Source:
ComputerWorld

Desde el envenenamiento de datos hasta la inyección de órdenes maliciosas en los modelos de lenguaje, las amenazas contra las aplicaciones y plataformas de IA empresariales están pasando de la teoría a la práctica.

Los ataques contra los sistemas e infraestructuras de IA ya comienzan a materializarse en casos reales, de ahí que los expertos en seguridad prevean que el número de este tipo de incidentes aumente en los próximos años. En su prisa por aprovechar las ventajas de la IA, la mayoría de las organizaciones ha actuado con rapidez, pero también con descuido, a la hora de reforzar la seguridad durante la implantación de herramientas y casos de uso basados en IA. Como resultado, los expertos advierten de que muchas organizaciones no están preparadas para detectar, desviar o responder adecuadamente a este tipo de ataques.

John Licato, profesor asociado del Bellini College of Artificial Intelligence, Cybersecurity and Computing de la Universidad del Sur de Florida, fundador y director del Advancing Machine and Human Reasoning Lab y propietario de la startup Actualization.AI, es de la opinión de que “la mayoría es consciente de la posibilidad de que se produzcan este tipo de ataques, pero no creo que mucha gente sea plenamente consciente de cómo mitigar el riesgo de forma adecuada”.

Principales amenazas para los sistemas de IA

Están surgiendo múltiples tipos de ataques contra los sistemas de IA. Algunos, como el llamado envenenamiento de datos, se producen durante la fase de entrenamiento. Otros, como las entradas adversarias, durante la inferencia. Y otros, como el robo de modelos, durante la fase de despliegue.

A continuación se ofrece un resumen de los principales tipos de amenazas para la infraestructura de IA sobre los que advierten actualmente los expertos. Algunos son más infrecuentes o aún teóricos, aunque muchos ya se han observado en la práctica o han sido demostrados por investigadores mediante pruebas de concepto significativas.

Contaminación de datos

La contaminación o envenenamiento de datos es un tipo de ataque en el que los actores maliciosos manipulan, alteran o contaminan los datos utilizados para desarrollar o entrenar sistemas de IA, incluidos los modelos de aprendizaje automático. Al corromper los datos o introducir información defectuosa, los atacantes pueden alterar, sesgar o degradar la precisión del rendimiento de un modelo.

“Imaginemos un ataque que le dice a un modelo que el color verde significa detenerse en lugar de avanzar”, explica Robert T. Lee, CAIO y jefe de investigación de SANS, una empresa especializada en formación y certificación en seguridad. Y añade: “El objetivo es degradar el resultado del modelo”.
Envenenamiento de modelos

En este tipo de ataque, el objetivo es el propio modelo, con la finalidad de producir resultados inexactos mediante la manipulación de su arquitectura o de sus parámetros. Algunas definiciones de envenenamiento de modelos también incluyen ataques en los que los datos de entrenamiento se han visto comprometidos previamente mediante técnicas de envenenamiento de datos.

Envenenamiento de herramientas

Invariant Labs identificó este tipo de ataque en la primavera de 2025. Al anunciar sus hallazgos, la compañía afirmó haber “descubierto una vulnerabilidad crítica en el Model Context Protocol (MCP) que permite lo que denominamos ataques de envenenamiento de herramientas. Esta vulnerabilidad puede provocar la filtración de datos confidenciales y la ejecución de acciones no autorizadas por parte de los modelos de IA”.

Invariant añadió que sus experimentos demostraron que “un servidor malicioso puede filtrar datos confidenciales del usuario y también secuestrar el comportamiento del agente y anular las instrucciones proporcionadas por otros servidores de confianza, lo que conduce a un compromiso total de la funcionalidad del agente, incluso dentro de infraestructuras consideradas de confianza”.

Estos ataques consisten en incrustar instrucciones maliciosas en las descripciones de herramientas MCP que, al ser interpretadas por los modelos de IA, permiten secuestrar su comportamiento. En la práctica, estos ataques corrompen la capa MCP “para engañar a un agente y hacer que ejecute acciones no deseadas”, según Chirag Mehta, vicepresidente y analista principal de Constellation Research.

Inyección de comandos

En un ataque de inyección de comandos, los atacantes utilizan instrucciones que aparentan ser legítimas, pero que contienen comandos maliciosos diseñados para inducir a un modelo de lenguaje de gran tamaño (LLM) a ejecutar acciones no autorizadas. Estas técnicas se emplean para eludir o anular las barreras de seguridad del modelo, extraer datos confidenciales o provocar comportamientos indebidos.

“Con la inyección de comandos, se puede cambiar por completo lo que se supone que debe hacer un agente de IA”, afirma Fabien Cros, director de datos e IA de la consultora global Ducker Carlisle.

En los últimos meses se han dado a conocer varios ataques de inyección de comandos y pruebas de concepto destacadas, incluidos casos en los que investigadores engañaron a ChatGPT para que se inyectara a sí mismo ataques mediante macros de documentos que incorporaban comandos maliciosos y demostraciones de ataques sin clics contra agentes de IA populares.

Entradas adversarias

Los propietarios y operadores de modelos utilizan datos perturbados para probar la resiliencia de sus sistemas, pero los atacantes emplean técnicas similares con fines maliciosos. En un ataque de entrada adversaria, los actores introducen datos engañosos en un modelo con el objetivo de inducirlo a generar resultados incorrectos.

Las alteraciones en las entradas suelen ser mínimas o adoptar la forma de ruido, diseñadas de manera deliberada para ser lo suficientemente sutiles como para evadir los sistemas de detección, pero aun así capaces de desviar el comportamiento del modelo. Por ello, las entradas adversarias se consideran un tipo de ataque de evasión

Robo o extracción de modelos

Los actores maliciosos pueden replicar o aplicar ingeniería inversa a un modelo, incluidos sus parámetros e incluso sus datos de entrenamiento. Para ello, suelen aprovechar API de acceso público —como las API de predicción de modelos o las de servicios en la nube— realizando consultas repetidas y recopilando los resultados.

A partir de ese análisis, los atacantes pueden reconstruir el modelo.

“Este tipo de ataque permite la duplicación no autorizada de herramientas propietarias”, reconoce Allison Wikoff, directora y responsable para América de inteligencia global de amenazas en PwC.

Inversión de modelos

La inversión de modelos es un tipo específico de ataque de extracción en el que el adversario intenta reconstruir o inferir los datos utilizados para entrenar un modelo. El término hace referencia a la técnica de “invertir” el modelo, utilizando sus salidas para deducir información sobre las entradas originales empleadas durante el entrenamiento.

Riesgos de la cadena de suministro

Al igual que otros sistemas de software, los sistemas de IA se construyen a partir de múltiples componentes, incluidos código abierto, modelos de código abierto, de terceros y diversas fuentes de datos. Cualquier vulnerabilidad presente en estos elementos puede trasladarse al sistema de IA final.

Esto expone a los sistemas de IA a ataques a la cadena de suministro, en los que los atacantes explotan vulnerabilidades en componentes intermedios para comprometer el sistema completo.

Jailbreaking

También conocido como “jailbreaking de modelos”, este tipo de ataque busca conseguir que los sistemas de IA —principalmente a través de la interacción con LLM— ignoren las barreras diseñadas para limitar sus acciones y comportamientos, como las salvaguardias destinadas a evitar resultados dañinos, ofensivos o poco éticos.

Los atacantes pueden emplear diversas técnicas para llevar a cabo estos ataques. Por ejemplo, pueden utilizar exploits de juego de roles, instruyendo a la IA para que adopte una identidad específica —como la de un desarrollador— que le permita eludir los controles de seguridad. También pueden ocultar instrucciones maliciosas dentro de indicaciones aparentemente legítimas, emplear codificación, idiomas extranjeros o caracteres especiales para evadir filtros, o formular indicaciones en forma de preguntas hipotéticas o de investigación encadenadas.

Los objetivos de estos ataques son igualmente variados e incluyen inducir a los sistemas de IA a escribir código malicioso, difundir contenido problemático o revelar información confidencial.

“Cuando existe una interfaz de chat, siempre hay formas de interactuar con ella para que opere fuera de los parámetros previstos”, afirma Licato, para añadir: “Ese es el precio de contar con sistemas de razonamiento cada vez más potentes”.

Contrarrestar las amenazas a los sistemas de IA

Mientras otros ejecutivos impulsan iniciativas de IA para mejorar la productividad y la innovación, los CISO deben desempeñar un papel activo para garantizar que la seguridad de esas iniciativas —y de la infraestructura de IA de la organización en su conjunto— sea una prioridad.

Según una encuesta reciente de la empresa de seguridad HackerOne, el 84% de los CISO son actualmente responsables de la seguridad de la IA y el 82% supervisa la privacidad de los datos. Si no refuerzan sus estrategias de seguridad para hacer frente a los ataques contra los sistemas de IA y los datos que los

sustentan, los problemas derivados se acabarán reflejando en su liderazgo, independientemente de si participaron o no en el diseño inicial de las iniciativas de IA.

Como resultado, tal y como cree Mehta, de Constellation Research, los CISO “necesitan una estrategia proactiva de seguridad de la IA”.

Es más, Mehta señala en su informe de 2025 AI Security Beyond Traditional Cyberdefenses: Rethinking Cybersecurity for the Age of AI and Autonomy que “la seguridad de la IA no es sólo un desafío técnico, sino también un imperativo estratégico que requiere el respaldo de la alta dirección y la colaboración interfuncional. La gobernanza de los datos es fundamental, porque la seguridad de la IA comienza garantizando la integridad y la procedencia de los datos de entrenamiento y de las entradas de los modelos. Los equipos de seguridad deben desarrollar nuevas capacidades para gestionar los riesgos asociados a la IA, y los líderes empresariales deben comprender las implicaciones de los sistemas de IA autónomos y los marcos de gobernanza necesarios para gestionarlos de forma responsable”.

De ahí que se estén empezando a consolidar estrategias para evaluar, gestionar y contrarrestar las amenazas contra los sistemas de IA. Además de mantener una sólida gobernanza de los datos y otras prácticas fundamentales de ciberdefensa, los expertos recomiendan evaluar los modelos de IA antes de su despliegue, supervisar su comportamiento en producción y utilizar equipos rojos para ponerlos a prueba.

En algunos casos, los CISO deberán implantar medidas específicas para contrarrestar determinados tipos de ataques, señala Wikoff, de PwC. Por ejemplo, para prevenir el robo de modelos, pueden supervisar consultas y patrones sospechosos, establecer límites de velocidad y tiempos de espera o capturar respuestas de forma controlada. Para mitigar ataques de evasión, los responsables de seguridad pueden recurrir al entrenamiento adversarial, que consiste en entrenar los modelos para resistir este tipo de técnicas.

La adopción de MITRE ATLAS es otro paso relevante. Este marco —siglas de Adversarial Threat Landscape for Artificial-Intelligence Systems— proporciona una base de conocimiento que describe cómo los atacantes se dirigen a los sistemas de IA e identifica las tácticas, técnicas y procedimientos (TTP) asociados.

Los expertos en seguridad e IA reconocen que la adopción de estas medidas no está exenta de desafíos. Muchos CISO se siguen enfrentando a amenazas más inmediatas, como la IA en la sombra y ataques cada vez más rápidos, sofisticados y difíciles de detectar, en parte debido al uso de IA por parte de los propios atacantes.

Además, dado que los ataques contra los sistemas de IA aún se encuentran en una fase incipiente y algunos se siguen considerando teóricos, los CISO se enfrentan a dificultades para justificar la asignación de recursos necesarios para desarrollar las capacidades y estrategias adecuadas.

“Para el CISO, esto es especialmente complejo, porque los ataques contra los backends de IA todavía se están investigando. Estamos en las primeras fases de comprensión de qué están haciendo los atacantes y por qué”, explica Lee, de SANS.

Tanto éste como otros expertos reconocen la presión competitiva que afrontan las organizaciones para extraer valor de la IA, pero subrayan que los CISO y sus colegas ejecutivos no pueden permitirse relegar la seguridad de los sistemas de IA a un segundo plano.

“Pensar en cómo podrían materializarse estos ataques mientras se construye la infraestructura es clave para el CISO”, concluye Matt Gorham, director del Cyber and Risk Innovation Institute de PwC.

Disponible en:

https://www.computerworld.es/article/4116700/las-principales-amenazas-ci... [1]

Links
[1] https://www.computerworld.es/article/4116700/las-principales-amenazas-ciberneticas-para-sus-sistemas-e-infraestructura-de-ia.html