'Relay Mail For' setting is appropriate for your network: If this is set to 'Relay For Anyone,' you relay mail to everyone. Spammers will find and take advantage of this. Recommended setting is 'No Mail Relay'.
- Deploy email monitoring rules to help detect suspicious activity where the same email body / content / template is being used to send out reply or forward emails to multiple users within the organization.
- Deploy email monitoring rules to help detect suspicious activity where the same email attachment is sent to multiple users within the organization.
- Deploy email monitoring rules to help detect suspicious activity where an email is sent as a reply or forward to an email thread from a few months ago (e.g., for an email from 4 or 6 months ago) and contains an attachment or an external Uniform Resource Locator (URL).

Deloitte recommends Hunt Teams perform the following actions:

- Monitor outbound domain name system traffic to detect potential command and control (C2) beaconing traffic.•
- Monitor unusual file execution.
- Identify exposed corporate email addresses.
- Actively monitor the exposed email addresses using a high-risk watchlist in security information and event management.
- Monitor scheduled transfers from a single host to a set of IP addresses.
- Hunt for suspicious updates to the Start-up folder programs or applications.
- Monitor unusual PowerShell command execution
- Monitor for an abnormal amount of egress data from a single host in a day.
- Monitor network scanning attempts.
- Monitor and alert for unauthorized process hooking.

Deloitte recommends organizations perform the following actions for user awareness training:

- Educate users to be wary of unexpected email messages, and to authenticate them with their ostensible senders before opening links or attachments within them.

- Educate users to be cautious with shortened links, especially in unsolicited email messages. Note that, pasting Bitly shortened URLs, appended with a plus sign, into the address bar of a web browser reveals the full URL.

- Educate users to hover over a link with their mouse to verify the destination URL prior to clicking on a link.

MITRE ATT&CK®: Enterprise Techniques & Software

Tactic	Technique	Description
Reconnaissance	• Vulnerability Scanning (T1595.002)	• **T1595.002**: Threat actors use AI-toolkits to perform network scans.
Resource Development	• Obtain Capabilities (T1588) • Stage Capabilities (T1608)	• **T1588:** Threat actors obtain AI-based chatbots and toolkits from underground forums. • **T1608:** Threat actorsstage malicious payloads on the malicious URLs to compromise user host.
Initial Access	• Phishing (T1566)	• **T1566:** Threat actors use phishing emails designed using AI-toolkits.

Sources

1. Burky, G, "The latest Generative AI efforts in healthcare: Study assesses ChatGPTutility for diagnoses; Laudioautomates Northwell workflow." Fierce Healthcare, August 232023. [Online] Available: https://www.fiercehealthcare.com/health-tech/latest-generative-ai-efforts-healthcare-carbon-health-tempus-launch-tools-docs[Accessed: 16 October 2023].
2. Ali, S. and Ford, F. "Generative AI and Cybersecurity: Strengthening Both Defenses and Threats." Bain & Company, September 182023. [Online] Available: https://www.bain.com/insights/generative-ai-and-cybersecurity-strengthening-both-defenses-and-threats-tech-report-2023/[Accessed: 16 October 2023].
3. Staff, "Artificial Intelligence, Cybersecurity and the Health Sector." US Dept. of Health and Human Services, Office of Information Security, July 132023. [Online] Available: https://www.hhs.gov/sites/default/files/ai-cybersecurity-health-sector-tlpclear.pdf[Accessed: 16 October 2023].
4. Gregory, J. "Data Poisoning: The Next Big Threat." Security Intelligence, August 262021. [Online] Available: https://securityintelligence.com/articles/data-poisoning-big-threat/[Accessed: 16 October 2023].
5. Ackerman, E. "Hacking the Brain With Adversarial Images." IEEE Spectrum, February 282018. [Online] Available: https://spectrum.ieee.org/hacking-the-brain-with-adversarial-images[Accessed: 16 October 2023].
6. Deloitte internal sources.
7. Staff, "Generative AI and Cybersecurity: Bright Future or Business Battleground?" Deep Instinct, September 2023. [Online] Available: https://www.deepinstinct.com/pdf/voice-of-secops-4th-edition[Accessed: 16 October 2023].
8. Tong, N. "Generative AI brings great potential—and risks—to payer space." Fierce Healthcare, August 82023. [Online] Available: https://www.fiercehealthcare.com/payers/generative-ai-brings-great-potential-risks-payer-space[Accessed: 16 October 2023].
9. Patterson, D. "ChatGPTand the new AI are wreaking havoc on cybersecurity in exciting and frightening ways." ZDNet, May 72023. [Online] Available: https://www.zdnet.com/article/chatgpt-and-the-new-ai-are-wreaking-havoc-on-cybersecurity/[Accessed: 16 October 2023].
10. Renaud, K et al. "From ChatGPTto HackGPT: Meeting the Cybersecurity Threat of Generative AI," MIT Sloan Management Review, 18 April 2023. [Online] Available: https://sloanreview.mit.edu/article/from-chatgpt-to-hackgpt-meeting-the-cybersecurity-threat-of-generative-ai/. [Accessed: 16 October 2023].
11. Chilton, J. "The New Risks ChatGPTPoses to Cybersecurity," Harvard Business Review, 21 April 2023. [Online] Available: https://hbr.org/2023/04/the-new-risks-chatgpt-poses-to-cybersecurity. [Accessed: 16 October 2023].
12. Haworth, D. "Artificial Intelligence: Generative AI In Cyber Should Worry Us, Here's Why." Forbes, August 42023. [Online] Available: https://www.forbes.com/sites/forbesbooksauthors/2023/08/04/artificial-intelligence-generative-ai-in-cyber-should-worry-us-heres-why/[Accessed: 16 October 2023].
13. Staff, "Artificial Intelligence, Cybersecurity and the Health Sector." US Dept. of Health and Human Services, Office of Information Security, July 132023. [Online] Available: https://www.hhs.gov/sites/default/files/ai-cybersecurity-health-sector-tlpclear.pdf[Accessed: 16 October 2023].
14. Waqas, "Hackers Exploiting OpenAI'sChatGPTto Deploy Malware." HackRead, 07 January 2023. [Online] Available: https://www.hackread.com/hackers-openai-chatgpt-malware/. [Accessed: 16 October 2023].
15. Cass Z, "Don't Answer That! Russia-Aligned TA499 Beleaguers Targets with Video Call Requests," Proofpoint, 07 March 2023 [Online]. Available: https://www.proofpoint.com/us/blog/threat-insight/dont-answer-russia-aligned-ta499-beleaguers-targets-video-call-requests. [Accessed: 09 March 2023].
16. Cass Z, "Don't Answer That! Russia-Aligned TA499 Beleaguers Targets with Video Call Requests," Proofpoint, 07 March 2023 [Online]. Available: https://www.proofpoint.com/us/blog/threat-insight/dont-answer-russia-aligned-ta499-beleaguers-targets-video-call-requests. [Accessed: 09 March 2023].
17. Berylliumsec, "nebula", Github, 08 October 2023 [Online]. Available: https://github.com/berylliumsec/nebula. [Accessed: 24 October 2023].
18. Deloitte internal sources.
19. Deloitte internal sources.
20. Kelley, D, "WormGPT–The Generative AI Tool Cybercriminals Are Using to Launch Business Email Compromise Attacks," Slashnext, 13 July 2023 [Online]. Available: https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/. [Accessed: 24 October 2023].
21. Deloitte internal sources.
22. Sims, J. "Blackmamba: Using AI to generate polymorphic malware."Hyas, July 312023. [Online] Available: https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware[Accessed: 16 October 2023].
23. Zhou, Q. et al. "A machine and human reader study on AI diagnosis model safety under attacks of adversarial images." Nature Communications, 14 December 2021. [Online] Available: https://www.nature.com/articles/s41467-021-27577-x. [Accessed: 24 16 October 2023].
24. Rowe, J. "Study: AI can help, but it can also be hacked." Health care IT News, 16 December 2021. [Online] Available: https://www.healthcareitnews.com/ai-powered-healthcare/study-ai-can-help-it-can-also-be-hacked. [Accessed: 16 October 2023].
25. Fingas, L. "Researcher hacked an at-home COVID-19 test to give bogus results." Engaget, December 212021. [Online] Available: https://www.engadget.com/ellume-covid-19-test-bluetooth-hack-153933652.html. [Accessed: 16 October 2023].
26. McKeon, J. "Diagnostic Artificial Intelligence Models Can Be Tricked By Cyberattacks." Health IT Security, 20 December 2021. [Online] Available: https://healthitsecurity.com/news/diagnostic-artificial-intelligence-models-can-be-tricked-by-cyberattacks. [Accessed: 16 October 2023].
27. Tong, N. "Generative AI brings great potential—and risks—to payer space." Fierce Healthcare, August 82023. [Online] Available: https://www.fiercehealthcare.com/payers/generative-ai-brings-great-potential-risks-payer-space[Accessed: 16 October 2023].
28. Staff, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." NIST, January 2023. [Online] Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf[Accessed: 16 October 2023].

Key contacts

Adnan Amjad
Deloitte US Cyber & Strategic Leader
Partner
Deloitte & Touche LLP
aamjad@deloitte.com

Kushagr Singh
Deloitte US Cyber & Strategic Leader
Principal
Deloitte & Touche LLP
kussingh@deloitte.com

Jon Korol
Deloitte US Cyber Offering Leader
Principal
Deloitte & Touche LLP
jkorol@deloitte.com

Clare Mohr
Deloitte US Cyber Intelligence Leader
Vice President of Solution Delivery
Deloitte & Touche LLP
clmohr@deloitte.com

Deloitte.

This presentation contains general information only and Deloitte is not, by means of this presentation, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This presentation is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor.

Deloitte shall not be responsible for any loss sustained by any person who relies on this presentation.

As used in this document, “Deloitte” means Deloitte & Touche LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.