ChatGPT,
¿la próxima gran amenaza de ciberseguridad?

El papel de la inteligencia artificial en la ciberseguridad está creciendo. Un nuevo modelo de IA destaca las oportunidades y los desafíos, porque tiene el potencial de revolucionar muchos aspectos de nuestras vidas, incluida la forma en que abordamos la ciberseguridad. Sin embargo, también presenta nuevos riesgos y desafíos que deben gestionarse cuidadosamente.

De hecho, hasta ahora se pensaba que la forma en que la IA se puede utilizar en ciberseguridad es a través del desarrollo de sistemas inteligentes que puedan detectar y responder a las amenazas cibernéticas. Pero ¿puede haber algo más?

ChatGPT en el foco de la atención mundial

En noviembre de 2022, OpenAI, una empresa de investigación y desarrollo de IA, presentó ChatGPT (Transformador Generativo Preentrenado) basado en una variación de su modelo InstructGPT, que está entrenado en un conjunto masivo de datos para responder consultas. Interactúa de manera conversacional una vez que se le da una indicación detallada, admite errores e incluso rechaza solicitudes inapropiadas. Aunque solo está disponible para pruebas beta en este momento, se ha vuelto extremadamente popular entre el público y OpenAI planea lanzar una versión avanzada, ChatGPT-4, en este 2023.

ChatGPT es diferente de otros modelos de IA en la forma en que puede escribir software en diferentes idiomas, depurar el código, explicar un tema complejo de múltiples maneras, prepararse para una entrevista o redactar un ensayo. Similar a lo que uno puede hacer a través de búsquedas web para aprender estos temas, ChatGPT facilita tales tareas, incluso proporciona el resultado final.

¿Qué es diferente en ChatGPT?

Quizá lo más importante es que ChatGPT es ajustado y entrenado constantemente por los usuarios, que pueden votar a favor o en contra de sus respuestas, lo que lo hace aún más preciso y poderoso, ya que recopila datos por sí mismo. De hecho, tal y como explican los profesionales que lo han probado, a diferencia de otros productos similares, puede participar activamente en una conversación y completar tareas complejas con una precisión asombrosa, al tiempo que ofrece respuestas coherentes y humanas.

Esto no quiere decir que ChatGPT no tenga sus limitaciones, porque todavía puede cometer errores, compartir información falsa y engañosa, malinterpretar las instrucciones de una manera cómica y ser manipulado para sacar la conclusión equivocada.

Pero el poder de ChatGPT no radica en su capacidad para conversar, sino en su capacidad casi ilimitada para completar tareas en masa, de manera más eficiente y mucho más rápida de lo que podría hacerlo un humano. Con las entradas y comandos correctos, y algunas soluciones creativas, ChatGPT se puede convertir en una herramienta de automatización inquietantemente poderosa.

Con eso en mente, no es difícil imaginar cómo un cibercriminal podría convertir ChatGPT en un arma. Se trata de encontrar el método correcto, escalarlo y hacer que la IA complete tantas tareas como sea posible a la vez, con múltiples cuentas y en varios dispositivos si es necesario.

Luces y sombras

Como ocurre con cualquier nueva tecnología, ChatGPT tiene sus propios beneficios y desafíos y tendrá un impacto significativo en el mercado de la ciberseguridad.

Por una parte, la IA es una tecnología prometedora para ayudar a desarrollar productos avanzados de ciberseguridad. Muchos creen que un uso más amplio de la IA y el aprendizaje automático son fundamentales para identificar amenazas potenciales más rápidamente. ChatGPT podría desempeñar un papel crucial en la detección y respuesta a los ataques cibernéticos y la mejora de la comunicación dentro de la organización durante esos momentos.

Por otra, ya hay voces que alertan de otros usos no deseados.

Por el momento, ChatGPT no escribirá un código de malware si se le pide que escriba uno, porque cuenta con protocolos de seguridad para identificar solicitudes inapropiadas. Pero, en las últimas semanas, algunos desarrolladores han intentado saltarse los protocolos y han logrado obtener el resultado deseado. Si un mensaje es lo suficientemente detallado como para explicar al bot los pasos para escribir el malware en lugar de un mensaje directo, responderá al mensaje, construyendo efectivamente malware bajo demanda. Si unimos esto a que existen grupos criminales que ofrecen malware como servicio, con la ayuda de un programa de IA como ChatGPT, pronto puede ser más rápido y fácil para los atacantes lanzar ataques cibernéticos con la ayuda de código generado por IA. Resumiendo, ChatGPT puede dar el poder a atacantes aún menos experimentados para poder escribir un código de malware más preciso, algo que anteriormente solo podían hacer los expertos.

Otro ejemplo de uso inadecuado de ChatGPT, es su capacidad para **responder a cualquier consulta de contenido,** algo que puede aplicarse cuando se combina con un ataque BEC (Business Enterprise Compromise). En este caso, los atacantes utilizan una plantilla para generar un correo electrónico engañoso para que el destinatario le proporcione al atacante la información o el activo que desea. Las herramientas actuales pueden detectar ataques BEC, pero con ChatGPT los atacantes podrían generar un contenido único para cada mensaje, lo que hace que estos ataques sean más difíciles de detectar.

Igualmente, escribir mensajes de phishing puede ser más fácil, sin ninguno de los errores tipográficos o formatos únicos que hoy en día a menudo son críticos para diferenciar estos

ataques de los mensajes legítimos. De hecho, muchos expertos en seguridad creen que la capacidad de ChatGPT para escribir correos electrónicos de phishing que suenan legítimos, el principal vector de ataque para el ransomware, hará que el chatbot sea ampliamente adoptado por los ciberdelincuentes, particularmente aquellos que no son hablantes nativos de inglés.

Más voces de alarma

A la vista de estas opciones, ya son varias las voces que se elevan para advertir de potenciales peligros. Pero, tal y como informa Bloomberg, la última advertencia es particularmente llamativa, porque proviene de la propia OpenAI. Así, dos de sus investigadores se encontraban entre los seis autores de un nuevo informe que investiga la amenaza de las operaciones de influencia habilitadas por la IA.

"Nuestro juicio final es que los modelos de lenguaje serán útiles para los propagandistas y probablemente transformarán las operaciones de influencia en línea", según el comunicado que acompaña al informe.
En el estudio sobre las operaciones de influencia habilitadas por IA, los investigadores destacan que una de sus principales preocupaciones es que las campañas podrían ser más baratas, más fáciles de escalar, instantáneas, más persuasivas y difíciles de identificar utilizando las herramientas de IA.
Las preocupaciones sobre los chatbots avanzados no se detienen en las operaciones de influencia. Los expertos en ciberseguridad advierten que ChatGPT y modelos similares de IA podrían bajar el listón para que los piratas informáticos escriban código malicioso para apuntar a vulnerabilidades existentes o recién descubiertas. De hecho, recientemente desde Check Point Software señalaron que los atacantes ya estaban reflexionando en foros de piratería sobre cómo recrear cepas de malware o mercados web oscuros utilizando el chatbot.

Otras amenazas

Además de los peligros que hemos ido detallando, los expertos han señalado otras posibles amenazas generadas con el uso de ChatGPT.
Uno de estos ejemplos pasa por algo, a priori, totalmente inocuo: construir un sitio web con ChatGPT. En internet es posible encontrar un montón de tutoriales que explican con gran detalle cómo hacer precisamente eso, pero el problema puede venir con las intenciones de la persona que quiere montar una página web. Ahí las posibilidades son casi infinitas, porque se podría clonar un sitio web existente con ChatGPT y luego modificarlo, crear sitios web de comercio electrónico falsos, ejecutar un sitio con estafas de scareware...
En esta misma línea, configurar un sitio web falso, ejecutar una página de redes sociales fraudulenta o crear un sitio de imitación, necesita mucho contenido, y debe parecer lo más legítimo posible para que la estafa funcione, y esto es lo que se facilita (tanto en la realización de la tarea como en el coste) con ChatGPT, porque ya no hacen falta desarrolladores y generadores de contenido.
Otro ejemplo es la desinformación, que se ha convertido en un problema importante en los últimos años. Las noticias falsas se propagan como un reguero de pólvora en las redes sociales, y herramientas como ChatGPT podrían incrementar este problema.

Fuente de información: itdigitalsecurity