OpenAI anunció el miércoles que ha desmantelado más de 20 operaciones y redes engañosas a nivel mundial que intentaron utilizar su plataforma para fines malintencionados desde principios de año.

Estas actividades incluyeron la depuración de malware, la redacción de artículos para sitios web, la creación de biografías para cuentas en redes sociales y la generación de imágenes de perfil mediante IA para cuentas falsas en X (anteriormente conocido como Twitter).

> *«Los actores maliciosos continúan adaptándose y probando nuestros modelos, pero no hemos observado pruebas de que esto haya resultado en avances significativos en su capacidad para desarrollar malware completamente nuevo o construir grandes audiencias virales»*, señaló la empresa de inteligencia artificial (IA).

También mencionaron haber frustrado actividades que buscaban generar contenido para redes sociales relacionado con elecciones en Estados Unidos, Ruanda, y en menor medida, India y la Unión Europea, sin que ninguna de estas redes haya logrado un alcance viral o audiencias sostenidas.

Entre estos esfuerzos se encontraba una operación realizada por una empresa israelí llamada STOIC (también conocida como Zero Zeno), que generaba comentarios en redes sociales sobre las elecciones en India, tal como fue revelado previamente por Meta y OpenAI en mayo.

Algunas de las operaciones cibernéticas destacadas por OpenAI incluyen:

- SweetSpecter, un grupo sospechoso de tener sede en China, que utilizaba los servicios de OpenAI para realizar reconocimiento basado en modelos de lenguaje (LLM), investigar vulnerabilidades, apoyar con scripts, evadir detección de anomalías y desarrollar nuevas herramientas. También intentaron sin éxito realizar ataques de spear-phishing a empleados de OpenAI para instalar el troyano de acceso remoto (RAT) SugarGh0st.

- Cyber Av3ngers, un grupo vinculado al Cuerpo de la Guardia Revolucionaria Islámica de Irán (IRGC), utilizó los modelos de OpenAI para investigar controladores lógicos programables.
- Storm-0817, otro actor iraní, empleó los modelos de OpenAI para depurar malware en Android diseñado para recolectar información sensible, utilizar herramientas de extracción de perfiles de Instagram con Selenium y traducir perfiles de LinkedIn al persa.

Por otro lado, OpenAI bloqueó varios grupos, incluidos A2Z y Stop News, que producían contenido en inglés y francés para publicarlo en diversos sitios web y cuentas de redes sociales.

> *«[Stop News] fue particularmente prolífica en el uso de imágenes. Muchos de sus artículos en la web y tuits estaban acompañados de imágenes generadas con DALL·E. Estas imágenes solían tener un estilo caricaturesco, con colores brillantes o tonos dramáticos para captar la atención»*, dijeron los investigadores Ben Nimmo y Michael Flossman.

Otras dos redes identificadas, Bet Bot y Corrupt Comment, utilizaban la API de OpenAI para generar conversaciones con usuarios en X y enviarles enlaces a sitios de apuestas, así como fabricar comentarios que luego eran publicados en X.

La divulgación ocurre casi dos meses después de que OpenAI prohibiera varias cuentas vinculadas a una operación de influencia encubierta iraní denominada Storm-2035, que empleaba ChatGPT para generar contenido enfocado en temas como las próximas elecciones presidenciales en Estados Unidos.

> *«Los actores malintencionados suelen utilizar nuestros modelos para realizar tareas en una fase intermedia de sus actividades: después de adquirir herramientas básicas como acceso a internet, direcciones de correo electrónico y cuentas de*

> *redes sociales, pero antes de desplegar productos finales como publicaciones en redes sociales o malware distribuidos a través de varios canales»,* explicaron Nimmo y Flossman.

La empresa de ciberseguridad Sophos, en un informe publicado la semana pasada, señaló que la IA generativa podría ser explotada para difundir desinformación personalizada a través de correos electrónicos dirigidos específicamente a ciertos usuarios.

Esto implicaría el uso indebido de modelos de IA para crear sitios web de campañas políticas, identidades generadas por IA de todo el espectro político y correos electrónicos dirigidos según los puntos de las campañas, lo que permitiría automatizar la propagación de desinformación a gran escala.

> *«Esto significa que un usuario podría generar desde material de campaña inocuo hasta desinformación intencional y amenazas maliciosas con solo una pequeña reconfiguración»,* comentaron los investigadores Ben Gelman y Adarsh Kyadige.

> *«Es posible asociar cualquier movimiento político o candidato real con el apoyo a una política determinada, incluso si no están de acuerdo. La desinformación intencional de este tipo puede hacer que las personas apoyen a un candidato que no desean o rechacen a uno que antes creían respaldar.»*