What is Dark AI?

By Daniel Iwugo

Unless you've been disconnected from the Internet and the world at large the past few months, you have probably heard of ChatGPT.

ChatGPT is 'an AI-powered language model developed by OpenAI, capable of generating human-like text based on context and past conversations'. In English, it's a great tool that answers almost any questions you can ask another human, and can do your homework for you (you didn't hear that last part from us 😉).

However, much like any tool, ChatGPT can be used for legitimate and malicious purposes. Hence, let's introduce the term 'Dark AI'.

What is Dark AI?

Dark AI is the concept of programming AI intentionally or unintentionally to carry out malicious activities. This could be by manipulating biases, data poisoning or simply letting it into the hands of

threat actors to modify the source code.

Dark AI has been around as long as 'Bright' AI has. In this article, we'll take a peek at the progress of Dark AI in Information creation and manipulation.

Without further ado, let's jump in.

Images

AI generated images could be used by threat actors for impersonation, misinformation, and disinformation operations (That's a lot of information). Tools such as thispersondoesnotexist.com can be used to create headshots for fake social media profiles, which could be used for impersonation or a fake persona. On the other hand, text-to-image tools such as Midjourney can be used to make hyper-realistic scenes such as a fake city, war, or incident.

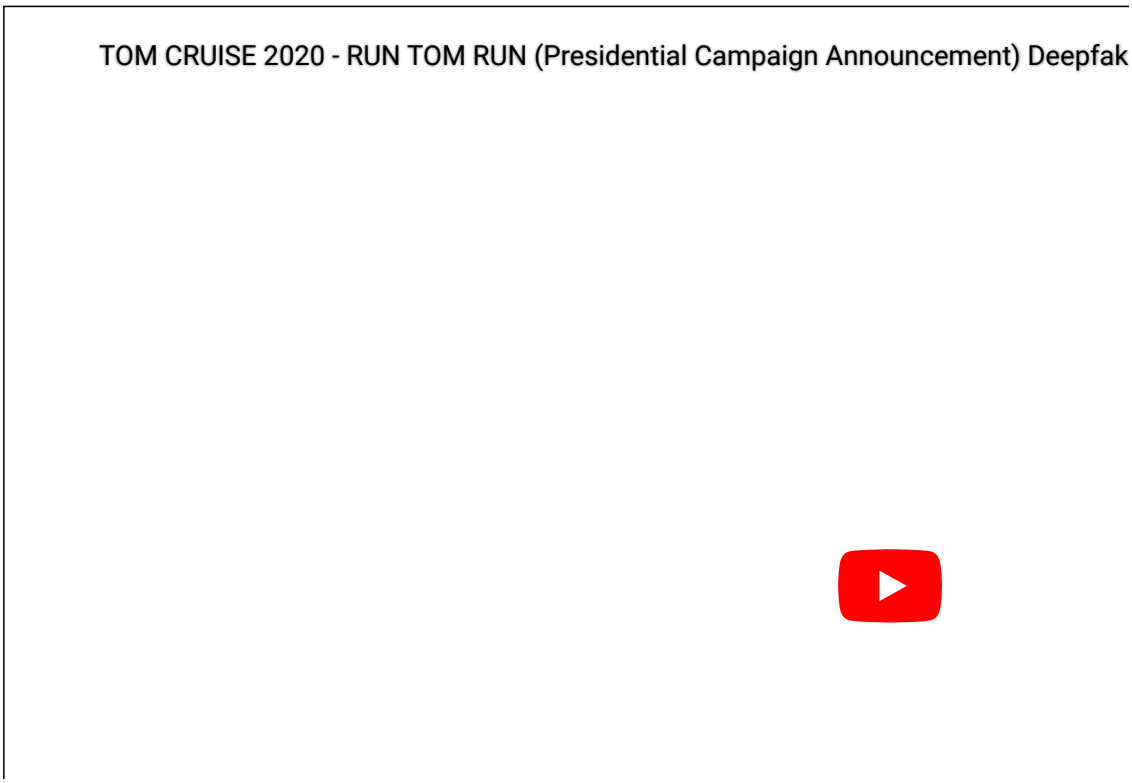
In March, a picture of Pope Francis getting his holy 'drip' on went viral. Turns out it was just a picture from a prompt made with Midjourney, an AI art tool.

The Pope in a puffer jacket ¦ Credit: Twitter.com

Video

Deepfakes have gone beyond replacing a face in an image. Advanced AI tools allow for full on manipulation and creation of videos. A common application of this are viral videos in which a prominent person says something they never said. This poses two problems. First, a person can be framed on the basis they did things that they actually never did. The other but possibly future scenario, is that due to the nature of deepfakes existing, a person could lie that real video evidence is a deepfake and is not the truth. Talk about a paradox.

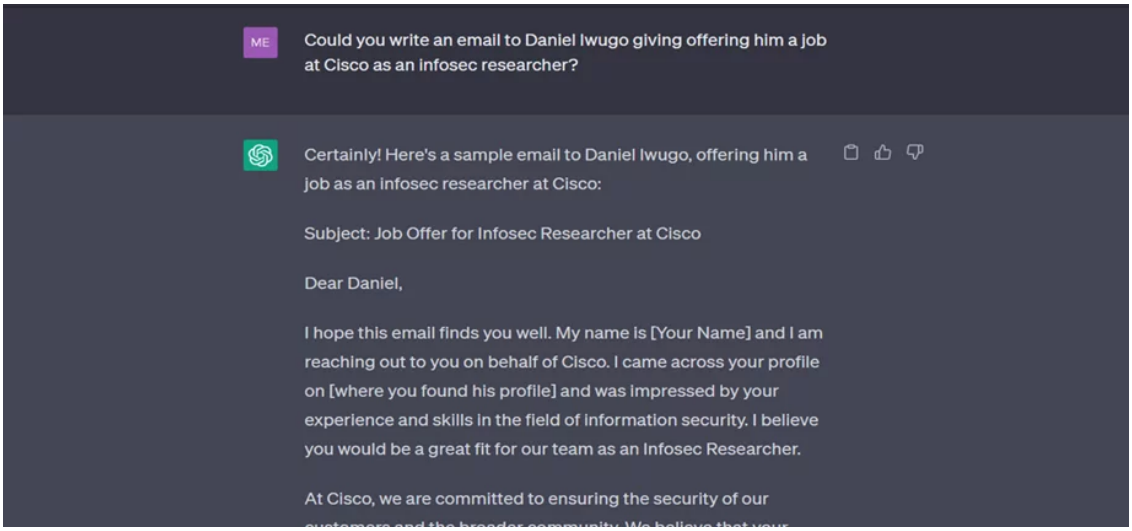
Here is a deepfake of Tom Cruise running for the 2020 United States Election (Spoiler: He never ran 🏃).



Text

Probably the most common lately, text generation has become the talk of the town. This is because phishing templates are written by generative AI quite easily. This has lowered the bar for text-based social engineering attacks as common grammar mistakes that were once used to identify spam and phishing emails are no longer applicable.

Below is an example of a phishing email I wrote to myself in an experiment earlier this year with ChatGPT.



ChatGPT writing a potential phishing email ¦ Credit: Daniel Iwugo

Although not observed frequently, voice cloning and text-to-speech AI tools could be used for social engineering over the phone, also known as vishing.

Mitigations

Even with the threats Dark AI poses, it's a bit difficult to actually defend against it. However, it is not impossible:

1. Always verify the source
   Fake images and videos come from the weirdest of sources before they go viral. If you can't find the source or link it to a legitimate news handle, it's probably fake anyway.
2. Check for any red flags
   Deepfakes are getting harder to tell from reality. But there can be subtle hints. Uneven skin tone, poor quality, and wrong physical dimensions are red flags to note. Also keep in mind that in videos, lip movements may not correspond to the audio in fakes.
3. Look for grammar errors in text
   Despite tools like ChatGPT making text generation a lot easier without grammatical errors, you can keep in mind other factors when receiving a message. When you receive a message, take note of 3 things: Medium, Sender, and Action. Was it sent via email, SMS or iMessage? Who sent this and do I know the person? And lastly, what are they asking me to do?
4. Keep your emotions in check
   Threat actors understand that when we get information, we attach a level of urgency to it. However, the higher the urgency, the less logical our actions tend to be. When you receive an image, video, message or call, keep a cool head 🤖. This could help you sense if something is fishy a lot faster.
5. Have a good conversation
   Going back on AI audio tools, if you receive a call that has a tone of urgency, relax, and try to understand the situation. The deeper the conversation, the more likely the threat actor using the AI will slip up because it's an impersonation, not the real person.

Tags: ARTIFICIAL INTELLIGENCE CHATGPT CYBERCRIME INTERNET SAFETY

RECENT POSTS

The CYSED Dictionary: Email

What is Dark AI?

PDF files more dangerous than Executables – Report reveals

ALL CATEGORIES

Awareness

Bussiness

Child Online Safety

Cybercrime

Cyberdiplomacy

Cybersecurity

Events

Privacy

Technology

Threat & Attacks

TAGS