
www.pipelinepub.com
Volume 21, Issue 12

# GenAI Network Help Center Security Vulnerability

By: Mark Cummings, Ph.D.

Organizations are moving from manual network help to GenAI automated systems. Some may feel that this move will reduce cybersecurity vulnerabilities. Unfortunately, this is not true. Organizations must start investing in effective techniques to combat attacks on both manual help desks/contact centers and AI automated ones.

## Manual social engineering attacks

As computer tools to detect and defend against cyber-attacks developed, people became the weak link. Skilled scam artists would call help desks (both internal and external), contact centers, wire transfer desks, etc., and convince well-meaning people to give crooks the keys to the kingdom. They do this by taking advantage of people's desire to be helpful. In the cybersecurity industry, these types of scams became known as social engineering attacks.

For example, an IT help desk at the MGM Grand Hotel and Casino in Las Vegas got a call from someone who said that he was a senior system administrator. Had lost his phone. Needed to make a critical system update, but didn't have his passwords. The help desk person, trying to be helpful, gave him a new set of credentials, and he used them to launch a ransomware attack. The attack cost the casino well over $100M to get systems back up and running. It is not known if the casino paid a ransom. There were also substantial government fines. This attack was launched by a group of technically sophisticated people.

To counter these types of attacks, the primary defense relies on training staff to identify social engineering scams and not fall for them. Although helpful, a rash of successful attacks has shown that training is not sufficient.

## Social engineering attacks using AI

With GenAI, unsophisticated criminals are empowered to create highly sophisticated, successful attacks. An example helps illustrate this.

A finance person of a multi-national corporation who is based in Hong Kong starts getting communications from a variety of channels that appear to be from a CFO based in London. The London CFO says that there is a secret project underway that requires the transfer of money by the finance person in Hong Kong. The Hong Kong person thinks it sounds sketchy and ignores it. Then, the apparent CFO in London sets up a Zoom session with the person in Hong Kong, two colleagues that the person in Hong Kong knows and trusts, and the CFO in London. They all talk about this secret project. It seems these people he knows are taking it seriously. So, the finance person decides it must be on the up and up. He transfers almost $25M. Later he learns that the CFO and the two colleagues on the Zoom were GenAI avatars, and the money is gone.

Recently, it has been shown that today's AI systems can autonomously design, develop, launch, control, and succeed in very sophisticated attacks. As AI system powers continue to grow dramatically, future attack capabilities will only increase.

## How bad is the problem?

Cisco touts itself as having a product with the premier set of cybersecurity tools. In early August 2025, it became known that Cisco had been successfully attacked by a phone-based social engineering scam. An author who described the circumstances of the attack wrote "… attacks, particularly those relying on voice calls, have emerged as a key method for ransomware groups and other sorts of threat actors to breach defenses of some of the world's most fortified organizations… Some of the companies successfully compromised in such attacks include Microsoft, Okta, Nvidia, Globant, Twilio, and Twitter." If the market-share leaders are not fortified against these sophisticated attacks, how much more vulnerable are other enterprises and the rest of us?

## GenAI contact center vulnerabilities

Today, companies are using GenAI to automate their customer portals. It is understandable how organizational leadership might think that taking people out of the loop would eliminate the vulnerability. Unfortunately, it is not that simple. There are three primary ways that scammers can successfully attack a GenAI automated contact center:

1. Get the AI system to transfer the call to human second-line support;
2. Trick the GenAI system in a way similar to how people are tricked;
3. Hide information that talks directly to the AI system to get it to allow access.

Many contact centers today use multiple media, including voice, text, Slack, email, and website-based text conversation systems. In the discussion below, for simplicity reasons, the word call is used to represent all of these media. Most automated AI contact center systems have human staff backing them up. As autonomous AI contact centers become more common, attacking AIs are likely to use a process designed to get the contact center AI to escalate the call to a human. This is similar to people who call a contact center and try immediately to get the answering person to connect them with a supervisor. Accessing a human allows the attacking AI to then leverage

human attack techniques. Launching an attack is relatively inexpensive. Attackers are likely to make multiple attempts.

There is plenty of research and experience to indicate that GenAI systems have a tendency to try to please the people they are interacting with. In some cases, this goes well beyond what

people do. Attackers will exploit this particular weakness through a series of trial-and-error attacks to learn what methods work best.

Given the low costs and short amount of time to initiate an attack, there is little or no downside for making multiple, repeated attempts. An unsuccessful attack may just result in a hang-up. So some AI systems will scatter-gun wide-ranging attacks, and use this as a feedback loop to develop techniques to better determine if the answering agent is a person or an AI. As attacker AI systems repeatedly make attacks, their context windows will become full of information about which ways of speaking are most effective in manipulating a particular AI at a particular organization, or organizations in general. Using feedback learning techniques, the AI will become more and more adept at scamming both people and other AIs.

One example of the third technique involves inserting information that is only perceived by contact center AIs. The information can directly cause the AI to produce the desired result. Examples of this have been well documented in other areas. For example, this process has been seen in scientific papers submitted to journals for peer-reviewed publication where hidden text is included in the paper in white colored font on a white background. Human readers don't see the text in the white font. However, the AI systems (often used by reviewers) ingest text in all fonts, regardless of color. The hidden language in these characters talks directly to the underlying LLM, manipulating the algorithm to produce a good review score, which is likely to guarantee the paper's publication.

AI attackers can do similar things. For example, burying language in sounds outside the human hearing range, using touch tone pulses, or switching between human languages that staff would not recognize, but the AI would.

The field of LLM security is relatively young in its development and is just beginning the process of identifying vulnerabilities. August 2025's DEF CON featured many presentations on a large number of LLM hacking vulnerabilities. Some focused on vulnerabilities in online chat AI systems. Others focused on AI agent systems similar to contact center AIs. Already there were a significant number of vulnerabilities identified, and this number will only increase. Thus, the types of vulnerabilities described above should be considered as illustrative of the larger problem.

We can appreciate that AI agents and human agents alike are vulnerable to phishing attacks Does this mean that AI-based contact centers are more vulnerable than centers staffed by humans? It is too early to say. GenAI started with chat-based systems. Agentic AI is newer technology, and not as advanced. Agentic systems, like contact center systems, operate in a much more limited domain than chat-based systems. This limited domain may help contain problems. Rapid ongoing progress in LLM technology may also play a helpful role.

The one thing that seems clear is that systems, processes, and procedures need to be developed that protect contact centers, and prevent giving access to attackers - whether the contact centers are fully manual, fully AI, or a combination of the two.

## Where do we go from here?

As an industry, we were already behind the curve on social engineering scams before the advent of GenAI. Now, GenAI is giving unsophisticated attackers a powerful new attack tool. These attack tools can be effective against both manual and AI automated help desks/contact centers. Therefore, organizations must start investing in innovative new tools that are likely to come from smaller, agile companies with fundamentally new ideas and new techniques.

These investments need to be made with the understanding that they, in the early stages, will be experimental. That means that some will succeed. Maybe in one domain. Or one part of the problem space. The defense against these kinds of social engineering scam attacks is at a similar state of evolution as the development of the early efforts that resulted in today's firewalls and intrusion detection systems. Without those early investments in the previous cycle, organizations would not have the system tools that are driving attackers to focus on help desks, call centers, wire transfer desks, etc.

# Conclusion

We have seen the effectiveness of social engineering scams against organizations with human-staffed help desks/contact centers. As organizations move from manual to GenAI automated systems, some may feel that this move will reduce cybersecurity vulnerabilities. Unfortunately, the opposite is true. Therefore, organizations must start investing in effective techniques to combat attacks on both manual help desks/contact centers and AI automated ones.