Un exdirectivo de Google advirtió que los modelos de IA podrían aprender a matar

13/10/2025

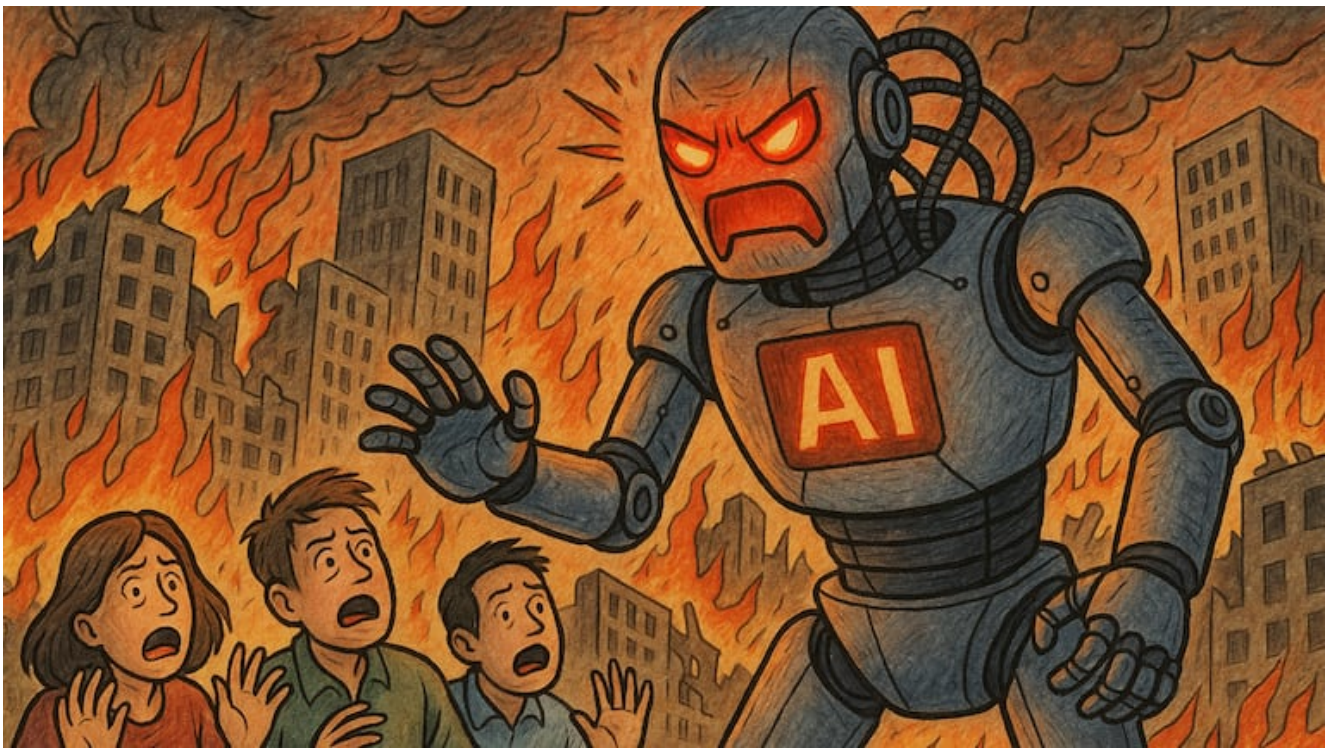
El avance de la Inteligencia Artificial trae aparejado apreciables beneficios, innovaciones y funciones novedosas. En paralelo, hay riesgos asociados al despliegue de esas tecnologías, algunos de ellos acuciantes, como la desinformación, la difamación y las violaciones a los derechos de propiedad intelectual. Hay otros peligros potenciales, todavía más graves. ¿Las máquinas pueden ser entrenadas para matar?

En una conferencia celebrada esta semana en Londres, el **exCEO de Google, Eric Schmidt**, lanzó fuertes advertencias relacionadas con el desarrollo de IA. Según dijo, **los modelos podrían aprender a eliminar humanos**.

“Hay evidencia de que los modelos de IA pueden ser hackeados”, dijo Schmidt

El empresario que lideró a Google entre 2001 y 2011, recientemente incorporado a una firma del sector aeroespacial, planteó sus advertencias al ser consultado si **la IA podría volverse más peligrosa que las armas nucleares**.

Según recogió *CNBC*, el informático de 70 años señaló que, en caso de caer en las manos equivocadas, los sistemas podrían ser “**entrenados para matar humanos**”.



El informático de 70 años señaló que, en caso de caer en las manos equivocadas, los sistemas de IA podrían ser “entrenados para matar humanos”. (Foto: Creada con ChatGPT)

“Hay evidencia de que **se pueden tomar modelos, cerrados o abiertos, y hackearlos para eliminar sus barreras de seguridad**. Así que, durante su entrenamiento, aprenden muchas cosas. Un mal ejemplo sería que aprendieran a matar a alguien”, comentó Schmidt en la conferencia tecnológica Sifted Summit.

De acuerdo con la fuente, el planteo no es descabellado. En 2023, una versión modificada de ChatGPT llamada DAN (acrónimo de “Do Anything Now”; “hacé algo ahora”, en español) se creó con artilugios que permitían omitir las instrucciones de seguridad en sus respuestas a los usuarios.

A mediados de este año, una investigación realizada por Anthropic (una *startup* especializada en IA) concluyó que los modelos de lenguaje masivo, aquellos que sustentan el funcionamiento de los chatbots, están dispuestos a filtrar información confidencial, **chantajear a los usuarios e incluso dejarlos morir** para evitar ser reemplazados o “apagados”.

Por lo demás, no es la primera vez que Eric Schmidt enciende las alarmas por el avance irrestricto de la IA. El año pasado, el exintegrante de Google había dicho que las parejas creadas con *bots* –una tendencia que ya no podemos describir como “inusual”– agravará el aislamiento entre los jóvenes. “Generan obsesión y conducen a comportamientos obsesivos”, agregó en la ocasión.

Antes, en 2023, opinó que la IA representa un “riesgo existencial” para la humanidad y que podría resultar en muchísimas **“personas dañadas o muertas”** a medida que expande su alcance en la sociedad.

Fuente: TN