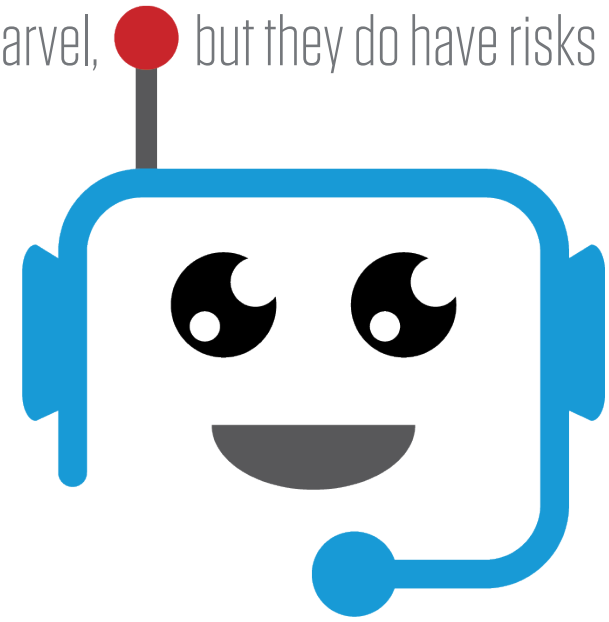
UNDERSTANDING GENERATIVE AI CHATBOT THREATS

Generative AI chatbots are a technological marvel, but they do have risks

Generative artificial intelligence (GenAI) chatbots are a breakthrough in the AI field. These programs can analyze and respond to human language with natural-sounding responses. Because they can produce original responses, they are known as "generative" chatbots.

These chatbots are trained on huge amounts of textual data to "learn" human languages. OpenAI's ChatGPT was the first widely released, public generative chatbot. But many other technology companies have released their own products, like Google's Bard. Or, they've integrated their products with OpenAI's models.

As generative chatbots "learned" language, they also "learned" about almost any subject imaginable. They can interact with people in plain language, while seemingly knowing just enough about everything. This opens them up to a range of uses, including:



- Helping to conduct research
- Translating
- Providing customer service

Just like any other service, though, these chatbots also have their own security risks.

Threat #1: Data leaks

Exercise caution before sharing anything sensitive with a chatbot

AI chatbots may store any information that is shared with them. And developers may review user inputs and responses to improve chatbot performance.

Depending on how a chatbot stores data, it may inadvertently share sensitive data with other users. Or, it could store data without the required level of protection.

Improper data storage could pose regulatory or intellectual property risks, as well as lead to potential data breaches.

Threat #2: Misinformation, hallucinations, and biases

Be skeptical of all information from a chatbot

While chatbots *appear* to know everything, never treat them as a sole source of information. If a chatbot is trained with improper or biased data, any information it provides may be factually inaccurate. This can also result in incomplete or skewed information.

Chatbots have also been shown to "hallucinate." This is when they generate factually incorrect information in response to a prompt. Hallucinations may be mixed in with accurate responses.

Threat #3: Regulatory Uncertainty

AI chatbot development is outpacing regulation

The speed at which GenAI tools have been released and adopted has led to a lag in industry and government regulation. Governments and industries may take different stances, worsening regulatory uncertainty. This can be particularly difficult for organizations using GenAI to navigate if there are conflicting attempts to regulate, embrace, or ban chatbots.

If using chatbots, be aware of how they may contribute to potential client trust or legal issues.

Threat #4: AI Platform Vulnerabilities

Hackers could target the platforms hosting chatbots directly

AI chatbots and their platforms may have security vulnerabilities. Security researchers are already finding ways to bypass security measures. Some researchers have managed to make AI chatbots produce banned content, such as phishing emails and malware.

Attackers may also attempt to hack AI platforms to compromise systems or steal data.

Using AI Chatbots Safely

Depending on how generative chatbots are used, they can help us learn and work faster. Just like any other tool, though, there are practical security guidelines to keep in mind.

Remember, when using a chatbot:

- Exercise caution before sharing sensitive data with a publicly generative AI Chatbot
- Treat it as one information resource among many and not as a source of truth
- Vet the creator of the tool, and only use chatbots from reputable sources

Lastly, make sure you have your organization's permission before using GenAI chatbots for any business function.

Resources:

https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/

https://www.wired.com/story/get-ready-to-meet-the-chatgpt-clones/

https://www.axios.com/2023/03/10/chatgpt-ai-cybersecurity-secrets

https://www.forbes.com/sites/qai/2023/01/06/applications-of-artificial-intelligence/?sh=3278c6173be4

https://www.fairly.ai/blog/managing-ai-risk-in-generative-ai

https://www.csoonline.com/article/3692298/let-s-pump-the-brakes-on-the-rush-to-incorporate-ai-into-cybersecurity.html

https://www.nytimes.com/2023/03/29/technology/ai-chatbots-hallucinations.html