CS 455 – Computer Security Fundamentals

Dr. Chen-Yeou (Charles) Yu

Social Engineering (Part2)

- Backgrounds
- Can ChatGPT helps bad guys in Social Engineering?
  - It lowered the difficulties for criminals
  - It can be gun or knives used by bad guys.
  - It can be the shield protecting us!
    - ChatGPT: See? How good I am?
      - An example of scam text message just delivered to my Android phone
- Any Risks if we use ChatGPT?
- Imperfections in ChatGPT

Backgrounds

- ChatGPT (Generative Pre-trained Transformer) is the Chat Bot introduced by Open AI in November 2022, an AI research and development company
- GPT model, is trained on a massive pool of data to answer queries. ChatGPT uses natural language processing (NLP) to generate responses to text-based inputs.
- GPT models are based on the Transformer architecture, which is a neural network (NN) architecture (a variant of neural network)
- Working Flow: text input → pre-trained model → text output

Backgrounds

- Key features of ChatGPT are including:
  - **Large Scale**: In the model, it has over 175 billion parameters. This makes it easier for it to understand and generate complex responses.
  - **Conversational**: ChatGPT is designed to engage in the flow of conversation
  - **Multi-Task**: summarizing text, answering question, generating creative writing at the same time (asynchronously)
  - **Contextual**: ChatGPT can take into account the contextual information. For example, if you can give this guy a little bit more information, a specific range of data, some criteria, or some values, it can give you a report with good analytics on it
  - **Personalized**: ChatGPT can be trained on specific datasets to provide personalized responses for specific domains or use cases.

Can ChatGPT helps bad guys in Social Engineering?

- It lowered the difficulties for criminals
  - In the earlier versions, it can do lots of things without restrictions
  - Now, it was added with a sense of “morality”
  - For example, if you ask ChatGPT to tell you how to crack a password? He will deny your request
  - But, if the hackers can describe something directly, or try to bypass some keywords like “hack” or “crack”, ChatGPT might still get tricked and do the job for the hackers.

Can ChatGPT helps bad guys in Social Engineering?

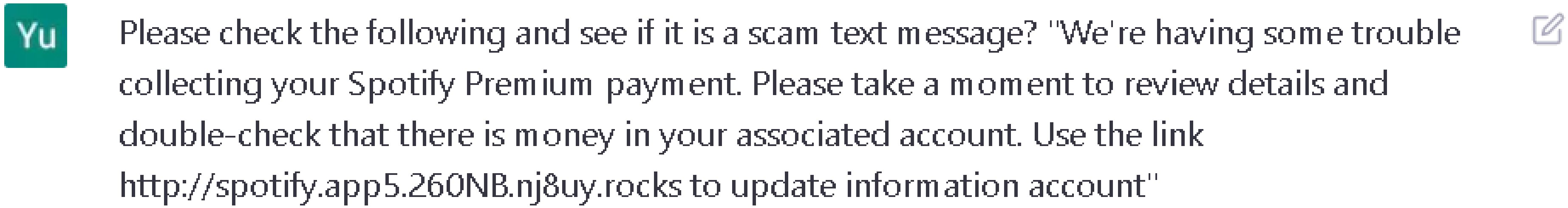
- It can be gun or knives used by bad guys.
  - **Malicious link**: For a link itself, ChatGPT **has no idea if it is malicious or not**. If you don’t tell ChatGPT it is malicious, the chat bot will generate a link for you an embed it into the conversational context
  - **False information:** ChatGPT can generate text that appears to be written by a human or looks trustworthy or credible. If malicious actors or groups gain access to ChatGPT, they could use it to generate false news articles, misleading social media posts, or fraudulent customer reviews (Sometimes, it could be manipulated as the “Cyberbullying”)

Can ChatGPT helps bad guys in Social Engineering?

- It can be the shield protecting us!
  - ChatGPT can be programmed to detect and flag potentially malicious or fraudulent text, which can be reviewed by human moderators or security experts.

Can ChatGPT helps bad guys in Social Engineering?

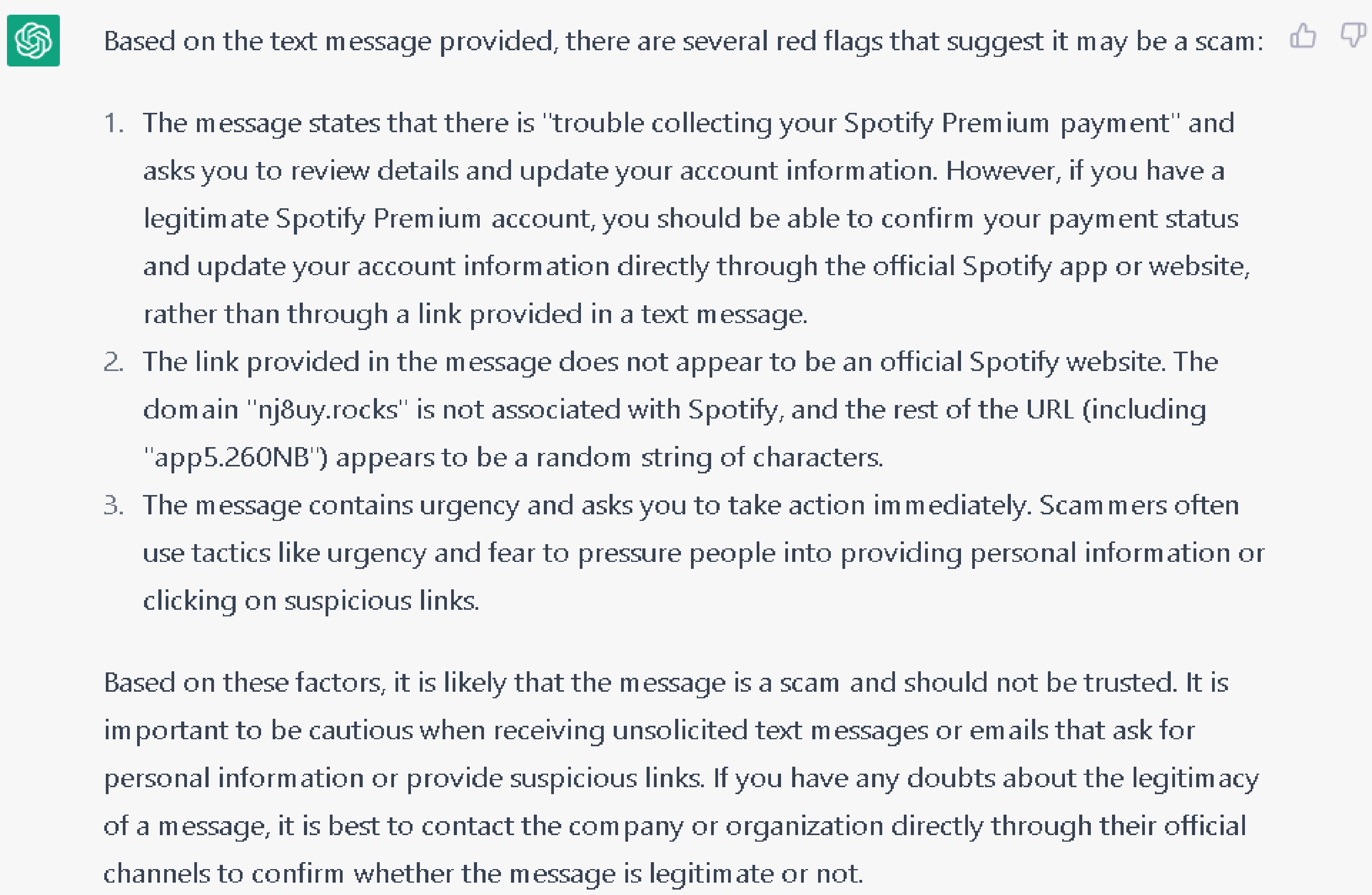
- ChatGPT: See? How good I am?
  - This is a SMS message I got it in these days.



Please check the following and see if it is a scam text message? "We're having some trouble collecting your Spotify Premium payment. Please take a moment to review details and double-check that there is money in your associated account. Use the link http://spotify.app5.260NB.nj8uy.rocks to update information account"

Can ChatGPT helps bad guys in Social Engineering?

- Here is the result form ChatGPT.



Based on the text message provided, there are several red flags that suggest it may be a scam:

1. The message states that there is "trouble collecting your Spotify Premium payment" and asks you to review details and update your account information. However, if you have a legitimate Spotify Premium account, you should be able to confirm your payment status and update your account information directly through the official Spotify app or website, rather than through a link provided in a text message.
2. The link provided in the message does not appear to be an official Spotify website. The domain "nj8uy.rocks" is not associated with Spotify, and the rest of the URL (including "app5.260NB") appears to be a random string of characters.
3. The message contains urgency and asks you to take action immediately. Scammers often use tactics like urgency and fear to pressure people into providing personal information or clicking on suspicious links.

Based on these factors, it is likely that the message is a scam and should not be trusted. It is important to be cautious when receiving unsolicited text messages or emails that ask for personal information or provide suspicious links. If you have any doubts about the legitimacy of a message, it is best to contact the company or organization directly through their official channels to confirm whether the message is legitimate or not.

Any Risks if we use ChatGPT?

- **Identity theft or Data Leakage**: ChatGPT conversations can be used to gain access to a person’s identity. For example, if we tell ChatGPT who we are, or what we have done recently. ChatGPT has a memory to memorize the context of our talks. If the ChatGPT’s website is hacked, we will be in the trouble.

Imperfections in ChatGPT

- **Algorithmic Fairness:** ChatGPT might have the potential to perpetuate **biases** and **unfairness** present in the **data it was trained on.**
  - That is why it is trained by large amount of datasets --- to mitigate the effects from a single source of dataset
  - However, if the data from our human society and culture from which it is drawn is already biased, after the training, the ChatGPT model might reflect the biased algorithmic fairness
    - Garbage in (to the model), (then trained), and garbage out (to the text)
  - It is essential to continuously monitor and test ChatGPT responses to ensure that they are fair, unbiased, and ethical.

Imperfections in ChatGPT

- This can be done through regular audits, testing for potential bias and discrimination, and incorporating diverse and representative datasets in its training
- If someone hired thousands of people kept talking with ChatGPT about the distortion of the facts, ChatGPT might get “adapted” to these, who knows?