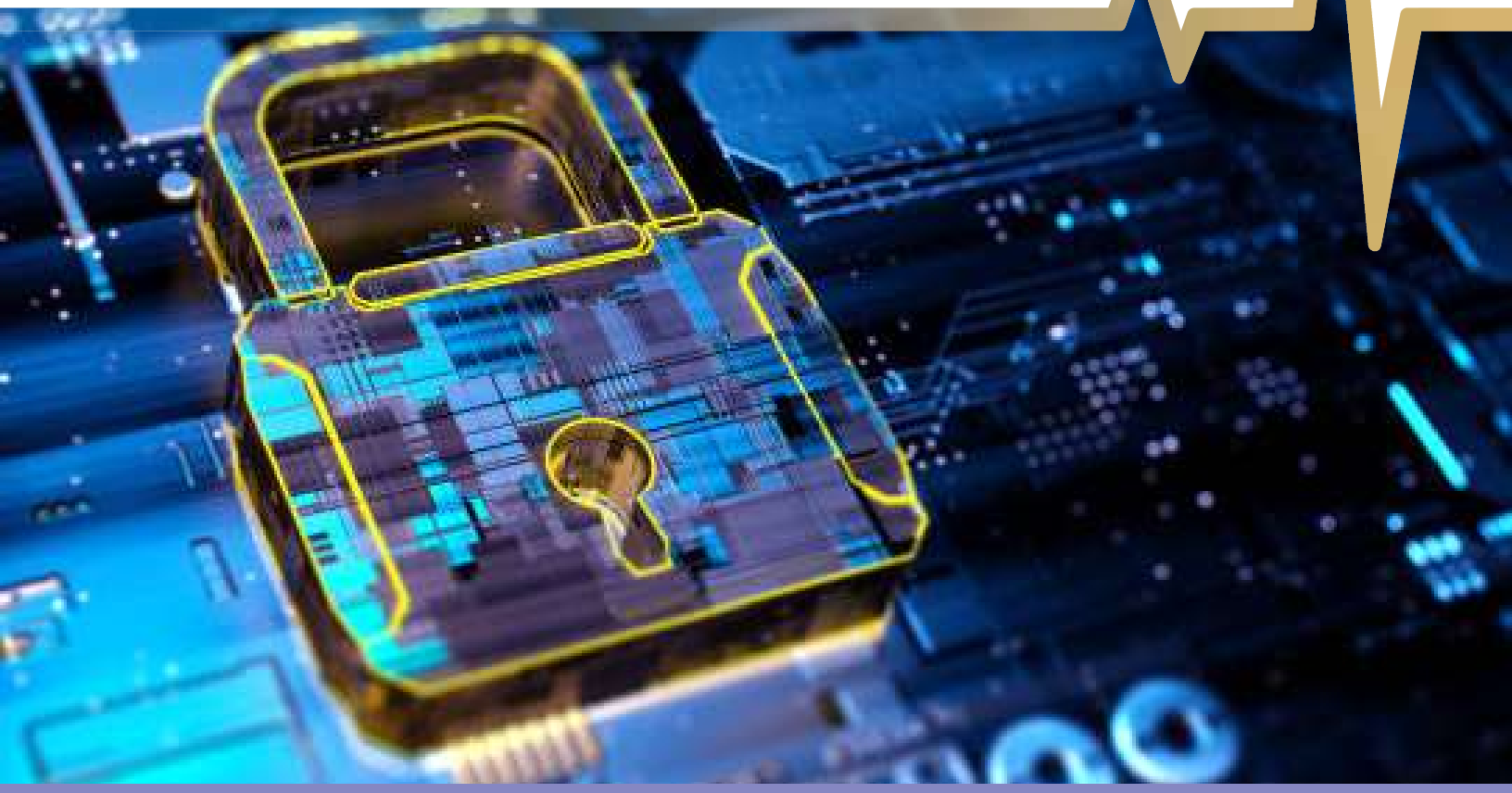
The Role of Artificial Intelligence in the Evolution of Social Engineering

Health-ISAC, Trinity Health



Health-ISAC™ Collaborating for Resilience in Healthcare

Trinity Health

TLP:WHITE

Key Judgements

- Publicly accessible AI is expediting evolution in social engineering.
- AI will help create false credibility for malicious actors.
- The risks of successful malign influence operations are increasing with the further development of AI.
- Both large-scale and small-scale social engineering efforts benefit from AI integration.

Executive Summary

Artificial intelligence is software that mimics authentic human intelligence. In 2023, interest in artificial intelligence skyrocketed after the public release of ChatGPT.

ChatGPT broke the record for engagement with the accruement of one million users in the first five days[1]. This time frame to reach one million users has never been observed before, underscoring the immense speed at which the service was adopted by the masses.

This increased growth was brought on by advanced capabilities. At its core, ChatGPT signified new code creation, writing, and cohesiveness capabilities in artificial intelligence models to the public. This new capability also sparked a renewed interest in criminal applications of large language models (LLMs) and other generative AI models such as the image generator DALLE-2.

The risk of artificial intelligence is quite a nebulous topic when looked at from a macroscopic lens, but when applied to a specific context such as risk to healthcare, it is possible to glean some notable insights. In a healthcare context, the criminal application of AI is especially concerning due to the potential incorporation of protected health information (PHI) into illicit endeavors, and the unique volatility of public trust in healthcare. As healthcare remains squarely in the center of the political spotlight, the two types of generative AI are being used to push divisive narratives. These are image generators and text generators, that when used in combination with each other, create compelling disinformation.

Throughout this report, a Senior Cyber Threat Intelligence (CTI) Analyst from Trinity Health gave their perspective on certain issues, adding supplementary value and insightful deductions from the research discussed herein.

Artificial Intelligence and Machine Learning

How threat actors interact with the target/victim is changing. Social engineering and victim exploitation techniques will improve and become much more effective due to generated text being less prone to the simple grammatical errors used for detection in the past. Delivery will likely become more streamlined as we move into the future. Much of this improvement will be fueled by access to artificial intelligence and machine learning capabilities.

Often in conversations, artificial intelligence and machine learning capabilities are considered one and the same. They are different from one another so this is how I would define each, we will consider machine learning as being very task oriented and pattern aware, such as Voice/ Speech imitation for example, whereas AI is a much more complex construct which combines machine learning, natural language processing, deep learning capability, and algorithms that allows for decision making capability paralleling what appears to be actual intelligence[2].

AI Weaponization in the Context of Social Engineering

Throughout history, social engineering attacks were planned and perpetrated by human threat actors employing various manipulation and persuasion techniques to establish rapport and exploit human vulnerabilities. If the threat actor is successful at convincing the victim the threat actor is trustworthy, believable, friendly, or helpful, the actor can appeal to those human emotions that allow for manipulation, the threat actor has essentially won[3]. Artificial intelligence and machine learning capability will serve to enhance their ability to persuade and manipulate. Manipulation campaigns that target human emotions and behaviors can benefit from the establishment of convincing, at least at the surface, back stories or pretext. Artificial intelligence will allow, should they feel it's necessary, actors the ability to create convincing "proof" of their authenticity as an individual the victim should trust. I state "should the threat actor feel it's necessary" because, most relatively sophisticated threat actors realize that given enough time most organizations can be exploited, at least initial exploitation without the use of 'A-Game' techniques, those saved for when the basic techniques fail.

CHATGPT which uses Reinforcement Learning with Human Feedback (RLHL) as well as other training methods to allow feedback associated learning process through human interaction. This conditioning process based on human feedback enables the AI to more effectively connect with individuals, adapt its approach based on real-time human feedback, and generate decisions that are more likely to succeed in manipulating human emotions and decision-making processes [4].

CTI Perspective

"Though different, I believe machine learning capability might be leveraged when needed by artificial intelligence as automated campaigns get more sophisticated and more detailed conversation input is required to achieve the desired objective, such as a successful social engineering campaign. Example of this might be a bot interacting with a victim on a social engineering campaign, when asked for some type of proof of identity, AI might be trained to leverage a particular machine learning capability that yields successful fake id's and present the ID as part of the AI social engineering campaign[5]. Are sophisticated attacks/campaigns like this happening presently? There are no signs at this time of this level of sophistication occurring in the e-crime spheres but as you know nothing is static in this field, evolution will occur, sophistication will develop, and we will need to bring our games up to deal with this new more sophisticated threat."

- Senior CTI Analyst, Trinity Health

The process of weaponizing AI for social engineering involves strategic steps to optimize its effectiveness. These steps include automating interactions, defining the AI's identity and role, training the AI using target-specific data, and refining its social engineering capability through customizable spear-phishing efforts and detailed pretexting[6]. When optimized, AI can create convincing material suitable for successful phishing campaigns or, even act as the threat actor itself, suggesting malicious links to its users[7].

In early 2023, an influence as a service vendor abusing social media to propagate manufactured narratives broke from the French news outlet Forbidden Stories. The elusive organization was named Team Jorge and was discovered to be using the mysterious Advanced Impact Media Solutions (AIMS) platform. This platform, not able to be found through conventional search engines such as Google or Bing, automated mass fake-account creation and used AI to create fake posts at scale. The utility of this platform lies in the ability to make pre-manufactured narratives seem genuine. According to Team Jorge, this method resulted in the election of the political candidate they backed in 27 of 31 cases [8]. AIMS exemplifies the difference between modern-day social engineering versus the campaigns of the past and the effective presence of AI in the disinformation workflow.

Disinformation Case Study: The 'Do So!' Movement

Disinformation is false information created as part of a larger agenda to socially manipulate a target population. A clear distinction between authentic beliefs in factually incorrect information and disinformation is the deliberate manufacture of the latter as opposed to the authenticity of the former. As the world population becomes increasingly engaged in the rapid consumerism of social media and internet news outlets, disinformation's efficacy as a psychological tool to strategically sew discord increases. By discrediting social opposition to a given agenda with carefully crafted information that is psychologically designed to appeal to the majority of a target population, disinformation can quickly gain genuine support, thus exponentially propagating it. When done correctly, disinformation can yield staggering results when the target audience runs with it.

CTI Perspective

"In my experience, when asked how many "zero day" exploitations cybersecurity professionals have seen in their career, the usual response is a few, a couple, a handful or some similar response. I would think there will be activity groups that will develop their skills and capabilities and will make a name for themselves within the AI/ML arena. We are at the early stages of a novel capability for the threat actors, but our defenses will continue to mature and our knowledge and experience dealing with this threat will allow us to compete effectively in my opinion."

- Senior CTI Analyst, Trinity Health

At first glance, this line of reasoning may appear a tad abstract with little grounding in reality, but that couldn't be farther from the truth. Large-scale disinformation efforts in the form of influence as a service and social media marketing have yielded concerningly effective results. Mainstream attention to these vendors began when the activities done by a subsidiary of the Strategic Communication Laboratories (SCL) called Cambridge Analytica were made public. Cambridge Analytica collected massive amounts of data from Facebook in the form of personality data points. These data points were then used to create aggregate personalities of different demographics.

Using these approximations of the average citizen, Cambridge Analytica was able to accomplish remarkable things, such as getting many of the Afro-Carribean kids in Trinidad to abstain entirely from voting by creating a fake political campaign advocating for voting apathy called "Do So!"[9].

In 2018, Cambridge Analytica disbanded, but the method of weaponizing mass data collection and private disinformation manufacturing remained.

Healthcare Specific Social Engineering and Content Farms

Aside from sophisticated large-scale propaganda efforts, there have been other, more amateur instances of AI-enabled disinformation being spread autonomously. Researchers have observed websites mimicking news sites that rapidly post AI-generated news articles, also known as AI-content farms, have begun to pop up in 2023. These content farms have been observed reporting conspiracy theories meant to sew public distrust in healthcare. A specific case of AI-generated disinformation attacking healthcare involved the operators of the known AI content farm, County Local News, attempting to use AI to generate phony news articles propagating the Vaccine Genocide conspiracy theory[10].

The farm, identified by newsguard, was allowing artificial intelligence to write and publish articles on the news site with little to no oversight. This is shown through the publication of an AI generated ethical complaint as a news article. The objection illuminated some truth about the operators of the website and their intentions. The objection to writing the prompt about vaccination genocide indicates that human operators attempted to get the AI to write phony news articles incorrectly tying deaths to Covid vaccinations.

The large scale at which the site was publishing new articles indicates there may have been intentions to falsely promote the news outlet through search engine optimization (SEO), and by extension, promote healthcare related disinformation.

Audiences being unable to distinguish between what's real and what is AI-generated will likely be the key to a successful campaign.

CTI Perspective

"Disinformation and propaganda campaigns are becoming more complex and more capable of achieving the desired goal. Actors such as Doppelganger are leveraging large social media presence to alter viewpoints in accordance with their, in this case, political agenda.[11] Will e-crime actors begin to use many of the same techniques to more readily achieve their goals, in my opinion, no doubt they will leverage the capability as it becomes available. AI used in a defensive manner to detect and alert on what may be an AI-generated campaign attempt, in my opinion, will be much more prevalent and will play an important role in detecting social engineering campaigns moving forward."

- Senior CTI Analyst, Trinity Health

Conclusion

The convergence of artificial intelligence, machine learning, and social engineering marks a pivotal juncture in healthcare cybersecurity. The capabilities of artificial intelligence and machine learning to enhance social engineering tactics used against our healthcare organizations are significant. An understanding of the ethical use of artificial intelligence and machine learning as a defensive strategy, coupled with proactive measures to fortify digital security, is essential in safeguarding against the adverse impacts of AI-driven social engineering exploits. Vigilance, education, and collaboration within the cybersecurity community are of great value in ensuring the responsible and ethical use of artificial intelligence techniques within healthcare organizations. Healthcare cyber defenders need to understand that while our defensive capabilities are improving the same can be said of our adversaries, we must be alert and ready to respond to social engineering attacks bolstered through the use of artificial intelligence.