AI-HELP & HINDRANCE TO IT

BY JAY BORDEN

AI, Artificial Intelligence, and its related technology LLMs, Large Learning Models, are expanding into everything.

Cybercriminals also recognize the benefits of AI and use it to create new attacks, improve phishing messages, discover new vulnerabilities, find companies using software with known vulnerabilities and exploit them quickly before patches can be applied.

Cybercriminals have also found ways to circumvent or disable the guardrails in AI systems. In some cases, they made a copy of an AI system and removed the guardrails entirely.

On the plus side, cybersecurity teams are using AI to discover vulnerabilities, analyze alerts, and more to make things safer.

However, AI systems still suffer from the two problems we have written about before, hallucinating and being hypnotized.

Hallucinations refers to the AI systems giving incorrect answers. This can be due to it being trained on bad information. It can be as simple as the training information being incorrect. Or it may be that the training information is biased due to intentional or unintentional biases on the part of the people assembling the training materials.

A third issue is the AI system simply makes up the answer without any factual basis for it. It just makes it up. Hence the term hallucinates.

onebrightlycyber.com
(888) 773-1920

AI-HELP & HINDRANCE TO IT

CONTINUED

Now, how does a hallucinating AI system affect cybersecurity?

Searching for threats and vulnerabilities can be performed more quickly by an AI system than by a person or even multiple people.

The AI system is faster at analyzing data than humans. The data can be used to train the system. Using real data tends to be bias free. All this is good.

But what about hallucinations? If the AI system misses something or "decides" something isn't a threat, it can lead to a breach.

That's one aspect of hallucinating. The other is the AI system creates an alert for an imagined threat. Or decides something isn't a threat. Either one will lower the cybersecurity team's trust in the AI system. Every time this occurs, trust in the AI system is diminished. What happens if the AI system isn't trusted? People won't use it or will not rely upon it.

If the AI system misdirects the team by creating alerts for hallucinated threats, the team will waste time and effort. Both are in very short supply by cybersecurity teams. Misdirection may cause them to miss a real threat resulting in a breach. Then trust in the AI system will be reduced to zero, or close to it.

Sadly, the cause of hallucinations in AI systems is still not known. Until it can be determined and fixed, the use of AI for cybersecurity will not reach its full potential.

To learn all the ways we can help make your company and family safer, visit onebrightlycyber.com, contact OneBrightlyCyber at info@onebrightlycyber.com, or call (888) 773-1920.