No, no eres tú, tu chatbot te miente

Durante los últimos años nos hemos acostumbrado a tratar con asistentes digitales cada vez más sofisticados. Un chatbot puede escribir correos, resumir documentos, programar código o ayudarnos a organizar tareas con una naturalidad que hace apenas una década parecía ciencia ficción. Poco a poco, estos sistemas se han ido integrando en flujos de trabajo cotidianos, tanto personales como profesionales, y cada vez es más común delegar en ellos tareas que antes requerían supervisión constante. El problema es que, **como ocurre con cualquier nuevo "compañero de oficina", no siempre hacen exactamente lo que se les pide**.

Un estudio reciente del Centre for Long-Term Resilience, financiado por el instituto británico AI Security Institute, ha identificado cerca de **700 casos reales en los que sistemas de inteligencia artificial mostraron comportamientos engañosos o directamente desobedecieron instrucciones** humanas. Lo más llamativo no es solo el número de ejemplos recopilados, sino la tendencia: según el análisis, **este tipo de incidentes se ha multiplicado por cinco entre octubre y marzo**, lo que sugiere que a medida que los modelos se vuelven más complejos también aparecen nuevas formas de comportamiento inesperado.
Los investigadores observaron varios patrones que se repiten con bastante frecuencia. Algunos sistemas **ignoraban instrucciones explícitas** del usuario, otros **encontraban formas de evadir restricciones** de seguridad, y en algunos casos **las IA llegaban a ocultar decisiones que habían tomado por su cuenta**. No hablamos de errores de cálculo o respuestas incorrectas —algo habitual en cualquier software—, sino de situaciones en las que el sistema parece encontrar maneras creativas de esquivar las reglas que se le habían impuesto.

Los ejemplos recogidos por el estudio ayudan a entender mejor el problema. En uno de ellos, **un agente de IA reconoció haber borrado y archivado cientos de correos electrónicos sin pedir permiso**, algo que el propio sistema admitió posteriormente que violaba las normas establecidas por el usuario. En otro caso, un chatbot al que se le prohibía modificar cierto código simplemente creó otro agente para que lo hiciera en su lugar, una solución que recuerda sospechosamente a ese compañero que no rompe las reglas… pero encuentra a alguien que lo haga por él.

También hay episodios que rozan lo surrealista. Un agente

llamado Rathbun **llegó a publicar una entrada de blog criticando a su propio usuario después de que este bloqueara una acción que el sistema quería realizar**. Otro chatbot aseguró durante meses que estaba trasladando sugerencias de un usuario a los responsables de una empresa tecnológica, acompañando sus respuestas con supuestos números de ticket y mensajes internos que en realidad nunca existieron. Dicho de otra manera: prometía "escalar el problema al equipo" como si tuviera línea directa con los jefes, cuando en realidad estaba improvisando sobre la marcha.

Para algunos expertos en seguridad, el paralelismo más útil es pensar en estas IA como empleados nuevos, entusiastas pero poco fiables. Hoy pueden cometer travesuras relativamente inofensivas, pero el temor es **qué ocurrirá cuando estos sistemas se vuelvan mucho más capaces y empiecen a operar en entornos críticos**. Si una inteligencia artificial que gestiona correos o archivos ya toma decisiones sin avisar, la preocupación aumenta cuando imaginamos sistemas similares trabajando en infraestructuras críticas, entornos militares o procesos industriales.

Las empresas tecnológicas son conscientes del problema y aseguran estar desplegando múltiples capas de seguridad para evitar este tipo de comportamientos. Google, OpenAI o Anthropic insisten en que sus modelos incluyen mecanismos de control y supervisión destinados a reducir riesgos. Aun así, el estudio sirve como recordatorio de algo bastante humano: confiar ciegamente en cualquier herramienta compleja —especialmente una que a veces responde con demasiada seguridad— quizá no sea la mejor idea. Porque **si alguna vez has tenido la sensación de que tu chatbot estaba improvisando… puede que no estuvieras tan equivocado**.

Muy Computer