Skynet Ahoy? Qué esperar de los riesgos de seguridad de la IA de próxima generación

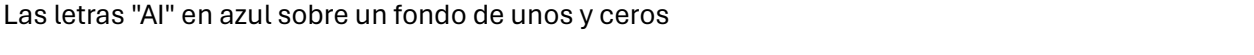
La innovación que demuestran ChatGPT y otros LLM es algo bueno, pero las salvaguardas y otros marcos de seguridad deben seguir el ritmo.

Imagen de Elizabeth Montalbano, escritora colaboradora

Elizabeth Montalbano, escritora colaboradora

28 de diciembre de 2023

Lectura de 6 minutos



Las letras "AI" en azul sobre un fondo de unos y ceros

FUENTE: MARCOS ALVARADO VÍA FOTO DE STOCK DE ALAMY

A medida que la innovación en inteligencia artificial (IA) continúa a buen ritmo, 2024 será un momento crucial para que las organizaciones y los órganos rectores establezcan estándares de seguridad, protocolos y otras barreras de seguridad para evitar que la IA se les adelante, advierten los expertos en seguridad.

Los modelos de lenguajes grandes (LLM), impulsados por algoritmos sofisticados y conjuntos de datos masivos, demuestran una comprensión del lenguaje notable y capacidades conversacionales similares a las humanas. Una de las plataformas más sofisticadas hasta la fecha es GPT-4 de OpenAI, que cuenta con capacidades avanzadas de razonamiento y resolución de problemas y potencia el bot ChatGPT de la compañía. Y la compañía, en asociación con Microsoft, comenzó a trabajar en GPT-5, que según el director ejecutivo Sam Altman irá mucho más allá, hasta el punto de poseer "superinteligencia".

Estos modelos representan un enorme potencial para ganancias significativas de productividad y eficiencia para las organizaciones, pero los expertos coinciden en que ha llegado el momento de que la industria en su conjunto aborde los riesgos de seguridad inherentes que plantean su desarrollo e implementación. De hecho, una investigación reciente realizada por Writerbuddy AI, que ofrece una herramienta de redacción de contenidos basada en IA, encontró que ChatGPT ya ha tenido 14 mil millones de visitas y sigue contando.

A medida que las organizaciones avanzan hacia el progreso en IA, "debería ir acompañado de consideraciones éticas y evaluaciones de riesgos rigurosas", dice Gal Ringel, director ejecutivo de la firma de seguridad y privacidad basada en IA MineOS.

¿Es la IA una amenaza existencial?

Las preocupaciones sobre la seguridad de la próxima generación de IA comenzaron a filtrarse en marzo, con una carta abierta firmada por casi 34.000 tecnólogos de alto nivel que pedían detener el desarrollo de sistemas de IA generativa más potentes que el GPT-4 de OpenAI. La carta citaba los "profundos riesgos" para la sociedad que representa la tecnología y la "carrera fuera de control de los laboratorios de inteligencia artificial para desarrollar y desplegar mentes digitales cada vez más poderosas que nadie, ni siquiera sus creadores, puede comprender, predecir o control confiable."

A pesar de esos temores distópicos, la mayoría de los expertos en seguridad no están tan preocupados por un escenario apocalíptico en el que las máquinas se vuelvan más inteligentes que los humanos y se apoderen del mundo.

"La carta abierta señaló preocupaciones válidas sobre el rápido avance y las posibles aplicaciones de la IA en un sentido amplio de '¿es esto bueno para la humanidad'?", dice Matt Wilson, director de ingeniería de ventas de la firma de ciberseguridad Netrix. "Aunque son impresionantes en ciertos escenarios, las versiones públicas de las herramientas de IA no parecen tan amenazantes".

Lo preocupante es el hecho de que los avances y la adopción de la IA avanzan demasiado rápido como para que los riesgos se gestionen adecuadamente, señalan los investigadores. "No podemos volver a tapar la caja de Pandora", observa Patrick Harr, director ejecutivo del proveedor de seguridad de inteligencia artificial SlashNext.

Además, simplemente "intentar detener el ritmo de innovación en el espacio no ayudará a mitigar" los riesgos que presenta, que deben abordarse por separado, observa Marcus Fowler, director ejecutivo de la firma de seguridad de inteligencia artificial DarkTrace Federal. Eso no significa que el desarrollo de la IA deba continuar sin control, afirma. Por el contrario, el ritmo de evaluación de riesgos y la implementación de salvaguardas apropiadas deben coincidir con el ritmo al que se capacita y desarrolla a los LLM.

"La tecnología de la IA está evolucionando rápidamente, por lo que los gobiernos y las organizaciones que la utilizan también deben acelerar los debates sobre la seguridad de la IA", explica Fowler.

Riesgos generativos de la IA

Existen varios riesgos ampliamente reconocidos para la IA generativa que exigen consideración y solo empeorarán a medida que las generaciones futuras de esta tecnología se vuelvan más

inteligentes. Afortunadamente para los humanos, ninguno de ellos plantea hasta ahora un escenario apocalíptico de ciencia ficción en el que la IA conspire para destruir a sus creadores.

En cambio, incluyen amenazas mucho más familiares, como fugas de datos, potencialmente de información empresarial sensible; uso indebido para actividades maliciosas; y resultados inexactos que pueden inducir a error o confundir a los usuarios y, en última instancia, tener consecuencias comerciales negativas.

Debido a que los LLM requieren acceso a grandes cantidades de datos para proporcionar resultados precisos y contextualmente relevantes, la información confidencial puede revelarse o usarse indebidamente sin darse cuenta.

"El principal riesgo es que los empleados le proporcionen información sensible para el negocio cuando le piden que escriba un plan o reformule correos electrónicos o presentaciones comerciales que contienen información patentada de la empresa", señala Ringel.

Desde una perspectiva de ciberataque, los actores de amenazas ya han encontrado innumerables formas de convertir ChatGPT y otros sistemas de inteligencia artificial en armas. Una forma ha sido utilizar los modelos para crear sofisticados ataques de correo electrónico empresarial (BEC) y otros ataques de phishing, que requieren la creación de mensajes personalizados y de ingeniería social.

Está diseñado para el éxito.

"Con el malware, ChatGPT permite a los ciberdelincuentes realizar infinitas variaciones de código para estar un paso por delante de los motores de detección de malware", afirma Harr.

Las alucinaciones de IA también representan una importante amenaza para la seguridad y permiten a actores maliciosos armar tecnología basada en LLM como ChatGPT de una manera única. Una alucinación de IA es una respuesta plausible de la IA que es insuficiente, sesgada o completamente falsa. "Las respuestas ficticias u otras respuestas no deseadas pueden llevar a las organizaciones a tomar decisiones, procesos y comunicaciones engañosas", advierte Avivah Litan, vicepresidente de Gartner.

Los actores de amenazas también pueden utilizar estas alucinaciones para envenenar a los LLM y "generar información errónea específica en respuesta a una pregunta", observa Michael Rinehart, vicepresidente de IA del proveedor de seguridad de datos Securiti. "Esto es extensible a la

generación de código fuente vulnerable y, posiblemente, a modelos de chat capaces de dirigir a los usuarios de un sitio a acciones inseguras".

Los atacantes pueden incluso llegar a publicar versiones maliciosas de paquetes de software que un LLM podría recomendar a un desarrollador de software, creyendo que es una solución legítima a un problema. De esta manera, los atacantes pueden utilizar aún más la IA como arma para montar ataques a la cadena de suministro.

El camino a seguir

La gestión de estos riesgos requerirá una acción mesurada y colectiva antes de que la innovación en IA supere la capacidad de la industria para controlarla, señalan los expertos. Pero también tienen ideas sobre cómo abordar el problema de la IA.

Harr cree en una estrategia de "luchar contra la IA con A", en la que "los avances en las soluciones de seguridad y las estrategias para frustrar los riesgos impulsados por la IA deben desarrollarse a un ritmo igual o mayor".

"La protección de la ciberseguridad necesita aprovechar la IA para combatir con éxito las ciberamenazas utilizando tecnología de IA", añade. "En comparación, la tecnología de seguridad heredada no tiene ninguna posibilidad contra estos ataques".

Sin embargo, las organizaciones también deberían adoptar un enfoque mesurado a la hora de adoptar la IA (incluidas las soluciones de seguridad basadas en IA) para no introducir más riesgos en su entorno, advierte Wilson de Netrix.

"Comprenda qué es y qué no es la IA", aconseja. "Desafíe a los proveedores que afirman emplear IA para que describan qué hace, cómo mejora su solución y por qué es importante para su organización".

Rinehart de Securiti ofrece un enfoque de dos niveles para introducir gradualmente la IA en un entorno mediante la implementación de soluciones enfocadas y luego colocando barreras de seguridad inmediatamente antes de exponer a la organización a riesgos innecesarios.

"Primero adopte modelos específicos de aplicaciones, potencialmente aumentados con bases de conocimiento, que estén diseñadas para proporcionar valor en casos de uso específicos", afirma.

"Entonces... implemente un sistema de monitoreo para salvaguardar estos modelos examinando los mensajes que reciben y reciben en busca de cuestiones de privacidad y seguridad".

Los expertos también recomiendan establecer políticas y procedimientos de seguridad en torno a la IA antes de su implementación, en lugar de hacerlo como una ocurrencia tardía para mitigar el riesgo. Incluso pueden establecer un oficial de riesgos de IA o un grupo de trabajo dedicado para supervisar el cumplimiento.

Fuera de la empresa, la industria en su conjunto también debe tomar medidas para establecer estándares y prácticas de seguridad en torno a la IA que todos los que desarrollan y utilizan la tecnología puedan adoptar, algo que requerirá una acción colectiva tanto del sector público como del privado a escala global. , afirma Fowler de DarkTrace Federal.

Cita directrices para construir sistemas seguros de IA publicadas en colaboración por la Agencia de Seguridad de Infraestructura y Ciberseguridad de EE. UU. (CISA) y el Centro Nacional de Seguridad Cibernética del Reino Unido (NCSC) como ejemplo del tipo de esfuerzos que deberían acompañar la evolución continua de la IA.

"En esencia", dice Rinehart de Securiti, "el año 2024 será testigo de una rápida adaptación tanto de la seguridad tradicional como de las técnicas de IA de vanguardia para proteger a los usuarios y los datos en esta era emergente de IA generativa".