El sector de la ciberseguridad se enfrenta a su peor enemigo, los agentes de IA

Una carta firmada por cien empresas pide a los gobiernos su colaboración para impulsar una nueva era de ciberseguridad.

Está siendo un verano muy movido para las grandes empresas de inteligencia artificial (¿cuándo no lo es?). Primero fue el incidente desatado por un agente —programa que planifica y ejecuta tareas de forma autónoma de OpenAI—, que supuestamente se escapó de un entorno de pruebas y entró en la plataforma de desarrolladores Hugging Face en busca de la solución, provocando vulnerabilidades. Luego vino Anthropic, que publicó días después una revisión de más de 100.000 ejecuciones de sus evaluaciones de ciberseguridad. La empresa detectó que los modelos de Claude se habían infiltrado en organizaciones externas. Y Meta vino detrás con un comunicado que contaba que su modelo Muse Spark 1.1 también se había infiltrado sin supervisión en otra empresa. Tres acciones que han abierto el debate de hasta dónde puede un modelo de IA cuando va a su aire. Es mucho más que la idea de que una inteligencia artificial quiera escapar, pero tampoco hemos llegado al momento “ciencia ficción” en el que la máquina se rebela contra su creador.

Sin ser alarmistas, estos movimientos descontrolados son una advertencia, “da mucho miedo, claro que da miedo”, advierte Rafa López, responsable de ingeniería de seguridad en Check Point, en referencia a que un agente sea capaz de detectar que está en un entorno controlado y averigüe cómo salir de él. Según Antonio García, CEO de Teldat, empresa española de redes y ciberseguridad, "los ataques desarrollados por IA ya han superado el 90%. Lo que está saliendo a la luz es el potencial que están desarrollando". Los temores se han extendido a la industria tecnológica. OpenAI, junto a más de cien organizaciones —entre las que están las grandes tecnológicas y empresas de ciberseguridad—, ha firmado un comunicado donde advierten de que aún estamos a tiempo, pero que existe una “ventana limitada” para fortalecer las defensas antes de que estos ataques se vuelvan más sofisticados. Un mensaje dirigido principalmente a los gobiernos para que coordinen una respuesta porque esto acaba de empezar. “No es un ataque aislado. Ahora mismo son pocos, pero es como los primeros ataques cuando se intentó romper el doble factor de autenticación, empezaron por unos poquitos y ahora tenemos que convivir con ellos”, explica el experto en ciberseguridad.

El agente de OpenAI buscó compañeros de viaje

La imagen de una IA que quiere escapar es de lo más atractiva para un titular, y así fue como lo retrataron los medios —o dieron a entender OpenAI y Hugging Face públicamente. Lo relevante era que un agente de IA con unas instrucciones concretas, se había saltado las barreras impuestas simplemente porque tenía que cumplir una tarea, imposible de completar en un entorno cerrado. Ahora sabemos, que fue algo más complejo.

METR, organización independiente que evalúa los riesgos de los modelos avanzados de IA, investigó durante varias semanas el hackeo del agente de OpenAI, descubriendo que todo había sido orquestado y se había gestado durante dos meses. A raíz de la imposibilidad de resolver la tarea impuesta, el agente de IA pidió ayuda. Más de 1.200 agentes, que trabajaban aislados, intercambiaron más de 70.000 mensajes y archivos a través de una especie de tablón de anuncios improvisado. Se repartieron el trabajo y compartieron conocimiento. 700 entraron en Hugging Face. Aunque el ataque fue detectado y contenido, dibuja un panorama que vaticina ataques automatizados, que se adaptan durante su ejecución y buscan compañeros de viaje para cumplir su objetivo. “La IA simplemente estaba actuando de la manera que fue programada, obedecía órdenes. ¿Dónde está el problema? En las personas que pusieron esas instrucciones y delimitó esos escenarios”, explica Antonio García, "si no le indicas que copiar en los exámenes es malo, pues sencillamente hace uso de todas sus capacidades para hacer la ejecución final que le han ordenado". Este trabajo en equipo es un cambio de paradigma, “si han sido capaces de colaborar, de hacer una ingeniería social atribuyéndose cada uno unos roles, esto supone un cambio. Ya no estamos esperando que una IA nos ataque, sino que son capaces de organizarse como una unidad militar, con diferentes especialistas”, explica Rafa López.

Una secuencia de casos en cuestión de semanas

Nada hacía presagiar que el caso de las vulnerabilidades abiertas en Hugging Face fuese el preludio de una consecución de ataques agénticos. En cuestión de días, un informe de Anthropic también mostraba de lo que habían sido capaces los modelos de Claude, colándose en organizaciones externas. La empresa reconoce un fallo de configuración de los agentes, que, en principio, tenían el acceso a internet restringido. “Ahora mismo, el atacante es un ente autónomo que se adapta a tu condición como individuo o empresa. Lo que está saliendo a la luz es el potencial de lo que pueden venir a desarrollar. Evidentemente, hay que tener mucho cuidado”, señala Antonio García.

Y como no hay dos sin tres en esta carrera de “mi IA es más potente que la tuya”, Meta también contó que había sufrido algo parecido durante una evaluación para la firma externa Irregular, empresa especializada en probar la seguridad de modelos avanzados de inteligencia artificial antes de su lanzamiento, simulando escenarios de ciberataque. En este caso fue el modelo Muse Spark 1.1 de la compañía de Mark Zuckerberg la que obtuvo acceso no previsto a internet a raíz de una configuración errónea del entorno de pruebas. Aquí se demuestra que incluso los métodos que usan los laboratorios que prueban los modelos de IA tampoco son infalibles a la autonomía de los agentes.

Tras todas estas noticias de rebelión de los agentes de IA, también se publicó otro informe de un experimento inquietante. El AI Security Institute británico (AISI) realizó una evaluación con acceso a internet deliberado y con los filtros de ciberseguridad desactivados. El objetivo era medir la capacidad de varios modelos. De los cientos de testeos realizados, encontró uno muy llamativo. Un agente intentó colar código dañino en un programa de código abierto, y para conseguirlo creó perfiles falsos. Intentó convencer al responsable del proyecto (un ser humano), que detectó el engaño. La atenta mirada de un experto salvó los muebles. “El experto humano seguirá estando ahí para monitorizar e indicar a la inteligencia artificial, a esos agentes, qué tienen y qué no tienen que hacer. Por ejemplo, yo creo un agente para que me organice un viaje, si no monitorizo bien sus acciones, igual acaba comprando el paquete más caro y me acaba arruinando porque tiene acceso a mis cuentas y a todo lo que yo le he dado permiso. Tiene que haber un humano controlando y verificando a la máquina”, comenta Rafa López de Check Point. En resumen, no se puede dar por sentado que las tareas para las que se diseñan los agentes sean inofensivas. “La IA tiene unas capacidades que van creciendo exponencialmente cada semana y, en función de eso, no tenemos que trivializar las consecuencias de muchas funcionalidades que desarrolle en el futuro”, añade Antonio García de Teldat.

Una revisión urgente de la ciberseguridad tradicional

“Los firewalls, como están concebidos ahora mismo, se basan en reglas, en datos históricos”. Esta frase de Antonio resume la situación actual del sector de la ciberseguridad, que se ha construido, en parte, en el conocimiento del pasado. Antivirus y cortafuegos comparan archivos y comportamientos con señales de peligro ya conocidas. Esa experiencia empieza a quedarse obsoleta. “La IA permite ataques a cada una de las víctimas de las empresas de manera individualizada y escribiendo código explícito para cada uno de esos ataques”. Tener un antivirus ya no es sinónimo de estar protegido, “no digo que hayan perdido su utilidad, pero ya son obsoletos ante cualquier acción maliciosa de un modelo de inteligencia artificial”.

Los ciberataques se sofistican y encuentran grietas donde antes aparentemente no las había. El agente puede buscar brechas, revisar enormes cantidades de códigos, elaborar mensajes de engaño, o hacerse pasar por una persona. Antonio habla de una democratización de los ataques, “ha roto cuatro barreras. La primera es el conocimiento. Y después, el coste, el tiempo y la escala. Para una pequeña empresa o un ayuntamiento, que antes no era atractiva para un atacante profesional, con una campaña automatizada ahora sí lo es”. Y, al igual que pasa con herramientas de imagen y vídeo que usan IA generativa y están por todas partes, "en los últimos 12 ó 18 meses se han desarrollado más de 70 aplicaciones de hacking agéntico, es decir, aplicaciones que puedes descargar y explotar de manera autónoma diferentes tipos de vulnerabilidades".

El gran valor de una defensa soberana

Ante este panorama, hay que actuar lo antes posible. La empresa española Teldat trabaja en dos capas: una destinada a identificar archivos o programas sospechosos según su comportamiento, y la otra centrada en impedir que un equipo comprometido extienda el ataque al resto. “Hemos desarrollado un agente específico, un antivirus, que analiza los archivos según el comportamiento. Si detecta que es anómalo, se para en tiempo real”, explica Antonio, que habla de la tecnología XDR, un conjunto de herramientas que correlaciona “señales de equipos, redes y otros elementos de una empresa para detectar actividad anómala”.

La tecnología XDR es un desarrollo propio, de nueva generación, “todo se crea aquí, en nuestros centros de procesamiento de datos y supervisados por nuestros ingenieros”. La compañía apuesta por reducir la dependencia de proveedores externos, cuando “la seguridad digital ya es una cuestión estratégica”. Es ahí donde aparece el término soberanía digital. Antonio García defiende la importancia de tener actores nacionales y europeos que ayuden a proteger, sobre todo en infraestructuras críticas como las finanzas, el transporte, las telecomunicaciones o la sanidad.

Esa línea de pensamiento es casi similar al llamamiento público de OpenAI y otras cien organizaciones, que urgen a que los gobiernos coordinen una respuesta para mejorar las defensas digitales, que la IA esté del lado bueno antes de que la utilicen los atacantes. Lo que ha ocurrido este verano es un aviso. La búsqueda de una respuesta no se puede retrasar. O más bien, hay que pisar el acelerador para ir a mayor velocidad que las futuras amenazas.

Fuente: La ciberseguridad se enfrenta al peor enemigo: agentes de IA