PANORAMA | Tecnología

La AEPD recibe la primera notificación en España de una brecha de datos ejecutada mediante un agente de inteligencia artificial

El sistema habría buscado vulnerabilidades de forma autónoma, accedido a una aplicación, modificado datos personales y consultado facturas; la Agencia advierte de que el caso aún debe ser analizado antes de extraer conclusiones definitivas

Redacción/Priego Digital
Lunes 21 de septiembre de 2026 - 22:00

La Agencia Española de Protección de Datos (AEPD) ha recibido 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, capaz de encadenar de forma autónoma distintas fases del incidente una vez puesto en funcionamiento.

La propia AEPD dio a conocer el caso el pasado 14 de septiembre a través de un artículo de su adjunto, Francisco Pérez Bes, en el que precisa que se trata de la primera notificación recibida por el organismo de estas características, una formulación importante porque no permite afirmar que sea necesariamente el primer ataque de este tipo ocurrido en España.

Según la información comunicada por la organización afectada, el agente utilizó un conocido modelo de lenguaje, cuya identidad no ha sido difundida, e inició una búsqueda de vulnerabilidades en archivos genéricos. Posteriormente realizó un inicio de sesión correcto y, una vez dentro del sistema, continuó explorando de forma autónoma posibles fallos de la aplicación. Esa actividad le habría permitido finalmente modificar datos personales y acceder a facturas.

Un agente que adapta sus acciones sobre la marcha

La principal diferencia respecto a otros usos maliciosos de la inteligencia artificial reside en el grado de autonomía.

Hasta ahora, los modelos generativos se han utilizado para tareas como redactar mensajes de *phishing*, traducir campañas fraudulentas, facilitar suplantaciones de identidad, analizar código o ayudar a localizar vulnerabilidades. En estos casos, normalmente actúan como herramientas manejadas de forma continua por una persona.

Un agente de IA, en cambio, puede recibir un objetivo general y, a partir de él, planificar tareas intermedias, utilizar distintas herramientas, ejecutar acciones, interpretar los resultados y modificar su estrategia según lo que va encontrando.

En el incidente comunicado a la AEPD, esa capacidad habría permitido que el sistema avanzase por diferentes fases del ataque sin que una persona tuviera que indicarle manualmente cada siguiente paso.

La AEPD pide prudencia: el incidente sigue bajo análisis

La Agencia introduce, no obstante, una cautela relevante. Los datos disponibles proceden inicialmente de la notificación realizada por la propia organización afectada y deberán ser examinados antes de establecer conclusiones definitivas sobre el funcionamiento y alcance del ataque.

Tampoco existe evidencia, a partir de la información difundida, de que el modelo de lenguaje utilizado o la infraestructura de su proveedor fueran comprometidos. El hecho de que un agente utilizase un determinado modelo no significa que esa herramienta hubiera sido creada con fines maliciosos ni que su desarrollador participara en el ataque.

La AEPD tampoco ha hecho pública la identidad de la organización afectada ni el modelo de inteligencia artificial empleado.

La IA no crea necesariamente nuevas amenazas, pero las acelera

Para la Agencia, la principal consecuencia de este tipo de incidentes no es tanto la aparición de técnicas totalmente nuevas como la capacidad de la inteligencia artificial para automatizar, acelerar y adaptar ataques que ya existen.

Según la AEPD, el uso de agentes puede aumentar la velocidad y escala de una ofensiva, reduciendo al mismo tiempo el margen disponible para que los equipos de ciberseguridad puedan detectar lo que está sucediendo y reaccionar.

Francisco Pérez Bes advierte además de que una única notificación no permite establecer todavía una tendencia estadística, aunque sí constituye una señal de que los ataques apoyados en agentes de inteligencia artificial están comenzando a materializarse sobre tratamientos reales de datos personales. La propia AEPD ha subrayado públicamente esta cautela.

Qué es una brecha de datos personales

La normativa de protección de datos define una brecha como un incidente de seguridad que provoca la destrucción, pérdida o alteración accidental o ilícita de información personal, o bien su acceso o comunicación sin autorización.

El Reglamento General de Protección de Datos (RGPD) establece que las organizaciones deben comunicar a la autoridad de control determinadas brechas cuando puedan suponer un riesgo para los derechos y libertades de las personas afectadas.

La AEPD dispone de un sistema específico para recibir estas comunicaciones y publica periódicamente estadísticas sobre los incidentes notificados.

Un cambio en la gestión de la ciberseguridad

El caso supone sobre todo un aviso para empresas y administraciones que utilizan servicios digitales y gestionan información personal.

Las medidas tradicionales —protección de credenciales, actualización de aplicaciones, control de accesos o monitorización de actividad— siguen siendo necesarias, pero la aparición de sistemas capaces de explorar y explotar vulnerabilidades de forma automatizada obliga a reducir también los tiempos de detección y respuesta.

La AEPD considera que esta nueva capacidad modifica el escenario de gestión del riesgo: un atacante puede

delegar determinadas tareas en sistemas capaces de decidir cómo continuar en función de las defensas que encuentran.

Por ahora, el organismo mantiene el episodio bajo análisis. Lo que sí puede afirmarse con los datos oficiales disponibles es que se trata de la primera brecha de datos personales notificada a la AEPD en la que un agente de inteligencia artificial habría ejecutado autónomamente diferentes fases de un ataque, accediendo finalmente a datos personales y facturas.