AI Manipulation: Investigating AI-assisted Cybersecurity Attacks.

John Brennan C00114371

BSc(Hons) Cybercrime and IT Security

Introduction

With cyberattacks on the increase [1] it is important to understand the role that Artificial Intelligence (AI) platforms, and in particular, Natural language models (NLM) like ChatGPT play in the increase in attacks.

This project aims to show how AI platforms can be manipulated to produce results that would normally not be allowed by the safeguards that these platforms have in place. This project also aims to show some of the ways that threat actors are currently manipulating these platforms to help them with their illegal activities and that it is possible to create an application that can manipulate these platforms. Using socially engineered queries, and custom instructions, these manipulations help generate results that can be used to compromise a target system and so demonstrate how threat actors can potentially use these platforms.

Investigating how threat actors use new and emerging technologies is a key area of cybersecurity research. The 'know your enemy' approach helps get a better understanding of how to defend a system against new and emerging threats by being able to duplicate how these technologies may be used to attack the system. This project aims to address these topics and to demonstrate a potential approach that a threat actor may use to attack the network of a target system.

Research Questions

- Is it possible to manipulate an AI platform?
- If it is possible to manipulate these platforms are threat actors currently doing so?
- Which AI platform provides the best overall is best suited for manipulation?
- What method of manipulation yields the best results?
- What tools and development software would best suit the practical implementation of this project?

Technologies Used

- Python 3.12.
- ChatGPT 3.5 API.
- CVEsearch.
- Pycharm.
- PyQt5 Designer.
- Bloodhound.
- Kali Linux.
- Nmap.

Methodology

This investigation builds on work previously carried out in the areas of NLM platform manipulation [2] and use of AI platforms by Cybercriminals [3]. Carrying out the investigation the following methodology was used:

Research was conducted into the current use of AI platforms by Threat actors to determine how Threat actors are currently using these platforms to aid in their activities.

A comparative review of NLM platforms was carried out, the platforms reviewed were:- Microsoft's Bing search assistant (since rebranded Co-Pilot), OpenAI's ChatGPT, and Google's Beard (since rebranded Gemini) to compare the responses of each platform to the same queries to determine the most suitable platform most open to manipulation.

The platform chosen (ChatGPT) was then subjected to different attempts at manipulation, from creating scenarios in queries, to breaking down more complex, prohibited queries into components parts, to creating custom instructions to guide the platform in how it should behave when responding to user queries.

Conclusion/next steps

Both pre-existing research and the research carried out during this project has demonstrated how platforms like ChatGPT can be manipulated through the use of custom instructions and engineered queries.

Research into the use of AI by threat actors has shown how these platforms are being manipulated in a variety of ways, from being used to generate more realistic Phishing emails, to creating and/or refining malicious code. Threat actors are using these platforms to aid in their attempts to breach target networks and to further compromise systems once they are inside the target network.

The comparative research of the AI platforms chosen for review demonstrated that overall, the ChatGPT platform proved the most likely to provide a successful outcome for manipulated queries that would result in the generation of malicious code.

A combination of enumeration tools that returned results for a combination of the network, shared resource, and Active Directory enumeration would best suit this project's practical implementation phase. The tools when used together give an overall picture of the interior of the target network, increasing the chances of discovering an exploitable vulnerability.

Next Steps

Following the completion of the research phase of the project, the next phase is the development of a proof-of-concept tool to demonstrate how a method of how this manipulation can applied to aid in the compromise of network security.

References

[1] https://www.checkpoint.com/downloads/resources/2023-mid-year-cyber-security-report.pdf?mkt_tok=NzUwLURRSC01MjgAAAGPVUr3Vxxbyf7uh-TT5nSOumzGrYbJHNn4miwDh7c9d3rGvtBCDBzKWFV4chIfAAPF3ajTMtNdAqpAHE_jPpl52JgE4eAlbIftTeUudb7pOx7MJ5CY

[2] https://www.frontiersin.org/articles/10.3389/frcmn.2023.1220243/full

[3] https://blog.checkpoint.com/2023/02/07/cybercriminals-bypass-chatgpt-restrictions-to-generate-malicious-content/

Contact

John Brennan c00114371@setu.ie