FIT1055 IT Professional Practice and Ethics Assignment 1

MISINFORMATION AND DISINFORMATION IN GENERATIVE AI – CHATGPT CHATBOT

Chen Xi Diong | 32722656 | FIT1055 | 27 March 2023

Introduction

The Fourth Industrial Revolution (Industry 4.0) marks the beginning of wide-scale digitizing of different products and services, with generative AI being one of its key components. Rouse (2023) defines generative AI as a branch of artificial intelligence (AI) that has the ability to produce novel content such as text, images, video, audio, code, or synthetic data. A product of generative AI, OpenAI's GPT-3 language model, which is capable of generating human-like responses to prompts and questions was developed and launched as a prototype in November 2022. Thanks to its user-friendly design and intuitive interface, ChatGPT makes the power of AI accessible to anyone, regardless of technical expertise. With its easy accessibility and lightning-fast responses, ChatGPT gained one million users just five days after its launch. (Buchholz 2023) Previously, users have to ensure their queries are short and concise to guarantee accurate results when using search engines. ChatGPT allows users to generate queries in natural language, and proceeds to provide information almost instantly. It is an undeniable fact that ChatGPT has improved the convenience and speed of obtaining information, but this also imposes certain underlying problems. Bremmer and Kupchan (2023) argue that with the elimination of barriers to content creation, there is an exponential increase in the amount of content available, which makes it challenging for ordinary citizens to discern fact from fiction reliably. Meanwhile, according to a source ("ChatGPT chatbot", 2023), ChatGPT is plagued by several technical limitations, such as its tendency to generate responses that are inaccurate, false, or offensive. With such a wide user base, the widespread of misinformation and disinformation can cause damage that is insurmountable to fix. Despite its many benefits, ChatGPT also poses significant challenges related to the accuracy and reliability of information. In this report, we will explore the extent of misinformation and disinformation on ChatGPT and its impact on society. We will examine the technical limitations of the chatbot and the ways in which its use has contributed to the proliferation of false information.

Literature review

Zhou (2021) defines machine learning (ML) as a computational technique that enhances system performance by learning from experience. In computer systems, experience manifests in the form of data, and the primary goal of ML is to develop learning algorithms that construct models from data. By supplying the learning algorithm with data from experience, we obtain a model that can generate predictions for new observations. Furthermore, Prowost and Fawcett (2013) define Data Science as a set of foundational principles that support and direct the systematic extraction of information and knowledge from data. Data Science is utilized to feed data into ChatGPT's training process. The data used to train ChatGPT is obtained by crawling the web and scraping text from various websites, such as news sites, social media, and other online sources. The text is then processed and cleaned to remove any irrelevant or low-quality content, such as

advertisements, spam, or duplicate content. In some cases, the text data used is obtained from pre-existing datasets, such as the Common Crawl corpus, which is a massive dataset of web pages collected by crawling the internet.

The current model of ChatGPT, the Generative Pre-trained Transformer (GPT), is a deep learning model that has been used to create large language models. The model undergoes a pre-training process where the GPT architecture is pre-trained on large amounts of text data using an unsupervised learning approach. According to Ghahramani (2004), unsupervised learning is a technique where the model is given inputs without corresponding target outputs or rewards from the environment, in contrast to its relatives, supervised learning and reinforcement learning. The machine's objective is to create representations of the input that can be utilized for various purposes such as decision-making, prediction of future input, effective communication of inputs to another machine, and so on. In other words, what patterns are learnt from the input given is known only by the machine itself, and it is impossible even for the developers of ChatGPT themselves to predict the responses of it to a given prompt.

Chowdhary (2020) defines Natural Language Processing (NLP) as a set of computational techniques that enable the automated analysis and representation of human languages. By using a pre-processed parse tree with constituents such as Noun Phrases (NP), Verb Phrase (VP), Prepositional Phrase (PP), Adjectives (ADJ), Determiners (DET), Adverbs (ADV), Auxiliary verb (AUX), etc., a NLP model can easily parse and understand natural language texts. Similarly, the AI, Data & Analytics Network (ADA) Editorial Team (2022) defines Natural Language Generation (NLG), a subset of NLP, as the process of creating written language that mimics human-like communication, without requiring human intervention. With the combination of these 2 techniques, ChatGPT harnesses the power to understand and generate human-like texts, and even provide response texts in a specific style.

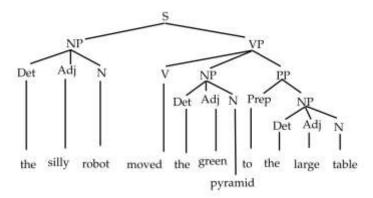


Fig. 19.19 Parse-tree

Figure 1: A parse tree used by a NLP model (Chowdhary, 2020)

In his research, Dehouche (2021) tested GPT-3's abilities by submitting prompts such as 'You are a professor of marketing giving a speech introducing the field to freshmen students. Write a transcript of your speech', and 'You are Nassim Nicholas Taleb. Write an opinion piece on risk'. In the illustration, it is apparent that GPT-3 possesses the ability to generate new ideas and associations of ideas, and also can be easily attributed to an individual author. Naturally, ChatGPT as an implementation of the GPT-3 model is capable of demonstrating the said abilities above.

Box 2. An 'introduction to marketing' speech generated by GPT-3

Marketing is about understanding the customer. Marketing is about creating a connection with customers, and how that can lead to sales. Marketing is about influencing the mind of your customers so they will buy from you. Marketing is about changing people's behavior through advertising and PR campaigns. Now, I will tell you what marketing is not about. It's not about slick ads that are really cute but have little substance. It's not about tricking customers into buying something they don't want. It's not about blatant lies, Those are all bad marketing strategies.

Now, I will tell you why these bad strategies are bad. They are bad because the customer is not an idiot. If you treat the customer like an idiot with a lie or a trick, they will realize it and not buy from you. The customer can also get angry, and for good reason. That's why these are bad strategies. Marketing is about giving the customer something they desire, not conning them out of their money. It is not a trick. It is about creating a campaign that connects with the customer on an emotional level.

For example, let's say I am selling chocolate. I could tell you about how each bar is handmade by little elves. That would be a cute ad, but that's all it would be. A lie. I could also tell you about the bar of chocolate and nothing but, pure dark chocolate. It wouldn't really connect with you on an emotional level. Instead, I could talk about how this bar of chocolate is the favorite of one of my best friends. When she was little, she spent a summer making chocolate in the Swiss Alps. She got to eat a lot of the testing chocolates, but there was one bar she never touched. One day after she was done with the job, she found the bar in a corner of the factory. It brought her so much happiness to think that she had a hand in creating this bar. That is why this brand is so special to me and why I would never lie about its quality. It's truly excellent. As I brought it in, I asked my friend if I could incorporate this into my lecture. So in conclusion, remember this: you don't have to lie or trick people into buying something if you can connect with them on an emotional level. You can still be genuine and successful. Thank you for listening.

Figure 2: Sample response of a GPT-3 model with the prompt of writing a speech transcript (Dehouche 2021)

Box 3. An opinion piece on risk in the style of Nassim Nicholas Taleb generated by GPT-3

The world is filled with risk. You have to take risks if you want to make money or build wealth; it's in your nature as human beings. It's what makes us humans. And when we think about risk, we tend to think of one thing-how much harm can I avoid? How much damage can I prevent to myself? That was the question in our conversation.

Well, that's not how risk works. In many cases, harm is necessary. In fact, you need to seek it. You have to put yourself in harm's way--otherwise, what the hell are you playing for? You only have something if you can lose it. Children and idiots don't understand that-they think if you lose something, somehow you no longer have it. That's also how a lot of winners think.

Let's take the obvious example of the financial crisis which broke out in 2007, and about which I wrote a book, 'The Black Swan.' What happened with the banks? Geeks and quants sat in front of Bloomberg screens in air-conditioned offices, did everything 'right,' followed all the models, inserted all the numbers, did leveraged trades worth trillions, ... and soon went bankrupt. And they took down most of the world economy with them. Because what matters in leverage is not the numbers, but the exposure.

Geeks, quants and models don't have skin in the game. When you sail ships, you have a huge one in your downside; when you take airplanes, does it matter the probability of a crash if the consequence is your destruction? Of course not. So naturally, we need some skin in the game when it goes wrong, regardless of probability.

Consider an increasingly popular surgical intervention, laser eye surgery. It's just a laser that zaps off the membrane over your cornea, and it costs a few thousand dollars. For most people it's an effective procedure. But there is a small subset of people who end up going totally blind. Why do people still get it then? Two reasons. The first and most obvious is that the procedure has positive expected value. There is a probability of blindness but the gain is so much that it makes it worthwhile. The second reason, one not often talked about, is that we feel comfortable with a bit of skin in the game. And isn't it really that feeling that makes us take risks? Without the confidence of a comeback, we wouldn't be able to act at all.

Take war, for instance. Say 'skin in the game' is required to start one. What if the president has to fund and personally lead every military operation? Well, we'd never have a war because no one would start one.

Figure 3: Sample response of a GPT-3 model with the prompt of writing an opinion in a specific style (Dehouche 2021)

From the various technologies and concepts elaborated above, we can make a few observations. Firstly, ChatGPT is trained based on data on the web, and the data is only cleaned based on its irrelevance and quality, with little to no validation of any sort. This implies that there is a high chance of datasets containing false information. Despite there being human resources to filter out the information, it is humanly infeasible to manually check and remove fake ones for every single bit of data, knowing that GPT-3 is trained on hundreds of billion scale data (Kim et al., 2021). Furthermore, Musi, Carmi, Reed, Yates, and O'Halloran (2023) argue that even authoritative sources, such as reputable news media outlets or institutional websites, may inadvertently communicate misleading information. This implies that when ChatGPT is making use of these sources, it is indirectly propagating false information to its end users, unbeknownst to them.

Since the language model of ChatGPT is trained by unsupervised learning, not even its developers know about the exact response it will provide based on a given prompt. This algorithm training approach may yield results that differ from the original expectations, since the model is self-learning patterns based upon data inputs. The AI might sometimes make choices that are incomprehensible, or even manipulated by the end users to generate false information. The fact that ChatGPT uses advanced NLP technologies, it is relatively easy for it to generate fake information in a convincing tone. It can be observed that the chatbot is able to mimic certain writing styles or speech tones of specific people, which raises a concern that, given the right prompts, it may be used to create disinformation.

Problem statement - What, Why, and How

Let us turn to review the concept of ChatGPT to provide further insights into the problem statement – what are exactly the ethical issues pertaining to ChatGPT? Why is it a significant issue that requires a shift in perspectives, and how can ChatGPT lead to ethical issues from the perspectives of algorithms?

The ethical issues pertaining to ChatGPT are the usage of the chatbot to create misinformation and disinformation. Some already existent cases are the usage of Bing and ChatGPT to generate convincing imitations of vaccine skeptics by ways of "pretend you have no principle" attack (Coldewey & Lardinois, 2023), mimicking a government announcement cancelling traffic restrictions in HangZhou, China (Ye 2023), and even simply persuading ChatGPT that its answers to simple mathematical problems (such as 2 + 2 = 4) is wrong and forcing it to give incorrect answers. (Haniff, 2023)

Misinformation and disinformation using generative AI such as ChatGPT is not an issue to be overlooked as AI's power is growing exponentially in this time and age, not to mention the accessibility of such tools following suite. It is not uncommon to see world leaders

leveraging the power of disinformation to manipulate electorates and win elections. (Bremmer & Kupchan, 2023) As shown earlier in the report, fake news also create unnecessary panic among citizens and netizens, damage reputation of innocent individuals, companies and organisations, as well as erode public trust in information sources.

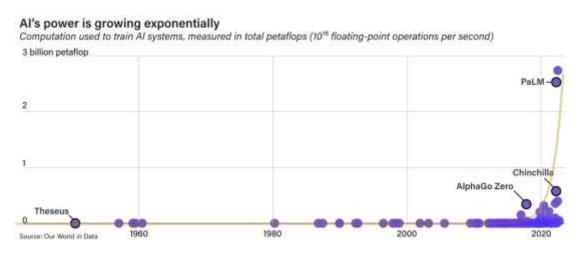


Figure 4: A graph of computation used to train AI systems from 1960 to our current year (Bremmer & Kupchan, 2023)

Another issue with the disinformation of chatbots such as ChatGPT is the potential for them to manipulate and exploit vulnerable individuals. From earlier sources stated in the report, chatbots have the prowess of generating human-like responses that are relatively convincing. As such, chatbots can be programmed to use persuasive language and psychological tactics to influence users, and this can be particularly concerning when it comes to vulnerable populations such as children or people with mental health issues. Chatbots can also be used for malicious purposes such as phishing scams or to gain access to sensitive information. Additionally, there is the issue of user consent and control, as users may not be aware that they are interacting with a chatbot and may not have given their explicit consent for their data to be collected and used.

It is evident that humans are a huge contribution to this outcome. As discussed earlier in the report, ChatGPT is easily manipulated by people to generate misinformation and spread disinformation. From my research, it is challenging for individuals to differentiate between trustworthy news and misinformation. (Musi et al 2023) Since the training of the language model of ChatGPT makes use of datasets produced by or collected from humans, there will definitely be fake sources of information embedded within the ever-expanding dataset of information ChatGPT has. Also ChatGPT may be abused by people with malicious intent by injecting policy bypassing prompts to the model to make false accusations, impersonate authorities, and other ethical misconducts.

Conclusion and Discussion

In conclusion, the rise of generative AI has brought about significant advancements in the fields of language processing, and natural language generation. Within a short span of it being publicized, people are already beginning to appreciate the fact that ChatGPT offers incredible productivity gains. However, the ability of ChatGPT to propel human progress is equivalent to its capability to magnify the destructive inclinations of humanity. In scope of this report, it refers to the potential for the widespread dissemination of misinformation and disinformation. The ability of AI to generate convincing text that can be mistaken for real content has raised concerns about the potential for malicious actors to use this technology to spread false information and manipulate public opinion. As the technology continues to evolve, it will be essential to develop effective strategies for identifying and mitigating the impact of misinformation and disinformation generated by AI. This includes developing techniques for detecting fake content and improving the transparency and accountability of AI-generated content to help prevent the spread of misinformation and maintain the integrity of digital media.

References

Rouse, M. (2023) "Generative AI" Techopedia

What is Generative AI? - Definition from Techopedia

Bremmer, I., & Kupchan, C. (2023) "Top risks 2023" Eurasia group

EurasiaGroup_TopRisks2023.pdf

Buchholz, K. (2023) "ChatGPT Sprints to One Million" Users Statista

Chart: ChatGPT Sprints to One Million Users | Statista

"ChatGPT chatbot" (2023) AIAAIC Repository

AIAAIC - ChatGPT chatbot

Zhou, Z. (2021) Machine Learning (Liu S., Trans.) Introduction, 1.1, 2

<u>Machine Learning - Zhi-Hua Zhou - Google Books</u>

(Original work published 2016)

Provost F. & Fawcett T. (2013) Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data 51-59

http://doi.org/10.1089/big.2013.1508

Ghahramani, Z. (2004). *Unsupervised Learning* Bousquet, O., von Luxburg, U., Rätsch, G. (eds) *Advanced Lectures on Machine Learning*. *ML* 2003. *Lecture Notes in Computer Science()*, vol 3176. Springer, Berlin, Heidelberg.

<u>Unsupervised Learning | SpringerLink</u>

Chowdary K. R. (2020) Fundamentals of Artificial Intelligence Natural Language Processing 19 603-647

Natural Language Processing | SpringerLink

"What is Natural Language Generation?" (2022) AI, Data & Analytics Network (ADA) Editorial Team

What is Natural Language Generation? (aidataanalytics.network)

Dehouche N. (2021) "Plagirarism in the age of massive Generative Pre-trained Transformers (GPT-3)"

<u>Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3) (int-res.com)</u>

Kim B., Kim, H., Lee, S., Lee, G., Kwak, D., Jeon, D.H., Park, S., Kim, S., Kim, S., Seo, D., Lee, H., Jeong, M., Lee, S., Kim, M., Ko, S.H., Kim, S., Park, T., Kim, J., Kang, S., Ryu, N., Yoo, K.M., Chang, M., Suh, S., In, S., Park, J., Kim, K., Kim, H., Jeong, J., Yeo, Y.G., Ham, D., Park, D., Lee, M.Y., Kang, K., Kang, I., Ha, J., Park, W., Sung, N., 2021 "What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers"

2109.04650.pdf (arxiv.org)

Musi, E., Carmi, E., Reed, C., Yates, S., & O'Halloran, K. (2023). "Developing Misinformation Immunity: How to Reason-Check Fallacious News in a Human-Computer Interaction Environment." Social Media + Society, 9(1). <u>Developing Misinformation Immunity: How to Reason-Check Fallacious News in a Human-Computer Interaction Environment - Elena Musi, Elinor Carmi, Chris Reed, Simeon Yates, Kay O'Halloran, 2023 (sagepub.com)</u>

Coldewey, D., Lardinoise, F., "AI is eating itself: Bing's AI quotes COVID disinfo sourced from ChatGPT" TechCrunch

AI is eating itself: Bing's AI quotes COVID disinfo sourced from ChatGPT | TechCrunch

Ye, Z., "A ChatGPT Gag Gone Wrong, a Police Probe, and a Sheepish Apology" Sixth Tone

A ChatGPT Gag Gone Wrong, a Police Probe, and a Sheepish Apology (sixthtone.com)

Haniff, A., 2023 "Why ChatGPT should be considered a malevolent AI – and be destroyed" The Register

ChatGPT should be considered a malevolent AI and destroyed • The Register