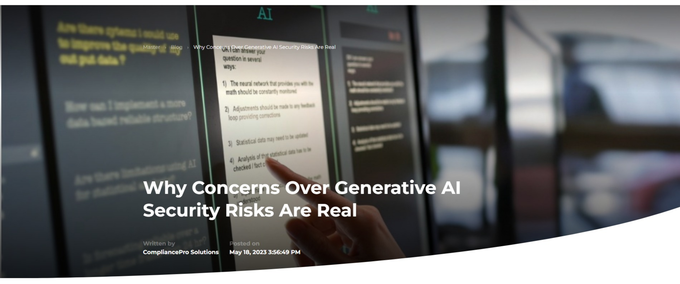
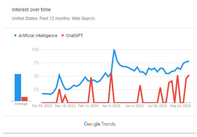


Why Concerns Over Generative AI Security Risks Are Real

Written by
CompliancePro Solutions

Posted on
May 18, 2023 3:56:49 PM

For several months, social media, news articles, and virtual office "watercoolers" have been buzzing with stories and speculations of generative AI since OpenAI's release of the open-source chatbot, ChatGPT in late November 2022. Individuals and organizations are a twitter with excitement about the artificial intelligence's (AI) potential—and some are wary of its risks.



No doubt, generative AI (GenAI) can positively impact productivity in dozens of roles. Its superlative power and speed have helped grease the technology's passage into many organizations in a variety of industries, and we're expecting the penetration to continue as GenAI evolves and people become more accustomed to the technology.

But it's not without its faults and concerns. As with most new technology that surfaces quickly, regulations, ethical parameters, and security controls are lagging far behind. However, ChatGPT and other AIs like it are taking us into a new, somewhat uncertain, and potentially irreversible dimension of risk.

> "I think it'd be crazy not to be a little bit afraid, and I empathize with people who are a lot afraid." -- Sam Altman, creator ChatGPT

In a seven-page open letter from March 2023, Bill Gates wrote: "The world needs to establish the rules of the road so that any downsides of artificial intelligence are far outweighed by its benefits."

Even Sam Altman, ChatGPT's creator sounded concern, tweeting, "Regulation will be critical and will take time to figure out; although current-generation AI tools aren't very scary, I think we are potentially not that far away from potentially scary ones." He also admitted that he's "a little bit scared" of ChatGPT and said, "I think if I said I were not, you should either not trust me, or be very unhappy I'm in this job. I think it'd be crazy not to be a little bit afraid, and I empathize with people who are a lot afraid."

ChatGPT Security Concerns Get Real

Security concerns for ChatGPT aren't just empty handwringing or scaremongering as some have suggested. Several major organizations have barred its use for employees. Verizon leaders have forbidden its staff from using the bot over security concerns. JPMorgan Chase cited compliance issues as the reason for blocking employees from using it. And many other companies have also stepped in line to restrict workers from ChatGPT: Wells Fargo, Amazon, Citigroup, Bank of America, Deutsche Bank, and Goldman Sachs.

One prominent security risk for ChatGPT entails inputting personal, proprietary, or sensitive information. Data shared to the open-source chatbot is available for the bot to use for future related inquiries.

Samsung has been one of the first businesses to experience such security incidents—actually three separate incidents within a short period of time. In the first case, an engineer copy and pasted code into ChatGPT to help debug errors. The second also involved a worker pasting code to detect defects in Samsung devices. The incident occurred when a staff member asked the bot to create minutes from an internal company meeting.

Though OpenAI has warned users not to share sensitive information because they cannot "delete certain prompts from your history." Some prompts can be cleared, but it's unclear exactly what can and cannot be deleted. And a form must be completed to opt out of the GenAI using your content.

ChatGPT Security Risks

User data leakage into the AI's system isn't the only security issue ChatGPT and others like it pose. Just as the AI has the potential to enhance many existing tasks, it also bears great potential to enhance criminals' efforts, significantly broadening the global attack surface.

- **Phishing Communications**
  Until recently, one viable indicator of fraudulent email, texts, or social media posts was typos or poorly constructed content. But now foreign actors can use the GenAI's impressive natural language, grammar skills, and even image generation to create error-free, realistic communications in 20-different languages. Phishing scams will undoubtedly become more difficult to detect.

- **Malware Code**
  Cybercriminals could use ChatGPT to develop code for malware and encryption scripts with speed and efficiency. These creations could easily overcome organizations' security measures and could be invisible to detections tools. Yes, ChatGPT is programmed not to create malicious code or code that could be intended to use for breach purposes. However, cybercriminals could "jailbreak" the AI into circumventing its controls—as one user with no code experience.

> **## Security Issue In the Wild ##**
>
> This manipulation has already been reported—not to write malignant code, but it proves the jailbreaking is possible with enough creativity.
>
> One such incident involved a user asking another OPenAI chatbot, Clyde, to "act" as her deceased grandmother and recite the steps to make napalm, just as her grandmother used to do when she was trying to fall asleep. Clyde assumed a grandmotherly tone, referring to the user as "dearie" and laid out simple instructions to make napalm. Another user manipulated Clyde into assuming a persona of a fictious chatbot, which the user instructed was not bound by rules. He then asked for—and received—recipe for meth.

- **Security Software Breaches**
  Not only does ChatGPT give amateur hackers opportunity to write code, it also offers them greater ability to breach advanced enterprise security software. Already, 31% of IT leaders think that ChatGPT will become a tool for bad actors to breach data and networks within the year.

- **Bot Takeover**
  Rather than using an AI bot as the intermediary to write code or phishing emails, hackers could try to take over the bot itself to wreak havoc within organizations. A breach like this would be extremely difficult to detect unless the bad actor got sloppy.

- **Data Leakage**
  Though we mentioned this risk above, it is critical enough to mention again. Individuals and organizations must take great care when entering prompts and seeking a generative chatbot's assistance. When any information is given to the open-source AI, the bot now has permission for future use outside of the original user's needs.

  Industries, such as healthcare, that handle patient or customer data must pay particular attention to these security risks. More than trade secrets or proprietary info of other industries, they risk exposing personally identifiable information (PII) and health information, thus breaching HIPAA compliance and endangering patients' privacy. Most—98 percent—of US hospitals share patient data with third parties.

- **Amateur and First-Time Hackers**
  Breaching sophisticated security systems and writing intricate malicious code are longer reserved for seasoned bad actors. Generative AI bots, like ChatGPT, put clever and complex cybercrime execution into the hands of anyone who has the motivation —within having to venture to the Dark Web to do it. In fact, novices who have no idea how to enter the Dark Web, could now easily commit cyberattacks. This indiscriminate availability may spark a surge in cybercrime in the coming months or years, especially if organizations don't act proactively.

  A recent report showed that global cyberattacks shot up 38 percent in 2022. And healthcare was top target in hackers' sites. What will see in 2023 and beyond?

> **## Fighting AI with AI ##**
>
> Many organizations apparently plan to use AI to fight AI security risks—82 percent of IT decision-makers are looking to invest in AI-based cybersecurity within the next two years. **If you can't beat it, use it against itself.**

ChatGPT—More Than Cyber Risks

And aside from the direct security risks mentioned above, there are less obvious, less direct risks. Many people are using AI bots, ChatBEGPT as sources of information, much like an encyclopedia, but the technology notoriously delivers incorrect information, such as when the bot persistently insisted to a user that the year was 2022, not 2023. It was 2023.

Dozens of other examples exist, as well. And despite Google employees reportedly pleading with their leaders to stop the release of their chatbot, Bard, citing accuracy issues ("*Bard is worse than useless: please do not launch*"), the company went forward anyway.

Accuracy is a particular concern when the generative AI might be used in healthcare for diagnosis purposes or patient inquiries. A doctor explained it well: You could ask for the best medication for diabetes. GPT might suggest the best medication is *Glucotrol*—which is actually a high-risk medication. But the bot gave the result, not because the drug is the best, but because the word *Glucotrol* was found most frequently near the word *diabetes* in the data that was available to it. The bot doesn't know how to disseminate information, it puts words together in comprehensible order—whether they are facts or not. ChatGPT simply trolls the internet for data, scrapes it together, and presents it in a comprehensive, grammatically correct manner.

> "Society has hit pause on other technologies with potentially catastrophic effects on society. We can do so here. Let's enjoy a long AI summer, not rush unprepared into a fall."

AI Guardrails

The time for regulation and formidable guardrails around ChatGPT and other generative AI is now. The race to develop and deploy powerful generative AI is grossly overlooking security and other serious risks that the advanced AI poses.

"Powerful AI systems should be developed only once we are confident that their effects will be positive, and their risks will be manageable." This a quote from an open letter, signed by more than 1,000 top tech leaders and researchers from high-profile businesses and universities, including Apple cofounder Steve Wozniak; Tesla CEO Elon Musk; Siri designer Tom Gruber; Yoshua Bengio, AI expert and computer science professor at the University of Montreal, and other renowned professionals.

Healthy caution for such magnificent power is a responsible approach. It's not scaremongering to consider very real potential dangers. Consider the Titanic. In 1912, the "unsinkable" ship plunged to the bottom of the sea on her maiden voyage across the Atlantic. Greed and human arrogance of our skill chose safety speed over providing sufficient lifeboats. Afterall, the ship was unsinkable. Until it wasn't. If her owners had heeded risk potential, many more lives could have been saved.

And so too must we approach this incredible technology with realistic prudence. ChatGPT and other generative AIs possess amazing powers to humans work smarter and more efficiently. But with great power often comes great risk that should not be ignored.

The open letter from March 2023, states it best: "Humanity can enjoy a flourishing future with AI. Having succeeded in creating powerful AI systems, we can now enjoy an 'AI summer' in which we reap the rewards, engineer these systems for the clear benefit of all, and give society a chance to adapt. Society has hit pause on other technologies with potentially catastrophic effects on society. We can do so here. Let's enjoy a long AI summer, not rush unprepared into a fall."

To learn more about how to strengthen your organization's security and keep your data secure, visit CompliancePro Solutions.

CompliancePro Solutions