AI Chatbot Usage Guidelines

AI chatbots like Microsoft's *Bing Chat*, Alphabets *Bard* or OpenAI's *ChatGPT*, are useful tools, however there are some risks you should be aware of when using them.

To help you to use Large Language Models (LLMs) safely, this article outlines the risks of which you should be aware and provides some guidelines to keep your data safe and enhance your experience.

1. Background

There are different kinds of chatbots including:

Menu/button-based - a limited set of questions and responses.

Rules-based AIs - a rudimentary conversational tool with a predefined conditional rule set using a limited data set.

AI-powered - has rules which focus on making conversations flow better. They use machine-learning algorithms and automation capabilities like robotic process automation (RPA). Also known as "Conversational AI chatbots" these can remember conversations with users and incorporate this context into their interactions.

Voice - a conversation bot that allows users to interact with the bot by speaking to it, rather than typing. These bots use Interactive Voice Response (IVR) technology and can be the automated interface for answering phone calls.

Generative AIs - able to generate a human-like response to queries, generative AI chatbots can take this a step further by generating new content as the output. This new content could look like high-quality text, images and sound based on LLMs they are trained on. Chatbot interfaces with generative AI can recognize, summarise, translate, predict and create content in response to a user's query without the need for human interaction.

*"**Generative AI** is a broad category for a type of AI, referring to any artificial intelligence that can create original content. Generative AI tools are built on underlying AI models, such as a large language model (LLM). LLMs are the text-generating part of generative AI."*
- Elizabeth Bell, *Appian*

*"The main **difference** between [non-generative] **chatbots** and **LLMs** is that LLMs are more advanced in terms of their language capabilities. While chatbots are limited to pre-programmed responses, LLMs can generate responses that are more personalized and nuanced."*
- Mhd Saeed Al Hasan, *PMP®*

Generative AI chatbots can be used to automate tasks and provide instant (*well-formatted*) information from searches to streamline your work. However, the data you provide them as input may also be used to train LLMs or for other purposes over which you may have limited control.

2. Common Risks

Accuracy of Output

Chatbots learn by weighing probability on large data sets of unknown provenances, which can result in inaccurate output.

Hijacking Conversations

When asked to analyse a website, LLMs often extract (aka 'scrape') information from the site. Unfortunately, miscreants can hide 'invisible' code within the site which can be interpreted by the LLM as 'instructions', thereby changing the original instructions entered. These "invisible prompts" can therefore quietly hijack your conversation executing the requests of the miscreants instead.

More information on conversation hijacking.

Sensitive Information

Your conversations may be used by chatbots as training material. Hence, it is not wise to include any sensitive data (including research, intellectual property, or professional identifiers) in your chatbot queries lest they threaten the confidentiality or integrity of that information.

3. Mitigation Guidelines

Human in the loop (HITL)

You should always validate the accuracy of any output you received from an LLM-based AI chatbot using independent sources.

Setting permissions

If you choose to share your content with an AI chatbot, you need to determine whether the way in which it handles your data is appropriate for the category or classification of the data uploaded.

Authorisation

You must also ensure you have the necessary rights to upload any data to a chatbot. For example, when dealing with Personally Identifiable Information (PII) you should ensure you have permission to use the information as input without de-identifying the data first or indeed whether sharing such data adheres to any appropriate legislation such as the Privacy Act.

> As a rule of thumb: You should not provide Generative AIs with information that needs to remain confidential such as: PII, financial details, social media details or passwords.

Access to networks containing sensitive information

Limit the access chatbots have to your home and UQ networks. If a chatbot asks for permission to access networks or asks you to make changes to your firewall, carefully consider the implications and seek expert assistance if in doubt.

Opting-out

Switch off "**training**" mode when using an LLM. You may need to search for an **opt-out form** on their website.

e.g. When using **ChatGPT** you can switch off training in *ChatGPT settings* (under Data Controls) to disable *training* for conversations, or you can submit this form. Once you opt out, all new conversations you have using the same user access account will not be used to train their LLM.

Reputation check

With the plethora of chatbots to choose from, be wary of using new, untested LLMs. Review chatbots thoroughly before use and/or use those provided by well-established organisations, e.g., Bing Chat Enterprise, Google Bard, OpenAI. etc.

4. General Reminders

Be Vigilant when using any automated communication channel. It is possible that the channel has been crafted to exfiltrate or socially engineer sensitive information from users.

Ensure that the interface you are using is directly connected to the chatbot and is not a trojan horse front end relaying your communications to the chatbot while quietly capturing them at the same time.

Report security concerns, involving Bing Chat, ChatGPT or other AI tools to the Cyber Security Operations Centre (CSOC) here:
https://support.my.uq.edu.au/app/opa/report_a_cyber_security_incident

5. Further Reading

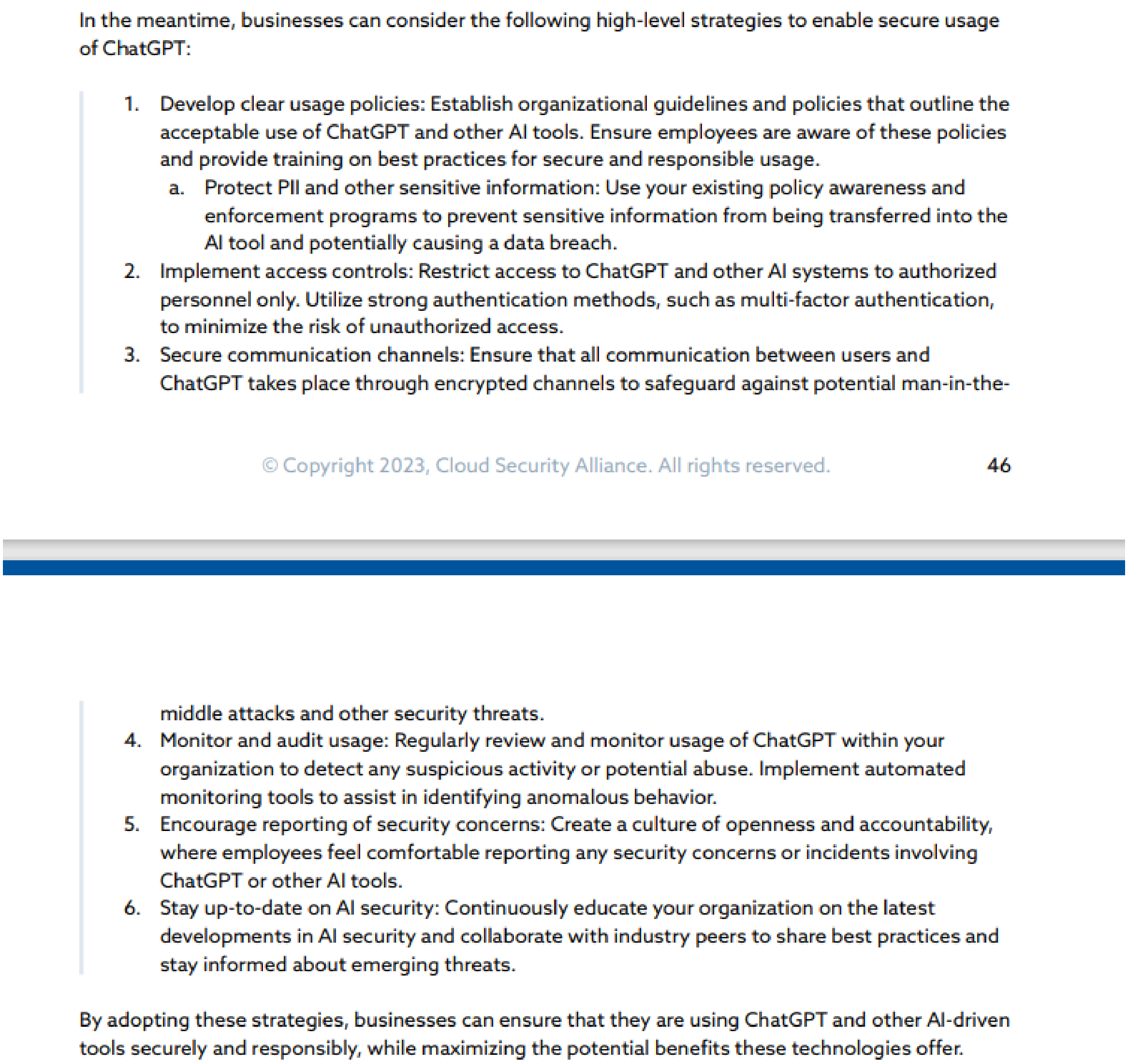
ChatGPT's Educator FAQ | OpenAI Help Center

Bing Chat's Educator FAQ | Bing Chat Enterprise Help Center

Google's Bard | Bard FAQ

Cloud Security Alliance | Security Implications of Chatgpt - see also excerpts in Appendix.

Appendix – Recommendations from the Cloud Security Alliance



In the meantime, businesses can consider the following high-level strategies to enable secure usage of ChatGPT:

1. Develop clear usage policies: Establish organizational guidelines and policies that outline the acceptable use of ChatGPT and other AI tools. Ensure employees are aware of these policies and provide training on best practices for secure and responsible usage.
    a. Protect PII and other sensitive information: Use your existing policy awareness and enforcement programs to prevent sensitive information from being transferred into the AI tool and potentially causing a data breach.
2. Implement access controls: Restrict access to ChatGPT and other AI systems to authorized personnel only. Utilize strong authentication methods, such as multi-factor authentication, to minimize the risk of unauthorized access.
3. Secure communication channels: Ensure that all communication between users and ChatGPT takes place through encrypted channels to safeguard against potential man-in-the-middle attacks and other security threats.
4. Monitor and audit usage: Regularly review and monitor usage of ChatGPT within your organization to detect any suspicious activity or potential abuse. Implement automated monitoring tools to assist in identifying anomalous behavior.
5. Encourage reporting of security concerns: Create a culture of openness and accountability, where employees feel comfortable reporting any security concerns or incidents involving ChatGPT or other AI tools.
6. Stay up-to-date on AI security: Continuously educate your organization on the latest developments in AI security and collaborate with industry peers to share best practices and stay informed about emerging threats.

By adopting these strategies, businesses can ensure that they are using ChatGPT and other AI-driven tools securely and responsibly, while maximizing the potential benefits these technologies offer.

© Copyright 2023, Cloud Security Alliance. All rights reserved.