Security Implications of ChatGPT: Preview of a Cloud Security Alliance Whitepaper

By Eric. E Cohen, CPA

Eric E. Cohen, CPA, is a technologist with a passion for collaboration toward the goal that "a piece of business information, once entered into any system, anywhere, never needs to be retyped as it moved through the business reporting supply chain." He's also a prolific author, engaged in virtually every effort to standardize accounting and audit data, a national expert to a wide variety of standards efforts, and co-founder of XBRL.

This article provides an overview of the forthcoming paper by the Cloud Security Alliance (CSA)[1] on the security implications of ChatGPT.[2] ChatGPT is an artificial intelligence (AI)-based chatbot that has gained widespread popularity and adoption, leading to the release of other similar products by competitors. While ChatGPT has many impressive capabilities, there are also concerns about its limitations, weaknesses, and potential security risks, and especially privacy risks. The CSA paper aims to explore these issues and their impact on the cybersecurity industry.

The topic of ChatGPT and its security implications is of great importance and relevance to the readers of *ThinkTwenty20*. As an emerging technology that in a very short period of time has already had a significant impact on the consumer market, ChatGPT is poised to disrupt many industries, including cybersecurity and accounting/auditing. The forthcoming CSA paper is a crucial resource for understanding the risks and opportunities associated with ChatGPT, and readers are encouraged to read and comment on the draft as it is made available.

As an emerging technology that in a very short period of time has already had a significant impact on the consumer market, ChatGPT is poised to disrupt many industries, including cybersecurity and accounting/auditing.

ChatGPT has already made waves in the media and popular culture, with its rapid adoption by users and the release of competing products. However, this popularity has also led to concerns about privacy, bias, and accuracy. For example, the recent ban of ChatGPT in Italy highlights the challenges of protecting user privacy in the age of AI. Similarly, the ongoing debates around the accuracy and reliability of ChatGPT's responses show the need for greater education and awareness about its limitations and weaknesses.

The CSA paper promises to provide a high-level overview of the implications of ChatGPT for the cybersecurity industry. While the paper is still in internal review and its contents subject to change, some key themes are likely to emerge. These include the need for greater transparency and accountability in the development and deployment of AI-based chatbots, the importance of educating users about the limitations and weaknesses of ChatGPT, and the potential impact of ChatGPT on businesses and industries. Readers can stay tuned for more updates on the CSA website and *ThinkTwenty20* blog and LinkedIn posts.

This article will cover three areas. It will begin with some background on ChatGPT and the new competitive AI environment that has disrupted the world since November 2022; it will then discuss the CSA paper. Finally, a call to the financial professional.

Background on ChatGPT

As noted, the Cloud Security Alliance (CSA)[3] will soon be publishing a draft paper for public comment on the topic of *Security Implications of ChatGPT*. The *ThinkTwenty20* community is strongly encouraged to read and comment on the draft when it is released, as well as to take a more active role in support of collaboration as CSA and other organizations continue to actively monitor and assess the benefits and risks of this emerging phenomenon. What is ChatGPT and what is this phenomenon?

On November 30, 2022, a new force was unleashed on the world, AI interaction with ChatGPT. AI, of course, is not new.

- AI has been a field of computer science for seventy years.[4]
- Apple showcased the vision they called Knowledge Navigator[5] in 1987. The vision was of a conversational personal assistant/chatbot, able to figure out the question rather than response to what was said, source and visualize resources, interact with the user and their contacts and much more.
- While the Apple Newton was not an incredible financial success, interactive AI with Apple's Siri, Google's Assistant and other voice-controlled personal assistants brought people closer to interacting with technology through voice exchange.
- The large firms have experimented with AI for contract review and journal entry work, amongst other projects, for many years.
- Companies like Mindbridge Analytics Inc.[6] have provided AI solutions for smaller audit firms; Mindbridge was founded in 2015.

As impressive as any of these milestones were, it was the release of OpenAI's ChatGPT that brought artificial intelligence to the masses and acted as the catalyst to the possibilities and challenges. ChatGPT did so in large part through the utter simplicity – by "putting the cookies on the bottom shelf where the kiddies can get at them," as one old radio preacher used to put it. ChatGPT made it so easy that anyone could do it, as long as they could get an account and the system wasn't overloaded with other requests. It did so by acting as an easily accessible "chat bot," a metaphor somewhat like instant messaging with a very smart, very fast, very considerate, but often fallible computer acting as friend and guide – the user types and the AI types back in an ongoing dialogue. Although anyone can use it, getting the best results from it does require some experience and skill.

Many users have been treating ChatGPT like a search engine, which ChatGPT is not; it is a “large language model” that leverages a large amount of training material to interact with a user “in a conversational way. ‘GPT’ stands for *generative pre-trained transformer*. The platform can take input from the user and information on which it is pre-trained and transform it to generate responses. Or, as ChatGPT notes when asked, “what do the individual words generative pre-trained transformer mean?”

> *Sure, I'd be happy to break down the meaning of each word in "Generative Pre-trained Transformer."*
>
> *Generative: This refers to the ability to generate new content, such as text, images, or audio. In the context of natural language processing (NLP), a generative model is one that can produce coherent and fluent sentences or paragraphs that resemble human writing.*
>
> *Pre-trained: This refers to a type of machine learning technique where a model is trained on a large dataset before being fine-tuned on a specific task. Pre-training allows the model to learn general features and patterns in the data that can be useful for a variety of downstream tasks.*
>
> *Transformer: This is a type of neural network architecture used in NLP that was introduced in the paper "Attention is All You Need" by Vaswani et al. (2017). Transformers use self-attention mechanisms to allow the model to focus on different parts of the input sequence, making them particularly effective for tasks such as language translation and text generation.*
>
> *So, when we put it all together, "Generative Pre-trained Transformer" refers to a specific type of machine learning model for NLP that is capable of generating new content, has been pre-trained on a large dataset, and uses a transformer architecture.*

“The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises and reject inappropriate requests.”[7] As it is not a search engine, but a conversationalist based on its training, ChatGPT would confidentially provide answers to just about any question that does not fall beyond the borders of what OpenAI’s guardrails for decency might consider appropriate. This means ChatGPT will confidently provide facts, references and other details that might be completely made up, although consistent with information in its training.

Nonetheless, more than just engage in conversational dialogue, ChatGPT has gone beyond most of our expectations, demonstrating its prowess in creating poetry and prose, organizing information, writing and debugging programming code, engaging in philosophical discussions, explaining and tutoring, and untold other capabilities. The author has personally engaged in theological discussion, British television trivia challenges, home repair advice, refreshing knowledge of a foreign language and translation of documents, creation of python programming code and XBRL instance documents, review of written materials, and a wide variety of other topics. ChatGPT has proven to be an amazing resource – but also has consistently made-up scholarly references, provided lyrics to songs it didn't know, fabricated biographies and otherwise given information that, if not treated with professional skepticism, could cause problems to the people who rely on the output.

The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises and reject inappropriate requests.

Nonetheless the rate of adoption by users of OpenAI's versions has been unprecedented, with the fastest-growing user base[8] of any consumer application in history. It has been featured heavily in the press, demonized and lauded in social media, and found its way into the plots of television shows. In March 2023, it was not only was the focus of an episode of the "irreverent" cartoon television series *South Park*, it was also given co-author credit[9] for an episode. So rabid has been the interest to gain access to ChatGPT, that hackers and scammers have flooded the market with ChatGPT extensions and apps that contain malware.[10]

This wild acceptance of OpenAI's ChatGPT has led to the acceleration of the release of other like products, both based on ChatGPT and licensed (such as Microsoft's use in Bing and Edge) and competitive, often with short-term detriment to the competitors[11] caught off-guard by the adoption of OpenAI's release and acceptance of ChatGPT. Microsoft licensed OpenAI's ChatGPT technology, as well as OpenAI's AI image creation tool DALL-E-2. Google released *Bard,*[12] Baidu released *Ernie*, Anthropic released Claude,[13] and numerous other alternatives have arisen.

With the quick rise and hype around the new tool, education around the limitations and weaknesses has been important to promote. All of the AI chatbot offerings[14] attempt to inform users that they are not magical front-ends to search engines, dealing only in the unvarnished and absolute truth. They "may occasionally" [ed: often] "generate incorrect information."[15] They "may occasionally produce harmful instructions or biased content." They may have limited knowledge, only as up to date as the materials they are trained on; the basic OpenAI ChatGPT has "limited knowledge of world and events after 2021."

Nonetheless, the news and social media are filled with users complaining about the perceived errors and problems. In addition, motivated users are trying to see if they can push ChatGPT beyond guardrails put in place to limit the impact of the known limitations. In addition, the training materials included personally identifiable information (PII) and copyrighted materials, which may be exposed during use. There is evidence, however, that despite attempts to protect users' privacy (concerns about issues such as privacy have led to ChatGPT being banned, at least temporarily, in Italy[16]). The people of Italy are finding ways to circumvent the protections in place for them, using virtual private network technology (VPN). Analysts note that driving users to VPN technology may increase the risk of privacy breaches as VPN solutions introduce their own incremental privacy risks.[17]

From an Enterprise point of view, management has been working overtime to develop policies related to the use of OpenAI and its alternatives. Employees are thrilled at the new efficiencies where OpenAI can create and review their code, help them in developing logical arguments and crafting presentations, organize activities, and accelerate and improve the quality of their processes. Despite guidance from OpenAI that anything typed may be reviewed by the AI trainers[18] or even become part of the training for future versions,[19] trade secrets have been exposed.[20]

Understanding that ChatGPT is only a subset of AI is important, as the accessibility of ChatGPT means we are trying to use it for tasks for which other areas of AI may be more suitable.

CSA Paper: Security Implications of ChatGPT

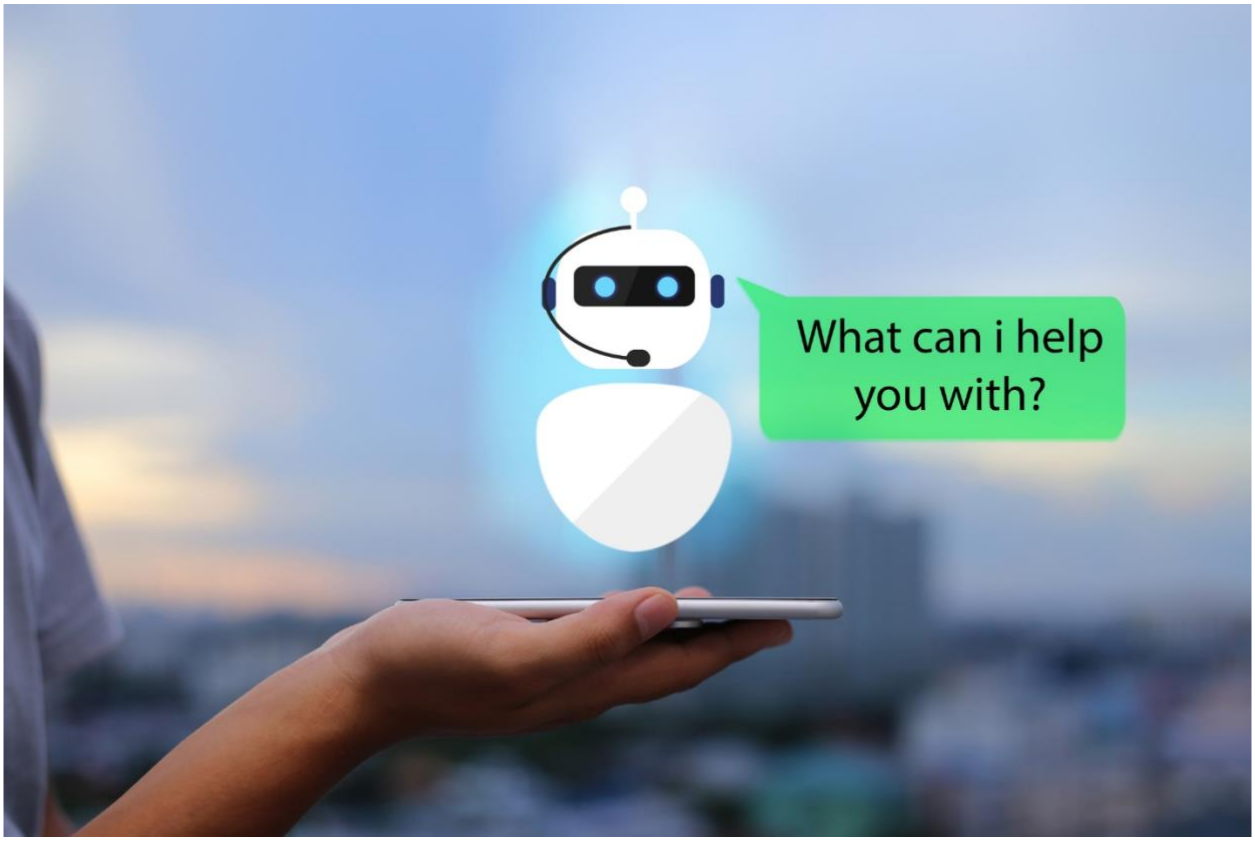
As a response to the phenomenon described above, the CSA provided an "open call to interested people to engage with them in the development of a paper that provides a high-level overview of the implications of ChatGPT in the cybersecurity industry." This paper is in final editing before a release for public comment. Quotes from the paper, as above, are based on an early internal release, and may not reflect the final draft. However, as the goal of this article is to provide background on the forthcoming CSA paper to help ThinkTweny20's readers to better assess it, preview quotes will be offered to help in that process. I will limit any larger excerpts to my personal contributions to the paper, although they are also subject to change or removal. Throughout this article, when I refer to "the author," I refer to myself; I am only one contributor of many to the CSA paper, fortunate to have collaborated with a stellar group of experts in security and AI.

A simplified outline (using my own words) of the paper:

- Introduction to ChatGPT
- How malicious actors can and are using ChatGPT
- How defenders can leverage ChatGPT
- How use of ChatGPT can be attacked
- How to help businesses cope with and use ChatGPT securely
- Things to know about generative AI – it has limitations and quirks of its own

- Conclusion
- References
- Appendices
- Additional resources

Let’s look at some of these sections.



Introduction to ChatGPT
The paper provides an overview of artificial intelligence and ChatGPT. Along with basics about what it is and how it works, the paper demonstrates that the place of ChatGPT in the larger world of AI, including GPT’s place as one of many generative foundational models.

Understanding that ChatGPT is only a subset of AI is important, as the accessibility of ChatGPT means we are trying to use it for tasks for which other areas of AI may be more suitable. As the old adage goes for financial professionals, we love spreadsheets, and so “Excel is the accountant’s hammer, and every task looks like a nail.” Even in the field of “generative AI,” the GPT model may not be the best model for the job.

This section ends with a brief discussion of other tools and solutions that shook up the security world. It is helpful to know that dissemination of both neutral and less benign tools has raised concern. While past performance is no promise of future results, the paper discusses a notable example of a hacking tool that evolved from concerns over becoming a pervasive enabler of exploits to a commercial provider of security solutions.

How Malicious Actors Can and Are Using ChatGPT

The second section focuses on "the potential risks associated with malicious actors using AI technologies to enhance their toolsets." ChatGPT has proven to be a powerful tool to write and review code. The good guys can use it to look for and remediate weaknesses in software patches, smart contracts and other code-based items of interest. That also means the bad guys can do the same. Threat actors can very efficiently analyze code and develop exploits. While the basic tools that unskilled malicious actors can use for breaching computers and networks have been broadly available for many years (leading to the term "script kiddies,"[21] ChatGPT has lowered the bar for customized malware.[22]

While many of the sections are aimed at an audience more IT technically oriented than the typical financial professional, each is an interesting read, with the most familiar to the reader being the potential of ChatGPT to facilitate phishing; in particular, "effortlessly craft[ing] legitimate-looking emils for various purposes." Our readers are probably aware that most scam emails will be filled with spelling mistakes and other inconsistencies; ChatGPT can customize, correct and create far more realistic communications.

How Defenders Can Leverage ChatGPT

The next section speaks about the steps management can take to leverage ChatGPT to defend against attacks. The author has appreciated the opportunity to hear from other participants in the paper's creation of their new efficiencies and successes with ChatGPT. Two of the sections of this part cover the ability to turn technical terms into English (or the language of the reader) and to explain updated security patches and change logs. We have all experienced having to "drink through a fire hose," perhaps most common when we try some new software or service and are faced with an agreement we must check before we can move forward. Having a tool that can summarize, look for red flags and otherwise find exceptions and explain it in easier-to-understand terms can take a lot of pain out of adopting and updating systems.

The author does not consider himself a programmer, although I teach students programming basics and a programming mindset. ChatGPT can do in seconds what used to take me a week when creating code. It can read a Python program and rewrite it in another programming language. Using simple textual prompts, ChatGPT can write regular expressions (regex) or otherwise take the complexity out of tasks.

In a space as rapidly changing and volatile as this – where almost every day a new opportunity or exploit or competitor to ChatGPT is named (at the time of this writing, an open-source attempt to make GPT-4 fully autonomous called Auto-GPT[23] has temporarily grabbed the attention of the press) – AI as the solution to keep up with cybersecurity in general and AI in particular makes a lot of sense – using "fire to fight fire," as it were.

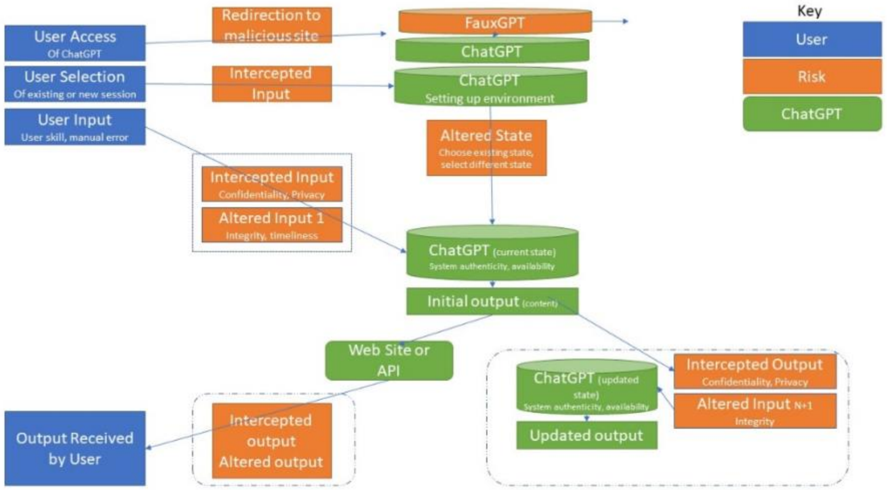
How Use of ChatGPT Can Be Attacked

As users, organizations and nations begin to rely on the output of tools like ChatGPT, attackers have more incentive to intercept or alter that output. There are many ways to affect the end results, from impacting the core of the training content it relies upon, to finding ways around the guardrails and protections, to intercepting and changing the sources, parameters, inputs and intermediate/final outputs between the user (whether a person typing directly into some manner of interface or using the application programming interface (API) to hook systems together.

In the diagram below, the author illustrates some of the touch points between the user and ChatGPT and the potential points for attack.

From my initial contribution, I note that "in the following illustration, many of the potential risk points to exploit the interaction between the user and ChatGPT are illustrated. This is highly simplified, but focuses on:

- Establishing a connection between the user and Cha.tGPT
- Starting a new conversation, or selecting an existing conversation to leverage the prior exchange.
- Entering user queries.
- Receiving and trusting that responses have maintained their integrity as the result of the query



During the intense and relatively brief time period we worked on drafting the paper, numerous changes to the ChatGPT environment have arisen. In late March, OpenAI announced initial support for "plugins" in ChatGPT. This enables ChatGPT to "interact with APIs defined by developers, enhancing ChatGPT's capabilities and allowing it to perform a wide range of actions."[24] However, it also opens even more access points for attack.

How To Help Businesses Cope with And Use ChatGPT Securely
The paper's next section includes initial guidance on security considerations while adopting ChatGPT. There is coverage, for example, of implementation leveraging Microsoft's Azure OpenAI Service for more control of the interactions and information.

Things to know about Generative AI – it has limitations and quirks of its own

The last section focuses less on attacks and more on the nature of generative AI technology itself. For example, the protections put into place by OpenAI, Microsoft or other solutions may sometimes limit the use. Many of us that have tried the Microsoft implementation of ChatGPT have asked what we may have thought was an innocuous question, but received feedback that "I prefer not" to continue the discussion from the chatbot. An AI expressing its "preference" for something is a state of anthropomorphism that may be troublesome, as helping users understand what AI is – and is not – is less stable when lines like that are blurred.

Conclusion, references, appendices, resources

Finally, the paper ends with a conclusion, references and resources, and appendices, including an interesting table illustrating areas of risks.

Call to Action

Will upcoming versions of ChatGPT take our jobs and lead to the extinction of mankind, as forecast by some, be the tool that will work hand in hand with tomorrow's professionals, or be a passing fancy? How can your organization or the ones you work with develop appropriate usage guidelines and policies? And how do we keep up with an area that is so rapidly changing?

This paper will be a great start and the CSA will appreciate your feedback. Plans for keeping this paper up to speed and beginning work on guidelines and policies are in place and can benefit from your help and support.

For more information, watch the CSA web site and the ThinkTwenty20 blog and LinkedIn posts. For a related panel presentation from CSA's March 28 *Cloud Threats and Vulnerabilities Summit 2023,* including the author as panelist, the video is available at the time of this writing from the CSA web site.[25]

Addendum:

This article was written during the writing of the paper and the final is now available at https://cloudsecurityalliance.org/artifacts/security-implications-of-chatgpt. My article does diverge a bit from the paper… but more to come. EEC

---

[1] https://cloudsecurityalliance.org/.
[2] https://chat.openai.com/.
[3] https://cloudsecurityalliance.org/.
[4] https://www.livescience.com/49007-history-of-artificial-intelligence.html.
[5] https://www.youtube.com/watch?v=umJsITGzXd0.
[6] https://www.mindbridge.ai/news/mindbridge-founded-to-prevent-the-next-bernie-madoff-using-artificial-intelligence/.
[7] https://openai.com/blog/chatgpt.
[8] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
[9] https://en.wikipedia.org/wiki/Deep_Learning_(South_Park), https://www.imdb.com/title/tt27035146/?ref_=ttep_ep4.
[10] https://www.digitaltrends.com/computing/chatgpt-extensions-apps-could-be-malware/.

---

[11] Google: https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/ Baidu: https://fortune.com/2023/03/16/baidu-ernie-bot-chatgpt-china-ai-share-price/.
[12] https://bard.google.com.
[13] https://www.anthropic.com/index/introducing-claude.
[14] See also Google's Bard and the help page at https://support.google.com/bard/answer/13275745?hl=en.
[15] https://chat.openai.com/chat.
[16] https://www.reuters.com/technology/italy-lift-curbs-chatgpt-if-openai-meets-demands-by-end-april-data-protection-2023-04-12/.
[17] https://techround.co.uk/news/after-italian-government-bans-chatgpt-vpn-searches-skyrocket/.
[18] https://help.openai.com/en/articles/6783457-what-is-chatgpt.
[19] https://help.openai.com/en/articles/7039943-data-usage-for-consumer-services-faq.
[20] https://adguard.com/en/blog/samsung-chatgpt-leak-privacy.html see also https://www.cnn.com/2023/04/06/tech/chatgpt-ai-privacy-concerns/index.html.
[21] https://en.wikipedia.org/wiki/Script_kiddie.
[22] https://www.recordedfuture.com/i-chatbot.
[23] https://github.com/Significant-Gravitas/Auto-GPT.
[24] https://platform.openai.com/docs/plugins/introduction.
[25] .

* * *