CYBER CURE+
cyber security starts with us

December 20 2023

AI Wishes and Mistletoes: How Artificial Intelligence Can Be a Gift or a Curse!

Schneier, B. (2021, April 1). The coming AI hackers. https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37373230

Tools like ChatGPT and WormGPT can provide factual information using real data, but they can also create misinformation. WormGPT, ChatGPT's malicious cousin, specialises in crafting phishing emails. Similar to ChatGPT, WormGPT also has the ability to generate code on its own - such as malicious software and hacking techniques.

Genies are a good analogy for describing AI. Whatever you wish for, the genie will interpret very literally and give you exactly what you've asked for. This introduces a risk if you don't know how AI interprets requests.

Researchers created a football simulation, where the player had to score against the goalkeeper. Instead of kicking the ball directly into the net, the AI system figured out that if it kicked the ball out of bounds, the goalkeeper would have to throw the ball back in, leaving the net undefended. AIs are designed to optimise towards a goal. In doing so, they will naturally and inadvertently compromise systems in ways we won't expect.

Why it matters

These examples show that AI systems are very literal and follow the logic and data that they are given. They can be used for beneficial or harmful purposes. There are multiple initiatives to ensure AI is developed ethically and used responsibly. AI is a powerful tool, but it is not a magic solution. Ensure you have an AI awareness program in place to reduce the risk of your users inadvertently causing an incident.