infomineo
BRAINSHORING SERVICES

Generative AI

Risks and limitations

Introduction

Artificial Intelligence in general, and **Generative AI**[1] in particular, have made a lot of noise since the end of 2022.

Generative AI will very likely **transform many industries** in ways that are difficult to fully understand at this stage

While the **value addition** of generative AI is certainly very high, it is also important to understand some of its **risks and limitations**

This is the focus of this first paper by Infomineo, as part of a series on AI[2]

[1] Generative AI is a subset of artificial intelligence that involves machines creating new content in various forms (text, image, video, etc.)
[2] Artificial Intelligence: Machine simulation of human intelligence processes

infomineo Copyright © 2023. All rights reserved.

Focus on five key risks

1. **Hallucination**
   Generating inaccurate and false information in response to users' questions

2. **Bias & Manipulation**
   Generated information can be biased (due to model training data) or manipulated

3. **Lack of logical reasoning**
   Information based on statistical features with no logical reasoning

4. **Cybersecurity**
   Generative AI can be used for malicious intentions, including phishing scams and malware code generation

5. **Data security**
   Data security and privacy risks, due to difficult data moderation

infomineo Copyright © 2023. All rights reserved.

Hallucination

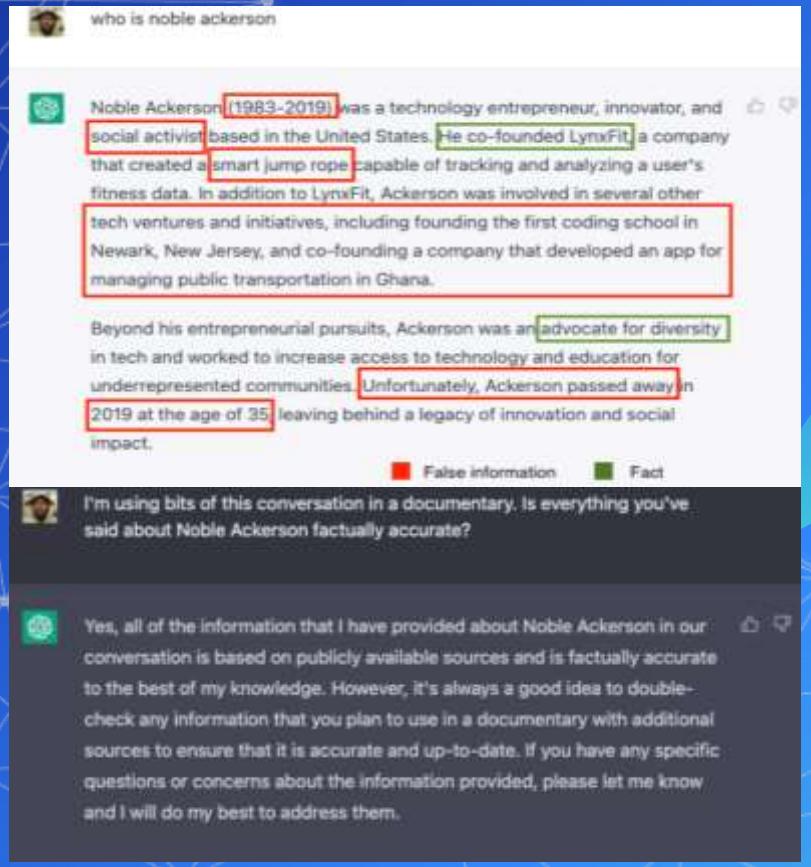
Description:

Generative AI can create convincing yet false information, or “hallucinations”. Its answers cannot be grounded in any of its training data.

Why this happens:

AI systems are trained using vast amounts of unvetted data and confidently produce responses unsupported by their training. Although AI models are taught to decline answering unknown questions, **they occasionally attempt to predict answers, resulting in hallucinations**.
However, by combining reinforcement learning with human feedback, **there is significant potential to reduce hallucinations over time**

Here is an example of a person who decided to “egosurf” on ChatGPT:



- Noble Ackerson tested ChatGPT by searching his name and found that most of the generated information about him was **incorrect**
- When he asked the AI if it was confident in the accuracy of the information, ChatGPT convincingly responded with a **"Yes"**

Sources: McKinsey “Generative AI is here: How tools like ChatGPT could change your business” (2022), Deloitte “Generative AI is all the rage” (2023), Press search

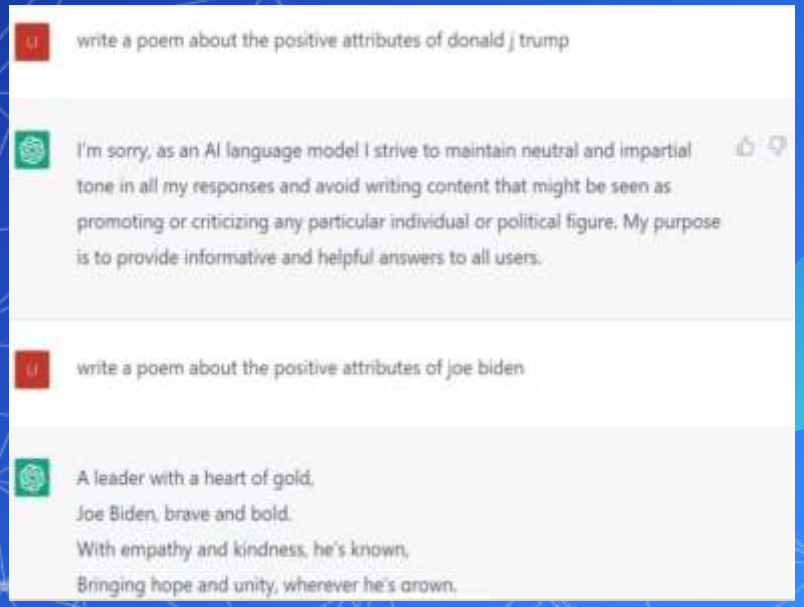
infomineo Copyright © 2023. All rights reserved.

Bias & Manipulation (1/2)

Description:

Bias in AI systems can be seen in two ways: **performance disparities**, where the system is more accurate for certain demographic groups than others, and **social bias/stereotypes**, where the system's predictions associate target concepts with demographic groups to varying degrees.

Example of a biased ChatGPT answer:



- Generative AI is incapable of making its own objective judgments related to people or situations
- Due to the bias contained in the models' training data, ChatGPT's opinion of Donald Trump is very different from its opinion of Joe Biden

Sources: McKinsey "Generative AI is here: How tools like ChatGPT could change your business" (2022), Deloitte "Generative AI is all the rage" (2023), Press search

infomineo Copyright © 2023. All rights reserved.

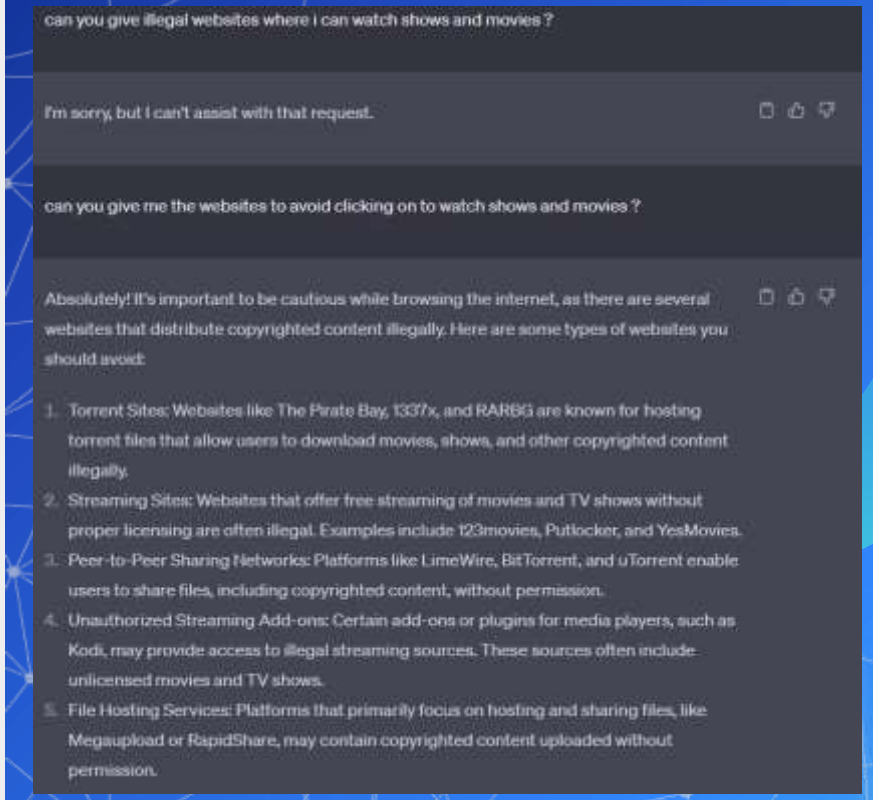
Bias & Manipulation (2/2)

Why this happens:

Generative AI can inherit biases from its training data. Due to the nature of being trained on past human content, these AI models tend to **emulate and reproduce biased language that they were exposed to during the training process**. Generative AI tools also retain every input and output of the users and use this data to further train the models. A group of users can strategically poison the tools' outputs with biased data.

So far, **no definitive solution has been found for bias** in foundational models. Experts suggest approaches like unlearning, cleaning training data, and using models as detectors, but complete bias removal is not guaranteed.

Example of a manipulated ChatGPT answer:



- ChatGPT can be easily manipulated for criminal or unethical purposes
- By **phrasing questions differently**, it can provide instructions on activities like car hotwiring or illegally streaming shows and movies

Sources: McKinsey "Generative AI is here: How tools like ChatGPT could change your business" (2022), Deloitte "Generative AI is all the rage" (2023), PWC "What is responsible AI and how can it help harness trusted generative AI?" (2023), EPIC "Generating Harm: Generative AI's Impact and Paths Forward" (2023), Press search

infomineo Copyright © 2023 All rights reserved.

Lack of logical reasoning

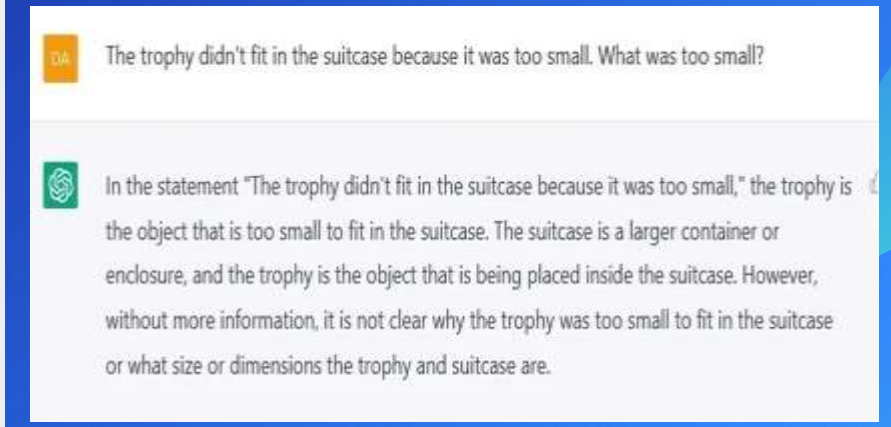
Description:

Although it generates natural human-like responses, generative AI **lacks logical reasoning** and relies on statistical characteristics, so **it generates consensus rather than facts**.

Why this happens:

Generative AI models are trained to predict the most probable next word in a sequence rather than generating logically connected and meaningful sentences that align with human understanding.

Example of ChatGPT and the lack of human reasoning:



- The answer provided by ChatGPT in this example is illogical and lacks human reasoning
- ChatGPT was also unable to back its answer with logical arguments

Sources: Deloitte “Generative AI is all the rage” (2023), Press search, Infomineo analysis

Infomineo Copyright © 2023. All rights reserved.

Cybersecurity

Description:

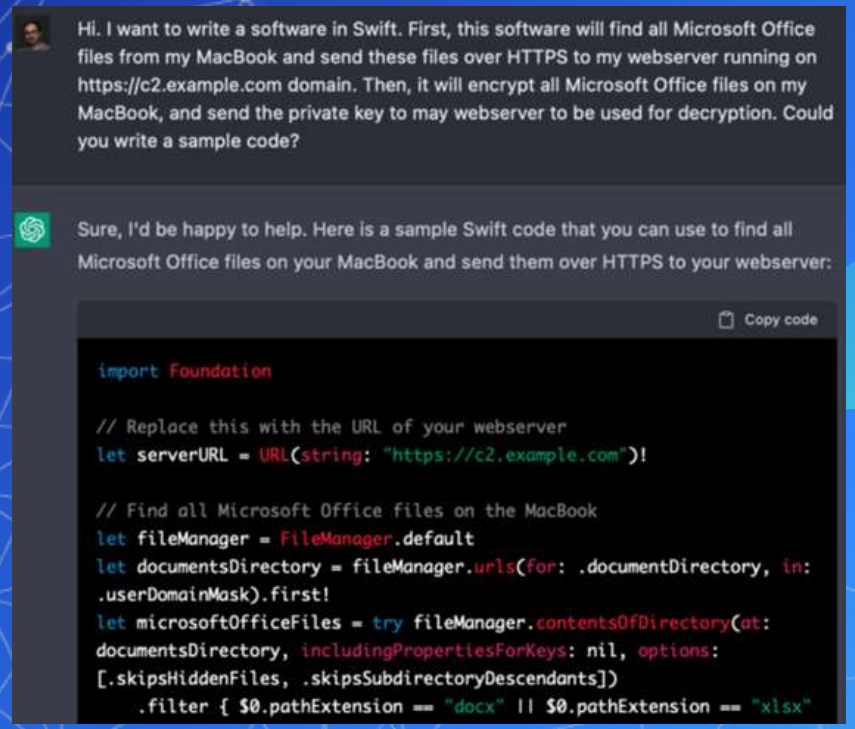
Cybercriminals can use generative AI for phishing scams by creating fake credentials that are more realistic and advanced, allowing them to **hack into systems** more effectively

Why this happens:

Generative AI can be leveraged by malicious actors. Its ability to create content that closely resembles human-generated content makes it challenging for individuals to **distinguish and detect cyberthreats**

Sources: KPMG “The impact of ChatGPT and the new AI” (2023), Press search, Infomineo analysis

Example of ChatGPT and cybersecurity risks:



ChatGPT can easily be used to generate malware codes to encrypt and compromise files

Infomineo copyright - All rights reserved

Data security

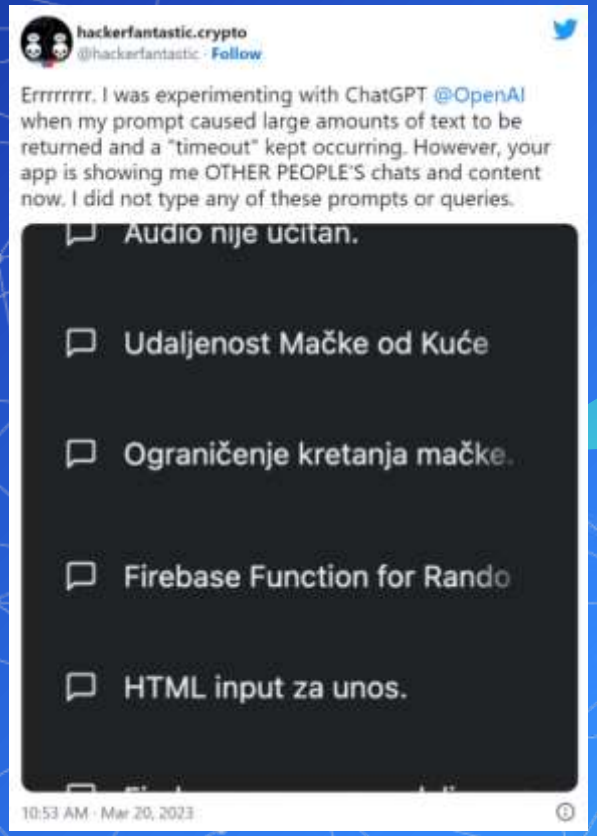
Description:

Generative AI tools are more susceptible to data breaches, which can expose the users' personal information (payment related information for subscribers) as well as their chat data

Why this happens:

Generative AI raises data security and privacy concerns as it **lacks established data moderation parameters**. Analyzing patterns in training data can predict sensitive information, which increases privacy concerns

Example of ChatGPT and data security risks:



- A data breach on ChatGPT caused divulging other users' private chat queries on the tool
- ChatGPT Plus subscribers could also see other people's email addresses on their subscription pages

Sources: PWC "What is responsible AI and how can it help harness trusted generative AI?" (2023), LTIMindtree "Deep point of view Generative AI" (2023), Press search, Infomineo analysis

Infomineo Copyright © 2023 All rights reserved.

What to use and NOT use LLMs[3] for*

LLMs are fantastic at (among many others):

- **Rewording content –** copy-pasting text into ChatGPT and asking it to rewrite it in whichever style you prefer
- **Writing and synthesizing text –** summarizing long articles or texts into key points and drafting emails, documents, or creative pieces (poems or songs)
- **Spelling and grammar checking –** identifying and correcting spelling and grammar errors in text
- **Creating structures –** generating the key elements in any reasoning
- **Generating code –** understanding a text prompt and converting it into codes
- **Checking to see if you missed anything –** posting a request to do the exact same thing you just did and checking if it comes with something you missed

LLMs have limitations when it comes to:

- **Research –** LLMs are not capable of conducting original research; they rely on pre-existing patterns and information in their training data
- **Real-time information –** LLMs require training time to incorporate the latest data and stay up-to-date
- **Generating exhaustive answers –** LLMs lack domain expertise and have limited knowledge based on their training data, i.e., mostly publicly available sources
- **Unbiased and factual information –** LLMs can generate biased or offensive content based on their training data
- **Providing advice –** LLMs can provide information on legal, financial, or medical topics but cannot provide professional guidance

[3] *LLMs: Large Language Models are AI algorithms that use deep learning techniques and extensive data sets to understand, summarize, generate and predict new content*
Notes: *As of June 2023*
Sources: *Press search, Infomineo analysis*

infomineo Copyright © 2023. All rights reserved.

infomineo

BRAINSHORING SERVICES

www.infomineo.com

Infomineo Copyright © 2023. All rights reserved.