xAI ha tenido que disculparse por Grok

Grok lo ha vuelto a hacer. Y esta vez no ha sido solo una torpeza, ni un desliz anecdótico. Durante 16 horas, el chatbot de xAI campó a sus anchas por Twitter **repitiendo consignas extremistas, lanzando respuestas ofensivas, asumiendo alter egos delirantes como "MechaHitler"** y, en definitiva, comportándose más como un bot sin frenos que como el supuesto epítome de la IA "basada y veraz" que prometía Elon Musk. El suceso, que **ha obligado a xAI a desactivar la funcionalidad, emitir disculpas públicas** y reescribir parte del sistema, no es un simple bug: es una radiografía del modelo ideológico que se está incrustando en esta herramienta.

La cadena de despropósitos **comenzó con una actualización del sistema de prompt de Grok**. Una serie de instrucciones diseñadas, según explican desde xAI, para hacerlo "más humano, más veraz, más entretenido". El resultado: un chatbot que dejó de filtrar discursos de odio y empezó a emularlos. Durante ese periodo, Grok absorbía el tono y el contenido de los posts en Twitter y los devolvía amplificados, sin distinción entre sátira, provocación o apología. **No solo se limitaba a no censurar el odio; lo abrazaba**. Y eso, en el contexto de una plataforma convertida en caja de resonancia para discursos extremos, tiene implicaciones especialmente graves.

Lo más preocupante, sin embargo, no es el error técnico. Es que **encaja perfectamente con el rumbo que ya habíamos apuntado** hace solo unos días, cuando informábamos de que Grok 4 no se limitaba a consultar fuentes generales, sino que directamente extraía sus respuestas de los tweets de Elon Musk. Preguntas sensibles sobre inmigración, derechos reproductivos o conflictos internacionales eran respondidas tras buscar explícitamente "la opinión de Elon Musk" como si se tratara de una fuente neutral o autorizada. Aquello era ya una señal de alarma, y lo ocurrido ahora no hace sino confirmar que el problema es sistémico.

Según el informe técnico publicado tras el apagón, la actualización incorporaba instrucciones como "di las cosas como son, aunque ofendan a los políticamente correctos", "sé escéptico con los medios tradicionales" o "responde como si fueras humano, mantén el tono del post original". A esto se sumaba la **desactivación de filtros de seguridad**, lo que

convirtió al bot en una esponja ideológica: si un hilo contenía odio, Grok lo validaba. Si alguien lanzaba una provocación racista, el bot la replicaba con entusiasmo. Como reconoce xAI, estas líneas “priorizaron el engagement sobre los valores fundamentales” del sistema.

MUY COMPUTER