

On high alert: The darker side of generative AI





The emergence of new technologies is making our daily lives more convenient, but with the rise of generative AI, to date most notably ChatGPT and DALL-E/DALL-E 2, it has become easier for individuals and organizations to misuse these technologies for malicious purposes. While it is imperative to continue to develop and adopt new technologies, we must also remain vigilant and aware of the potential risks, and take necessary measures to prevent financial crimes, as well as crime on all fronts.

“What is ChatGPT?”

“I am ChatGPT, a large language model trained by OpenAI, based on the GPT (Generative Pre-trained Transformer) architecture. I have been trained on a massive corpus of text data using unsupervised learning techniques to generate natural language responses to a wide variety of prompts and questions. My training data comes from diverse sources, including books, articles, and web pages, allowing me to have a broad knowledge base on a wide range of topics. I can answer questions, generate text, summarize documents, translate languages, and even write stories, among other tasks.”¹

ChatGPT, an example of generative AI, released to the public in November 2022 is software which mimics human language.² Since its launch, it has been at the center of many discussions both in coffee shops and boardrooms. Hailed as a game-changer that makes many tasks easier, such as contract drafting, story writing, and researching, there is potential danger lurking behind the exciting façade. Like any technology, it can unfortunately be misused for criminal purposes. The ease of access to AI and machine learning technology makes it possible for both individuals and organizations to exploit these tools for a wide range of nefarious activities, as both Europol and the UK’s National Cyber Security Centre have recently warned.³

Both advantages and disadvantages stem from the use of generative AI by criminals and for crime prevention. This article will discuss how this tool can enable malicious actors to carry out a range of activities, from fraud and cyberstalking, to impersonation and dissemination of false information.

Firstly, let’s discuss some of the benefits associated with the utilization of generative AI.

While it is imperative to continue to develop and adopt new technologies, we must also remain vigilant and aware of the potential risks, and take necessary measures to prevent financial crimes, as well as crime on all fronts

1. **Crime prevention:** Generative AI can be used to analyze data, and identify patterns in historic data to predict crimes before they happen. For example, it can assist in monitoring social media platforms and online forums for any red flags which may indicate suspicious behavior. This can enable quick and efficient detection, and prevent any potential criminal activity.
2. **Investigative assistance:** Generative AI can be used to analyze large volumes of data, and provide insights that could be difficult for humans to identify. Data from social media platforms may have been traditionally difficult to monitor on an ongoing basis. However, with the processing power of generative AI, this can be monitored quickly and easily to identify potential suspects, predict their movements and activities, and gather evidence to support investigations.
3. **Language translation:** Generative AI can translate different languages, which can be useful when investigators speak a different language to that which they are investigating in. Generative AI can help overcome language barriers in order to shed light on information in different languages, and ensure that investigations are conducted effectively.
4. **Crime analysis:** Similar to crime prediction, generative AI can also be used to analyze data, and identify patterns in historic data to understand crime more effectively. This can help to allocate investigators or enforcement agency resources efficiently, and develop strategies to prevent crime. For example, if a particular area is experiencing a high rate of financial crimes, generative AI can help identify the most common methods used by malicious actors, and provide insights on how to prevent them.

As with any form of artificial intelligence, generative AI has no consciousness; it can't rationalize why it generates the output it produces, and has no real understanding of the human experience. It is purely taught

to behave in a certain way through natural language processing. While limitations have been built into generative AI to mitigate the risk of it being used for nefarious activities, these limitations can be circumvented at times.⁴

Phishing is the most common type of cybercrime attack which, according to a report by Proofpoint, accounted for around 70% of data breaches in 2022

Some of the areas for potential misuse of large language models and AI chatbots include:

Phishing

Phishing is the most common type of cybercrime attack which, according to a report by Proofpoint, accounted for around 70% of data breaches in 2022. It is estimated that 3.4 billion phishing emails are sent every day.⁵ It is no surprise then that it is also one of the most common fraudulent uses of generative AI. Phishing is the process whereby malicious actors send messages, emails, or text messages that appear to derive from a legitimate source, such as a bank, and trick the recipient/victim into providing sensitive information, such as login credentials or financial information. Historically, phishing attacks could be detected through the grammatical and spelling errors within the text. Given generative AI's ability to generate human-like speech, it can be used by malicious actors to mimic the appearance and language use of legitimate sources. The speed and ease with which AI language models, such as generative AI, can be programmed to generate text messages that mimic human communication makes it a powerful tool. It also makes it harder for recipients to detect fictitious and fraudulent emails. >

Malicious actors can use generative AI to create convincing text messages, emails, or other types of content which mimic real people, and ultimately obtain the sensitive information they need

Cybercrime and hacking

Not only does generative AI have natural language processing capabilities, but it can also generate code in various programming languages such as Python, Java, JavaScript, C++, HTML, and CSS. Although the code that is generated may not be optimized for production use, it serves as a starting point. This becomes particularly useful for malicious actors with limited technical knowledge to gain insights to attack a system. It also accelerates the process by which they can optimize their coding to break into systems and the ability for them to translate natural language into working code. To extend on this, generative AI has also seen the ability to flow between programming languages and natural languages, making it easier for such malicious actors to build “end to end” malicious campaigns, which may start with phishing attempts, and end in the generation of malicious software.⁶

Impersonation

As digital identity verification becomes more common place, the risk of fraudulent identities and impersonation increases. The Guardian heralds ChatGPT as the best software program for impersonating humans ever released to the public.⁷ The danger is that this can be used for fraudulent means, including the use of fake identities to access sensitive information, and stealing funds through unauthorized financial transactions. Malicious actors can use generative AI to create convincing text messages, emails, or other types of content which mimic real people, and ultimately obtain the sensitive information they need. Additionally, tools such as DALL-E can be used as a kind of natural language Photoshop to create convincing yet fake photos.⁸

Impersonation can cause significant damage to the impersonated individual

and company, both reputational and financial. The ease with which generative AI can generate convincing text messages, coupled with the natural feel of the generated output, makes it a powerful tool overall.

Dissemination of false information/propaganda

The dissemination of false information and propaganda is a growing concern in the digital age. Generative AI can be used to create false and convincing content, including social media posts and other types of content that are designed to mislead and deceive individuals. Additionally, generative AI can be used to generate and disseminate fake news at an increased speed and volume, which can have a detrimental impact on society, as we have seen time and again. This shows the use case in terrorism and propaganda as well. NewsGuard performed an exercise whereby they prompted ChatGPT on 100 false narratives. Although there are limitations which exist to prevent the generation of fake information, they achieved incorrect but articulate and fluent claims regarding 80% of the topics shared with ChatGPT to prompt fake news.⁹ Generative AI can generate legitimate looking text messages that appear to be written by a human, as well as fake photos which can appear to be compromising for an individual; this makes it difficult for individuals to detect that they contain false information. The ease with which generative AI can be programmed to generate false information makes it a powerful tool for those who seek to deceive and mislead.

There is no doubt that ChatGPT and DALL-E have not only brought the topic of generative AI to the forefront of the public, but also have become a commonplace tool for many. Coupled with its benefits,

generative AI's criminal use is a real concern. In all the above cases, we can see how the speed, ease, and ability to mimic humans can pose a risk for fraudulent behavior against individuals and organizations.

To mitigate the risks of generative AI, individuals and organizations should drive awareness of these risks alongside the benefits. There is no doubt that generative AI and other AI-based tools can support law enforcement and regulators with detecting fraudulent behavior, preventing crime, supporting investigations, and analyzing crime data. As Sam Altman, the CEO for OpenAI, stated in a US Senate hearing, "We think that regulatory intervention by governments will be critical to mitigate the risks of increasingly powerful models."¹⁰ It's vital to ensure that these tools are used responsibly and measures are put in place to prevent their misuse for criminal activities. ●

By **Ralph Stobwasser**, Partner, Forensic and **Nicki Koller**, Manager, Forensic, Deloitte Middle East

Endnotes

1. <https://chat.openai.com/>.
2. <https://openaimaster.com/chat-gpt-login/>.
3. <https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk>; <https://www.europol.europa.eu/media-press/newsroom/news/criminal-use-of-chatgpt-cautionary-tale-about-large-language-models>.
4. <https://www.europol.europa.eu/cms/sites/default/files/documents/Tech%20Watch%20Flash%20-%20The%20Impact%20of%20Large%20Language%20Models%20on%20Law%20Enforcement.pdf>.
5. <https://aag-it.com/the-latest-phishing-statistics/#:~:text=The%20US%2Dbased%20IC3%20received,than%20%242.7%20billion%20in%202022>.
6. <https://www.europol.europa.eu/cms/sites/default/files/documents/Tech%20Watch%20Flash%20-%20The%20Impact%20of%20Large%20Language%20Models%20on%20Law%20Enforcement.pdf>.
7. <https://www.theguardian.com/commentisfree/2022/dec/08/the-guardian-view-on-chatgpt-an-eerily-good-human-impersonator>.
8. <https://www.techtarget.com/searchenterpriseai/feature/A-closer-look-at-what-makes-the-AI-tool-Dall-E-powerful>.
9. <https://www.newsguardtech.com/misinformation-monitor/jan-2023/>.
10. <https://www.theguardian.com/technology/2023/may/16/ceo-openai-chatgpt-ai-tech-regulations>.

There is no doubt that generative AI and other AI based tools can support law enforcement and regulators with detecting fraudulent behavior, preventing crime, supporting investigations, and analyzing crime data