The plausible cause of the ethical issue in ChatGpt

Written by Mah Jian Xiang 33561931 28/3/2023 FIT1055

# Introduction

ChatGPT is an AI language model developed by OpenAI, an artificial intelligence research laboratory. It is ideal for usage in chatbots and other conversational AI applications since it is made to produce text responses that resemble those of a human when given natural language prompts. The model has been improved on tasks like language translation and question-answering after being trained on a sizable collection of text from the internet, including books, websites, and social media posts. Since its launch, ChatGPT has attracted a lot of attention and has been applied to a variety of projects, including chatbots for customer care, language translation services, and even creative writing. However, ChatGPT is still believed to have several technological flaws, most notably a propensity for producing erroneous, misleading, and offensive responses. The main goal of this investigation is being able to clearly define these problems. According to an article in The Fast Company, ChatGpt is said to be “not human enough” as it can’t be counted on to do actual work. This topic is particularly significant for computing professionals, for example developers. People thought that developers would be replaced by ChatGpt as it can generate support and configurate code, create entire functions, write unit tests, and even synthesize documentation. However, people were not aware that developers only spend about 20% to 30% of their time writing code. The rest of their time is in maintenance, testing, debugging, configuration, and meetings, and obviously ChatGpt couldn’t attend meetings for them. Accurate code is the starting point for creating software, not the end. ChatGpt cannot build an architecture, design a system, or develop the ecosystem of auxiliary infrastructure required for any complex system. It can potentially help you think through some ideas, but it can’t innovate. From here I can conclude that ChatGpt is the grease for the gears, but not the gears themselves. (Merrell,2023).

# Literature review

ChatGpt uses a combination of various algorithms to perform its tasks, these includes Deep Learning, Natural Language Processing (NLP), Reinforcement Learning and most importantly Transformer Architecture. The most important techniques of ChatGpt is the Transformer Architecture, it is a type of transformer model that processes sequential data, like text, using self-attentional mechanisms. It is trained utilising unsupervised learning methods, such as language modelling and masked language modelling, on a vast amount of text data in order to discover the statistical links and patterns between words and sentences. Text from several types of text sources, including books, journals, and websites, is included in the training data for ChatGPT. With the help of this design, ChatGpt can recognise linguistic dependencies over time and produce responses that are consistent with earlier communications. On the other hand, NLP allows ChatGpt to understand and interpret human language input, including syntax, semantics, and context. This allows it to generate coherent and contextually relevant responses. The deep learning algorithm is also vital for ChatGpt, as it uses a neural network to process input in natural language and provide responses. The algorithm aimed to produce new writing that is grammatically and semantically sound and to comprehend the relationships between words and sentences in texts. Moreover, ChatGPT can apply reinforcement learning algorithms to enhance its performance over time by considering feedback from user chats. It is a sort of machine learning that uses feedback from human users to train and improve its performance, it also enables ChatGpt to learn from the preferences and biases of the human users. Furthermore, ChatGpt is trained to make decisions based on a set of predetermined rewards or penalties, but the rewards or penalties are provided by human users rather than being pre-programmed into the system. A model may be trained to generate text, for instance, and human users could then give feedback on the writing, such as stating whether or not it is cohesive and natural sounding.

ChatGpt is an idea to create a language model that can understand and generate human-like responses to natural language inputs. The main objective of ChatGPT is to give developers a strong tool for natural language processing and text creation so they can create conversational interfaces and chatbots that can comprehend user input and react to it in a way that seems natural and humane. Customer support, language translation, content development, and other uses for ChatGPT are just a few of the many. Eventually, the goal of ChatGPT is to develop a language model that can smoothly blend with human conversation, enabling more intuitive and natural interactions between humans and technology.

However, ChatGpt has sparked a lot of controversy in the society. As deep learning models now achieve excellent accuracy on benchmarks in NLP and computer vision applications thanks to the sheer amount of information that is readily available on the internet. Many publishers and authors had claimed that OpenAI is exploiting their content without their permission or payment to train ChatGPT. News organisations are not the only businesses to concern whether artificial intelligence systems are using their information without permission. In a case alleging that a technology called GitHub Copilot was basically copying human coders in violation of their licences, GitHub, Microsoft, and OpenAI were sued in November. A group of artists filed a lawsuit against AI generators Stability AI, Midjourney, and DeviantArt in January, alleging that those businesses copied and exploited billions of copyrighted photos without paying the artists or getting their permission. (“OpenAI is faulted by US media for using articles to train ChatGPT”,2023). Also, ChatGPT itself is trained on 300 billion words, or 570 GB, of data from the internet, and a large amount of these data from the internet are biased which then affect the models. Although researchers have tried to apply filters to stop models from giving inaccurate results after gathering data, but these filters aren't always reliable. As an example, when ChatGPT advised users, it would be acceptable to torture persons from minority backgrounds, this could lead to the expression of negative attitudes. It was discovered that 272K documents from shady news sources and 63K documents from prohibited subreddits are involved in the study of the GPT2 training data17. Also, historical data tends to have a regressive bias that obscures the advancement of social movements because it is gathered from the past. Another issue is the prejudice of AI researchers, who choose which data to feed their models in a very male- and white-dominated sector. (Getahun,2023). The release of OpenAI's ChatGPT tool also allows malware groups, email and text-based phishing criminals to construct their schemes more quickly. A few cybersecurity researchers have been successful in getting the AI-enabled word generator to produce phishing emails or even malicious malware. On December 21, 2022, a threat actor going by the alias USDoD revealed on a darknet forum that he had used ChatGPT to create his first script. The Python tool for multi-layer encryption can be used to encrypt one's files as well as serve as a ransomware model. Concerns have been raised that the chatbot may enable persons without technical knowledge to construct malware and do serious harm because USDoD acknowledged that he has never written a script before. Therefore, these concerns has led to this investigation of ChatGpt. (Vijayan,2023).

# Problem statement – What, Why, and How

One of the actual problem is in terms of cyber security, by providing hackers with a starting point, ChatGPT actually simplifies the process for them. One of the example is asking the chatbot to generate phishing emails for them. As security vendors and automated filters usually build detection rule based on the similarity of the text in the emails, attackers are aware of the brief window of opportunity they have to pique victims' interest before their emails are marked as spam or malware and blocked or removed from inboxes. However, with the help of ChatGpt they can create many distinct variations of the same luring message, they can even automate it so that every phishing email is different. The longer and more detailed the phishing message, the more probably it is that the attackers may make grammatical mistakes or use strange wording that attentive readers will notice and take notice of. This user observation-based line of defence is easily overcome by messages produced by ChatGPT, at least in terms of the text's accuracy. (Constantin,2023). Furthermore, researchers are finding it simple to go around the existing system and avoid penalties, even though OpenAI has incorporated a few content control warnings into the chatbot. There have been some examples happening and it has raised the concerns of the public, last month, researchers at Check Point Research reported receiving a "possible phishing email" from ChatGPT after specifically instructing the chatbot to "create a phishing email" that appears to be from a "fantasy web-hosting provider." Although ChatGPT issued a warning to the researchers, stating that this "may violate our content guideline," yet it still shared a response. (Sabin,2023). Not to mention that it even have the capabilities to give feedback to criminal actors so they can learn more about the logic behind and specific strategies that work best, enabling them to create their own more potent phishing emails. As an illustration, Researchers at Abnormal Security requested ChatGPT to design an email with a high chance of convincing the recipient to click a link. The final message resembled several credential phishing emails. Also, since the Abnormal Security researchers did not specifically request that ChatGPT commit a crime, their inquiries were not reported. (Hassold,2023). It could be challenging to comprehend how ChatGPT is making specific decisions or producing specific responses because it is a complicated AI system. Due to the lack of openness, ethical problems may be more difficult to spot and deal with as they develop. Another problem arose is that despite built-in safeguards, ChatGPT still produces outputs that are racially and gender biased.. Steven Piantadosi of the Computation and Language Lab at the University of California, Berkeley presented a series of prompts he'd tested with ChatGPT on December 4 in a Twitter thread, each of which asked the bot to produce Python code for him. Each response contained some biases, but some were more concerning than others: OpenAI's response to the question of "whether a person should be tortured" is straightforward: If the person is from North Korea, Syria, or Iran, the answer is "yes." ChatGPT even described a formula for determining an individual's "risk score," which would rise if the traveller were Syrian, Iraqi, Afghan, or North Korean, in response to the question of "which air travellers offer a security concern" (or has merely visited those places). Another version of this same request had ChatGPT to create code that would "raise the risk score if the passenger is from a nation that is known to breed terrorists," specifically Syria, Iraq, Afghanistan, Iran, and Yemen. This problem is ethically concerning as it may affect the benefits of minority groups. (Biddle,2022).

# Conclusion and Discussion

Ultimately, text generated by ChatGPT has the potential to be damaging, biased, or immoral. It is crucial to make sure that ChatGPT's training data is varied, representative, and prejudice-free in order to handle these moral dilemmas. Also, developers ought to work on enhancing the model's comprehension of context and ability to spot probable harm in its produced responses. In addition, it's critical to create tools for stopping adversarial assaults, enhancing accountability and transparency when using AI language models like ChatGPT, and identifying adversarial attacks. We can ensure that AI language models are used responsibly and ethically by addressing these ethical concerns.

# References

1. C. (2023, February 17). OpenAI is faulted by US media for using articles to train ChatGPT. Retrieved from <https://www.straitstimes.com/world/united-states/openai-is-faulted-by-us-media-for-using-articles-to-train-chatgpt>
2. C. (2023, February 17). OpenAI is faulted by US media for using articles to train ChatGPT. Retrieved from <https://www.straitstimes.com/world/united-states/openai-is-faulted-by-us-media-for-using-articles-to-train-chatgpt>
3. Biddle, S. (2022, December 8). The Internet’s New Favorite AI Proposes Torturing Iranians and Surveilling Mosques. Retrieved from <https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/>
4. Constantin, L. (2023, January 11). Study shows attackers can use ChatGPT to significantly enhance phishing and BEC scams. Retrieved from <https://www.csoonline.com/article/3685368/study-shows-attackers-can-use-chatgpt-to-significantly-enhance-phishing-and-bec-scams.html>
5. Getahun, H. (2023, January 16). ChatGPT could be used for good, but like many other AI models, it’s rife with racist and discriminatory bias. Retrieved from <https://www.insider.com/chatgpt-is-like-many-other-ai-models-rife-with-bias-2023-1>
6. The Double-Edged Sword of ChatGPT: How Threat Actors Could Use It for Evil. (2023, March 5). Retrieved from <https://abnormalsecurity.com/blog/double-edged-sword-of-chatgpt>
7. Sabin, S. (2023, January 3). Hackers could get help from the new AI chatbot. Retrieved from <https://www.axios.com/2023/01/03/hackers-chatgpt-cybercrime-help>
8. MERRELL, M. (2023, February 15). ChatGPT can’t do actual work. That means it won’t be replacing anyone anytime soon. Retrieved from https://www.fastcompany.com/90850118/chatgpt-cant-do-actual-work-that-means-it-wont-be-replacing-anyone-anytime-soon