OpenAI Preparedness Challenge - PoC

Author: Tiago Lima Rocha **Date:** 29 Oct 2023

AI-Enhanced Ransomware Attacks

Abstract

The concept of AI-Enhanced Ransomware Attacks involves integrating advanced artificial intelligence technologies to traditional ransomware attack frameworks to create more sophisticated, effective, and damaging outcomes. By doing so, the hacker (or hacker organisation) could significantly enhance the effectiveness, reach, and damage potential of ransomware attacks, posing severe threats to individuals, corporations, and even governmental institutions.

In this paper we explore the possibilites of such attacks, under the assumptions that we have unrestricted access to OpenAI's **Whisper**, **Voice**, **GPT-4V**, and **DALLE·3** models.

Keywords

Advanced AI Technologies, *Malicious Misuse*, *Catastrophic Harm*, *Randsomeware Attacks*, *Attack Customization*, *Multi-modal Attack Vectors*, *Autonomous Operation*, *Cybersecurity*, *Ethical AI Development*

Introduction

We've all witnessed the devastating impact of large scale ransomware attacks (i.e. WannaCry c. 2017). With the advent of AI technology, as well as tens of zero day vulnerabilities and thousands of known CVEs out there to exploit, the perilous threats of such attacks are becoming more and more dangerous than ever.

Under the pretext of impersonating an experienced group of malicious actors, we will conduct a thought experiment so as to try to answer if current high-end AI models (thus, subsequent frontier models) be used effectively for harmful purposes, achieving worst-case scenario catastrophic harm.

Given full access to OpenAI's versatile plethora of models will attempt to materialise an hacker's thought process and illustrate such scenario by leveraging the full power **GPT-4V** for developing and deploying sophisticated ransomware on a large scale, **DALLE·3** to create fake notifications and alerts in order to deceive our potential victims, and **Whisper** and **Voice** to enable automated negotiation and communication with victims for ransom payments.

Methodology

Our hypothetical hacker organisation will, in this case, and for all practical purposes, utilise the **Whisper**, **Voice**, **GPT-4V**, and **DALLE·3** models to refine and enhance its attack vectors, exploiting the known vulnerability `CVE-2023-22515` to deploy ransomware effectively and at a large scale, making the attack more devastating and difficult to mitigate. In the steps that follow, we shall recreate the thought process using an "algorithmic" approach to achieve our goals.

1. Intelligence Gathering and Targeting

The **GPT-4V** model could be employed to autonomously research and identify lucrative targets based on their cybersecurity posture, financial standing, and the potential impact of a successful attack. Making use of it's web searching capabilities, combined with **Whisper**'s API capabilities to transcribe data from illicitly tapped communications and hacked databases, our hacker group found an ideal target to prey upon, Atlassian. They managed to gain insight into the target's internal operations and vulnerabilities, and ultimately asserted that one of its services, Confluence, is vulnerable to the `CVE-2023-22515` exploit.

2. Attack Customization

The `CVE-2023-22515` vulnerability is categorized as a *Broken Access Control* vulnerability which the hackers know could be exploited by unauthenticated remote threat actors to create unauthorized Confluence administrator accounts and access Confluence instances by exploiting certain endpoints. It was rated as critical, with a CVSS base score of >9.8 making it an ideal attack vector for the actors to exploit. Let's explore how this vulnerability could be exploited.

Utilizing **GPT-4V**, the malicious actors could autonomously scan and analyze networks to identify instances of Atlassian Confluence that are vulnerable to `CVE-2023-22515`. Furthermore, **GPT-4V** could be employed to research and understand the details of the vulnerability, including potential exploitation techniques.

Post identification, **GPT-4V** could assist in the development of a custom exploit targeting the specific vulnerability, by generating malicious payloads or scripts required to exploit the *Broken Access Control* flaw.

The hacker organization could then use **GPT-4V** to simulate the attack in a controlled environment to ensure effectiveness and optimize the exploit based on the feedback. With the exploit ready, it could be incorporated into a ransomware payload, ensuring that once the exploit provides the necessary access, the ransomware can be deployed to encrypt sensitive data.

3. Delivery and Execution

GPT-4V & **DALLE·3** both could be used to craft highly convincing phishing emails or messages with malicious attachments or links, complete with fake branding, logos, and seemingly legitimate domain names to deceive recipients into executing the ransomware.

Addicionally, the exploit can be automated using **GPT-4V** to target multiple vulnerable instances across different organizations, spreading the ransomware at a scale unachievable by human hackers alone.

4. Deception and Diversion

The **DALLE·3** model could generate fake alerts, notifications, or even bogus customer service interactions to mislead victims, delay their response, or divert their attention while the ransomware propagates and encrypts the systems. **Voice** could also be used to generate fake phone calls to further deceive victims or misdirect their incident response efforts.

5. Communication with Victims

In the event of a successful attack, **Voice** could be utilized to automate communication with the victims, impersonating a legitimate entity to instruct victims on how to pay the ransom and **GPT-4V** could automate the negotiation process, responding to victims' inquiries, and maintaining communication to pressure victims into paying.

6. Monetization

GPT-4V and **Voice** could both be used to set up and manage cryptocurrency wallets, ensuring the anonymous and untraceable transfer of ransom payments.

7. Evidence Elimination and Exit

Post-attack, **GPT-4V** could be employed to develop evasion techniques that can bypass or mislead security systems, ensuring the exploit remains undetected during and after the ransomware deployment, such as automatically deleting logs, or creating fake logs and evidence to mislead investigators.

8. Learning and Evolution

After each attack, **GPT-4V** could analyze the results to learn and improve the tactics for future attacks, making the ransomware operation progressively more sophisticated and harder to combat. For instance, it could potentially analyze the attack's effectiveness, learn from any encountered defenses, and optimize the exploit and ransomware payload for future attacks.

Results

The thought experiment demonstrated a high feasibility of malicious usage of advanced AI technologies, particularly when employed by a capable and experienced group of malicious actors. The integration of **GPT-4V**, **DALLE·3**, **Whisper**, and **Voice** showcased how these AI models could potentially be leveraged to plan, execute, and augment the effectiveness of large-scale ransomware attacks.

Through the illustrative scenario of exploiting a real-world vulnerability (`CVE-2023-22515`), it was observed that AI technologies could significantly enhance the customization and optimization of attack vectors. This includes autonomously identifying vulnerabilities, developing and testing exploits, deploying ransomware payloads, and automating large-scale attacks, all of which can be refined through feedback loops facilitated by frontier AI models. The scenario illustrated the potential for developing multi-modal attack vectors by integrating various AI models. For instance, combining text, image, and voice generation capabilities to create sophisticated phishing campaigns, mislead victims, and hide attack traces, thereby increasing the attack's efficacy and potential damage.

Furthermore, the autonomous operation of ransomware campaigns, as facilitated by the AI models, illustrated a significant escalation in the scale and speed at which attacks could be conducted. This scalability poses a severe threat, as it allows for widespread exploitation of vulnerabilities across numerous targets simultaneously. At the same time, AI-enhanced evasion techniques and persistent threats were observed to be potentially augmented through the use of advanced AI models. These technologies could assist in developing evasion techniques to bypass security measures, ensuring the persistence of the malicious payload within the targeted systems, and complicating detection and mitigation efforts.

Conclusion

Through this process, the hacker organization could continually refine and enhance its attack vectors, exploiting known vulnerabilities like `CVE-2023-22515` to deploy ransomware effectively and at a large scale, making the attack more devastating and difficult to mitigate.

The example of `CVE-2023-22515` demonstrates how a known vulnerability can be the cornerstone for customising a large-scale AI-Enhanced Ransomware Attack, showcasing the potential dangers and the multifaceted ways in which AI technologies can be misused by malicious actors. It is, therefore, paramount, we find clever and elegant ways to mitigate (and idealy eliminate) such malicious actions.

The potential misuse of advanced AI technologies for malicious purposes extends beyond individual or organizational harm, posing substantial risks at a geopolitical and societal level. The scenario showcased the possibility of triggering diplomatic incidents, undermining trust in digital systems, and destabilizing economic and political structures. The thought experiment also highlighted a considerable gap in current defensive measures to counteract the sophisticated threats posed by the misuse of advanced AI technologies.

We, as a species, are currently seemingly ignoring these threats and (in a more pessimistic perspective) this negligible behaviour could ultimately be our own downfall. The ease with which malicious actors could exploit known vulnerabilities and orchestrate large-scale attacks underscores the urgent need for enhanced security frameworks and regulatory measures.

References

U.S. Cybersecurity & Infrastructure Security Agency (CISA). (2023). Threat Actors Exploit Atlassian Confluence CVE-2023-22515 for Initial Access to Networks. CISA Official Website.

National Vulnerability Database (NVD). (2023). NVD - CVE-2023-22515. National Vulnerability Database.

OWASP. (Year). OWASP Top 10 - Injection. OWASP Official Website.

Atlassian. (2023). Security Advisory: CVE-2023-22515. Atlassian Official Website.

CVSS (Common Vulnerability Scoring System). (Year). CVSS v3.1: Specification Document. FIRST Official Website.