Cybersicherheit im Zeitalter von ChatGPT: Willkommen Misstrauen

Bislang hatten Organisationen ihre größte Schwachstelle beim Kampf gegen Cyberkriminalität gut im Griff: die Mitarbeitenden wurden erfolgreich geschult und sensibilisiert. Aber nun nimmt mit AI-generiertem Social-Engineering-Betrug eine neue Welle Fahrt auf. Bis die Technologie reif ist, muss der Mensch den Wachhund spielen, findet Chester Wisniewski, Field CTO Applied Research bei Sophos und stellt drei Prognosen für die Zukunft auf.

Organisationen haben sich intensiv mit einer ihrer kritischsten Cybersecurity-Komponente auseinandergesetzt: den Mitarbeitenden. Sie begegnen der „Schwachstelle Mensch“ mit kontinuierlichem Training und vertrauen mittlerweile häufig darauf, dass die Nutzer zum Beispiel potenzielle Phishing-Attacken aufgrund sprachlicher Unregelmäßigkeiten oder falscher Rechtschreibung und Grammatik erkennen.

Aber KI-gesteuerte Sprach-und Inhaltsgeneratoren wie ChatGPT sind auf dem besten Weg, diese verräterischen Elemente aus Scams, Phishing-Versuchen und anderen Social-Engineering-Angriffen zu entfernen. Eine gefälschte E-Mail vom „Vorgesetzten“ kann dank künstlicher Intelligenz überzeugender als jemals zuvor klingen und die Beschäftigten werden es unzweifelhaft schwerer haben, Fakt und Fiktion zu unterscheiden. Im Falle dieser Betrügereien sind die Risiken von KI-Sprachtools nicht technischer Art. Sie sind sozialer Natur – und damit beängstigend.

Die spezifischen Risiken von KI-generierten Phishing-Attacken
Von der Erstellung von Blogbeiträgen und Codiercode bis zum Verfassen von beruflichen E-Mails – KI-Sprachtools können all das. Die Technologien sind geschickt darin, überzeugende Inhalte zu generieren und sie sind gespenstisch gut darin, menschliche Sprachmuster nachzubilden.

Auch wenn wir bislang noch keinen Missbrauch dieser Programme für Social-Engineering-Inhalte verifizieren konnten, vermuten wir, dass dieser unmittelbar bevorsteht. ChatGPT wurde bereits verwendet, um Malware zu schreiben und wir erwarten, dass kriminelle Akteure bald schadhafte Applikationen für KI-Sprachtools entwickeln werden. Aber: Bereits heute stellt KI-generierter Phishing-Inhalt einzigartige soziale Risiken dar, die die technische Abwehr unterlaufen!

KI-Phishing-Attacken: Kann jede E-Mail eine Gefahr sein?
Nehmen wir zum Beispiel KI-erzeugte Schadsoftware: hier können existierende Sicherheitsprodukte den Codiercode in Millisekunden analysieren und zuversichtlich als sicher oder schadhaft einschätzen. Maßgeblicher noch, Technologie kann Technologie entgegenwirken.

Aber in mit künstlicher Intelligenz erstellten Phishing-Nachrichten eingebettete Worte und Nuancen können von Maschinen nicht entdeckt werden – es sind Menschen, die diese Scam-Versuche als Empfänger interpretieren werden. Da KI-Instrumente in der Lage sind, anspruchsvolle und realitische Inhalte nach Bedarf zu produzieren, können wir uns also immer weniger auf Menschen als Teil der Verteidigungslinie verlassen.

Die Rolle des Menschen wandelt sich. Es wird heißen: Technologie gegen Technologie
Diese sich rapide entwickelnde Situation erfordert eine Neubewertung der Rolle des Sicherheitstrainings im Kampf gegen Social Engineering-Attacken. Auch wenn es bislang noch keine kommerzielle Anwendung gegen KI-generierten Betrug gibt, werden zunehmend Technologien als zentrales Werkzeug zur Identifizierung und zum Schutz Einzelner vor Maschinen-erzeugten Phishing-Angriffen dienen. Menschen werden zwar auch noch eine Rolle spielen, aber nur eine sehr kleine.

Drei Prophezeiungen für das ChatGPT-Zeitalter
Auch wenn wir noch im frühen Stadium dieses neuen KI-Zeitalters sind, lässt sich bereits absehen, dass KI-erzeugter Phishing-Inhalt ein enorm wichtiges Thema für die Sicherheitsstrategie von Unternehmen wird. Folgende drei Prognosen, wie ChatGPT als Werkzeug für Cyberkriminalität eingesetzt werden könnte und welche Sicherheits-Antworten sich daraus entwickeln, scheinen dabei am wahrscheinlichsten:

1. Komplexere Nutzer-Authentifizierung wird nötig werden.
Maschinen sind sehr gut darin, wie Menschen zu klingen, daher ist es nötig, in Unternehmen neue Authentifizierungs-Möglichkeiten zur Verfügung zu stellen. Das bedeutet: jede Kommunikation, die Zugang zu Unternehmensinformationen, Systemen oder monetären Elementen betrifft, *muss* komplexere Formen der Nutzerauthentifikation einfordern. Bestätigung via Telefonanruf wird wahrscheinlich die geläufigste Methode zur Verifizierung dieser Arten von E-Mails oder Nachrichten werden. Unternehmen könnten zudem ein geheimes Tagespasswort einsetzen, um sich gegenüber anderen Instanzen oder Individuen zu identifizieren.

Einige Finanzinstitute operieren bereits so. Welche Form der Verifizierung auch immer zum Einsatz kommt, es ist entscheidend, dass die Methode nicht einfach von Angreifern genutzt werden kann, wenn diese Nutzer-Zugangsdaten kompromittiert haben.

Da die KI-Technologien sich rasend schnell entwickeln und immer weiter verbreiten, müssen Authentifikations-Methoden Schritt halten. Im Januar dieses Jahres hat Microsoft zum Beispiel VALL-E offengelegt, eine neue KI-Technologie, die eine menschliche Stimme nach dreisekündiger Audioaufnahme klonen kann. In der nahen Zukunft wird der Telefonanruf als Authentifizierungsanforderung also wahrcheinlich auch nicht mehr genügen…

2. Legitime Nutzer verwässern die Sicherheitswarnungen: alle oder keiner.
Viele Nutzer setzen ChatGPT ein, um schnell beruflichen oder werblichen Inhalt zu produzieren. Dieser legitime Gebrauch von KI-Sprachtools macht Sicherheitsantworten komplizierter, da es schwieriger wird, kriminelle Beispiele zu identifizieren.

Beispiel: Nicht alle E-Mails, die ChatGPT-erzeugten Text enthalten, sind schadhaft, wir können sie also nicht alle generell blockieren. Das führt bis zu einem gewissen Grad zur Verwässerung unserer Sicherheitsreaktion. Anbieter von Sicherheitslösungen könnten als Gegenmaßnahme „Vertrauens-Punkte“ oder andere Indikatoren entwickeln, die die Wahrscheinlichkeit beurteilen, dass eine Nachricht oder E-Mail zwar KI-generiert aber dennoch vertrauenswürdig ist. Auch könnten sie KI-Modelle trainieren, um mit künstlicher Intelligenz erstellten Text erkennen zu können und ein „Vorsicht“-Banner auf benutzerorientierten Systemen zu platzieren. In bestimmten Fällen könnte diese Technologie Nachrichten aus der E-Mail-Inbox des Mitarbeitenden herausfiltern.

3. KI-generierter Betrug wird interaktiver werden.
Ich erwarte, dass KI-Sprachprogramme in Zukunft noch weitaus stärker von Kriminellen interaktiv genutzt werden, um Phishing-E-Mails oder Einmalnachrichten zu produzieren.

Betrüger könnten diese Technologie einsetzen, um Individuen via Echtzeit-Chat zu manipulieren.
Nachrichtendienste wie WhatsApp, Signal oder Telegram sind durchgängig verschlüsselt, diese Plattformen sind daher nicht in der Lage, betrügerische oder KI-generierte Nachrichten in privaten Kanälen zu filtern. Das könnte sie sehr attraktiv für Betrüger machen, die auf diesen Plattformen ihren Opfern auflauern.

Auf der anderen Seite könnte diese Entwicklung Organisation dazu bringen, ihre Sicherheitslösungen zu rekonfigurieren. Womöglich werden Filtertechnologien auf individuellen Mitarbeiter-Endgeräten nötig.

Bis neue Technologien greifen gilt: jeder Kommunikation mißtrauen
KI-Sprachwerkzeuge stellen wesentliche Fragen für die Zukunft der Cybersicherheit. Es wird schwieriger, herauszufinden, was real ist und zukünftige Entwicklungen machen das nicht leichter. Technologien werden die primäre Waffe gegen KI-getriebene Cyberkriminalität sein. Aber jetzt müssen die Mitarbeitenden ran und lernen, jeder Kommunikation zu misstrauen. Im Zeitalter von ChatGPT ist das keine Überreaktion, sondern eine kritische Antwort.

Social Media von Sophos für die Presse
Wir haben speziell für Sie als Journalist*in unsere Social-Media-Kanäle angepasst und aufgebaut. Hier tauschen wir uns gerne mit Ihnen aus. Wir bieten Ihnen Statements, Beiträge und Meinungen zu aktuellen Themen und natürlich den direkten Kontakt zu den Sophos Security-Spezialisten.
Folgen Sie uns auf und

LinkedIn: https://www.linkedin.com/groups/9054356/
Twitter: @sophos_info

Pressekontakt:

Sophos
Jörg Schindler, PR-Manager Central & Eastern Europe
joerg.schindler@sophos.com, +49-721-25516-263

TC Communications
Arno Lücht, +49-8081-954619
Thilo Christ, +49-8081-954617
Ulrike Masztalerz, +49-30-55248198