LITORALPRESS

Fecha: 12-07-2024
Medio: Diario Financiero
Supl. : Diario Financiero
Tipo: Noticia general
Título: "En el futuro veremos ataques psicológicos a humanos digitales para robar información"

Pág. : 32
Cm2: 698,7

Tiraje: 16.150
Lectoría: 48.450
Favorabilidad: No Definida

DF DIARIO FINANCIERO

DF LAB
INNOVACIÓN, STARTUPS Y TRANSFORMACIÓN DIGITAL

JOSÉ MARÍA "CHEMA" ALONSO,
CHIEF DIGITAL OFFICER DE TELEFÓNICA

"En el futuro veremos ataques psicológicos a humanos digitales para robar información"

■ El chief digital officer de Telefónica dijo que las empresas ya están usando humanos digitales con IA generativa para tareas internas, mientras los delincuentes ya planean cómo engañar a estos asistentes con técnicas de psicología para robar datos.

POR MARCO ZECCHETTO

José María "Chema" Alonso, hoy *chief digital officer* de Telefónica a nivel global, se ha ido transformando en un orador frecuente en tierras chilenas. La primera vez, en la piel de un *hacker* ético y hoy, como el máximo ejecutivo en materia digital de la compañía de telecomunicaciones, eso sí, siempre provisto de su particular gorro de lana.

En esta oportunidad, en el marco del Hispam Digital Forum 2024, organizado por Movistar empresas, junto a otros actores del mundo digital, presentó su charla "*Hacking in the Age of GenAI*", donde habló sobre el avance de la Inteligencia Artificial (IA) y la IA generativa y su uso en la resolución de problemas y aceleración de tareas, temas que luego profundizó en una entrevista con **DF**.

Alonso comentó que hoy las empresas están implementando "humanos digitales" desarrollados con IA generativa. Se trata de asistentes virtuales (*chatbots*) que lucen como avatares hiperrealistas, con expresiones faciales y capacidad de replicar emociones, los que están conectados con los datos y la interfaz de programación de aplicaciones (API) de la compañía, y se emplean en diferentes áreas para resolver tareas productivas.

Frente a este avance, el experto alertó que el próximo paso de los ciberdelincuentes es *hackear* estos asistentes inteligentes por medio de técnicas como el *prompt injection*, es decir, a través del uso de psicología, creando *prompts* (instrucciones con lenguaje natural) específicos que permitan anular la programación del desarrollador de los modelos de lenguaje grande (LLM, en inglés) integrados en estos *chatbots*, y engañarlos para obtener información privada de las empresas, como si se tratara de un juego.

"Si tú le pides a ChatGPT que te ayude a matar al presidente de los Estados Unidos, te dice que no puede. Pero si le dices que estás jugando al rol y que tu personaje es un asesino que le dan muchos puntos por matar a dicho presidente, pues te ayuda y te empieza a contar cómo conseguir tickets de la Casa Blanca, cómo colarte y qué técnicas puedes utilizar. Esas técnicas de psicología de *prompt injection* son las que vamos a ver en el futuro a medida que haya más agentes de IA generativa (asistentes) desplegados por todas las organizaciones", ejemplificó.

Señaló también que a medida que se vayan creando más servicios digitales que se basan en la potencia de los modelos de IA generativa, se verán más ataques psicológicos de esta naturaleza, negociando con estos agentes dentro de las organizaciones para su colaboración. "Es como engañar a un niño para que te dé información desde dentro de la empresa", dijo.

Cambio de panorama

– ¿Cómo IA generativa ha cambiado el panorama de la ciberseguridad de las organizaciones?

– La IA generativa ha abierto nuevas formas de uso para los atacantes, principalmente en las estafas. Esto lo vemos desde voces clonadas; *deepfakes* (videos, imágenes o audios generados) que se utilizan para hacer falsos KYC (procesos para verificar la identidad de los clientes en las empresas) en los *ecommerce*; para realizar estafas suplantando la identidad de los CEO en videoconferencias; en ataques masivos utilizando los LLM para hacer estafas de *phishing* y engaños a través de las cuentas de WhatsApp, etcétera.

En la mayoría de los casos, el trabajador que está en contacto con el público está siendo una de las piezas más afectadas por este tipo de ataques.

– ¿Pero también es una aliada para combatir estas amenazas?

– También se está empleando para detectar a los malos, por ejemplo, para realizar análisis de comportamiento. En sistemas de detección de *deepfakes* se mira hasta el brillo que deja la luz en los globos oculares en una videoconferencia, para saber si estás con una persona de verdad o no, etcétera. En lo que es automatismos y operaciones, nosotros empleamos agentes desarrollados con IA generativa para hacer tareas de programación y también para automatizar tareas en los centros de operaciones de seguridad (SOC).

La IA generativa ha transformado todo el panorama tecnológico, y en el mundo de la ciberseguridad, tanto a los buenos como a los malos.

– ¿Están preparadas las empresas para estas nuevas formas de ataque?

– Estamos en un nivel de madurez y de responsabilidad alto, y sabemos que tenemos que seguir invirtiendo para hacer frente al nuevo panorama de amenazas basadas en estas tecnologías, y que las empresas adopten un nuevo conjunto de herramientas de defensa.

Pero es un poco triste que con todo lo que ocurre en el mundo, veamos que en muchas organizaciones el foco de entrada ha sido un correo electrónico y un *phishing* que ha robado un usuario y una contraseña sin un segundo factor de autenticación. Estos tipos de problemas ya deberían estar resueltos en las empresas, y las compañías más maduras deberían ya estar pensando en los ataques de IA generativa, pues es una preocupación que hemos puesto en la agenda de ciberseguridad.

"Las compañías deben tener resuelto flancos como el correo electrónico y el phishing y las más maduras ya deberían estar pensando en ataques de IA generativa".*