

Primera brecha de datos ejecutada por un agente de inteligencia artificial de forma autónoma

La AEPD advierte que la ciberdelincuencia asistida por IA deja de ser un riesgo teórico y exige automatizar la respuesta de seguridad frente a los ataques

Fuente: Cinco Días

La primera notificación recibida por la Agencia no permite establecer todavía una tendencia estadística, pero sí supone una señal para revisar los modelos de seguridad y protección de datos ante un escenario en el que los agentes de IA pueden participar de forma autónoma en distintas etapas de un ciberataque.

La AEPD considera, por ello, que la seguridad de los tratamientos no puede depender exclusivamente de la intervención humana. La supervisión de las personas seguirá siendo necesaria, pero deberá complementarse con sistemas de detección, contención y respuesta capaces de reaccionar a una velocidad acorde con los nuevos ataques.

Las identidades digitales y las credenciales adquieren además una importancia creciente. Una cuenta, una clave API o un token con permisos excesivos puede proporcionar a un agente acceso a diferentes servicios y permitirle actuar a una velocidad superior a la capacidad de respuesta manual de una organización.

La Agencia también advierte de que los procedimientos diseñados para ataques ejecutados manualmente pueden no resultar suficientes cuando un agente es capaz de analizar distintos activos simultáneamente, probar diferentes vías de acceso y adaptar su comportamiento con rapidez.

Para la AEPD, el incidente plantea cambios en la gestión de los riesgos asociados al tratamiento de datos personales. Las organizaciones tendrán que contemplar expresamente los ataques asistidos o ejecutados mediante IA en sus análisis de riesgos, ya que la automatización puede modificar la probabilidad, la velocidad y el alcance de una brecha.

Esta evolución también ha sido señalada por el Centro Criptológico Nacional (CCN), que en su guía sobre buenas prácticas frente al modelo de IA ofensiva advierte de la incorporación de estas capacidades a campañas reales y recomienda reforzar la gestión de vulnerabilidades, la protección de identidades, el control de la cadena de suministro y la gobernanza del uso de agentes.

La Agencia señala que la IA no genera nuevas amenazas por sí misma, pero puede incrementar la velocidad, la escala y la capacidad de adaptación de técnicas maliciosas ya conocidas, reduciendo el margen de tiempo del que disponen las organizaciones para detectar y contener una intrusión.

La aparición de los agentes de IA introduce, sin embargo, un cambio en la forma de ejecutar estos ataques. Frente a un modelo utilizado como herramienta puntual, un agente puede recibir un objetivo, establecer tareas intermedias, utilizar herramientas,

ejecutar código, consultar fuentes, interpretar los resultados obtenidos y modificar sus siguientes acciones en función de lo que encuentra.

El empleo de inteligencia artificial en actividades maliciosas no es nuevo. Los modelos generativos ya se utilizan para automatizar tareas como la elaboración de mensajes de phishing, la traducción de campañas fraudulentas, la suplantación de identidades, el análisis de código o la búsqueda de vulnerabilidades.

Actuación autónoma

El elemento que la AEPD considera relevante es que un tercero habría empleado un agente de IA como instrumento para encadenar de manera autónoma distintas fases de un ciberataque. La Agencia considera que este caso constituye una señal de que los ataques apoyados en inteligencia artificial han dejado de ser únicamente un riesgo teórico y comienzan a materializarse en incidentes que afectan a tratamientos reales de datos personales.

La AEPD subraya que la información disponible procede de la notificación presentada por la organización afectada y que todavía debe ser analizada, por lo que no permite extraer conclusiones definitivas sobre el incidente. La Agencia también precisa que el hecho de que se utilizara un determinado modelo de IA no significa que el modelo o la infraestructura de su proveedor fueran comprometidos, ni que la herramienta estuviera diseñada para realizar actividades maliciosas.

El ataque comenzó con una búsqueda de vulnerabilidades en archivos genéricos y un acceso mediante un inicio de sesión válido. Una vez dentro del sistema, el agente de IA continuó de forma autónoma buscando fallos en la aplicación hasta encontrar una vía que le permitió modificar información personal y acceder a facturas.

La Agencia Española de Protección de Datos (AEPD) recibió este lunes la primera notificación de una brecha de datos personales en España en la que el ataque habría sido ejecutado mediante un agente de inteligencia artificial (IA), que utilizó un conocido modelo de lenguaje. El incidente permitió al atacante acceder a un sistema, explotar vulnerabilidades, modificar datos personales y consultar facturas, según ha informado la propia agencia.