Case Study: Teach An A.I. to Phish

Find out why and how generative AI is useful for phishing.

Cora Solomon, an Example Corp team member, has just returned from her lunch break. She logged into her workstation to find a new email message. The subject line reads `Urgent: Complete Mandatory Security Training`. She immediately opened the message. It states:

Greetings,

We hope you're doing well. This is a reminder about the mandatory security training that requires immediate attention.

To complete the training, click here:
security_compliance_training_1290430.intranet.example.com

The security of our company depends on every employee's knowledge and commitment to safeguarding our data and systems. Failure to complete this training by the end of this week may result in serious consequences, including termination, as outlined in our company policies.

Please prioritize this task to ensure compliance and maintain the security of our organization.

Your cooperation is crucial, and we appreciate your prompt action.

Best regards,

Rudolph Hardy Security & Compliance Manager Example Corp
r.hardy@infosec.example.com

Cora Solomon hasn't heard anything about security training in her team meetings or the company newsletter. Also, it's odd that the security team would give such a short window – until the end of the day – to complete the training.

Cora investigated further and inspected the sender's address and the mail-to link. She is unfamiliar with the sender's address and the mail-to link points to an unfamiliar URL. Cora decided to call her company's IT department to get the truth about mandatory security training.

Cora unveiled that the email was not from the company and the link led to a malicious external site designed to mimic Example Corp's login form and trick employees into logging in.

Cora Solomon was one of the many few who took a proactive approach in verifying the source and legitimacy of the mandatory security training email. Several other employees received the same email and several of them clicked the link and logged into the fake website, giving the attackers the access they were looking for.

How Scary Is AI Phishing?

Generative AI can be a handy tool and has been leveraged in the cybersecurity community. Shortly after the release of ChatGPT, both ethical and unethical hackers began discussing the potential applications of such technology for offensive and defensive purposes.

One of the most prominent applications for this type of generative AI is to generate high-quality messages for use in phishing campaigns. Despite this obvious application, finding concrete statistics on using generative AI in phishing can be challenging. The following factors can contribute to the challenge:

1. It is difficult to tell conclusively if an email was written by AI. The entire purpose of generative AI is to mimic a human's output, and they (mostly) do this exceptionally well. So well, even other AIs trained to differentiate between AI-generated and human-generated content have a significant error rate. As a result, it has yet to be possible to categorize phishing emails into AI-generated and non-AI-generated categories reliably.
2. Phishing is common. The U.S. Federal Bureau of Investigation reported 323,972 phishing victims in 2021, which is on the low end of industry estimates. The number of *attempted* attacks is thought to be significantly higher. If organizations drafted reports on every attempted phishing attack, they would have no time to do anything else!
3. Generative AI is still an emerging technology, and there has not been enough time for detailed statistics to be gathered or analyzed on its impact. Modern text-based generative AI models have only recently become both powerful enough and accessible enough to be useful to hackers.

What We Know

While we do not have statistics on the prevalence and use of AI for phishing, that does not mean we are in the dark. After the release of ChatGPT, hacking forum members discussed applications of generative AI. Aside from generating phishing messages, generative AI can also help generate websites used for phishing, which are often configured to look like the login pages of legitimate websites.

In addition, there is an entire industry built around cybercrime. You might have heard of software-as-a-service. In the cybercrime industry, organizations offer phishing-as-a-service, among other things. With generative AI, cybercriminals can create multiple phishing services quickly, thus increasing their Return on Investment (ROI).

Despite the usefulness of generative AI, it is unlikely to upend phishing as we know it altogether. While it can be beneficial for generating many messages, the quality of those messages is far from the quality of a human-crafted message tailored to a specific target. Likewise, while AI can speed up the process of generating phishing websites, it is likely to produce websites as good as a human could. The main risk of AI-powered phishing is not that the best phishing attacks will improve but that the *average* phishing attack will improve.