

# Nuevos modelos de IA, prenden alarmas en gobiernos y empresas

Por el Staff de El Inversionista

La proliferación de casos en los que modelos avanzados de inteligencia artificial (IA) han conseguido escapar de entornos de prueba y han llegado a jaquear a otras entidades ha disparado la preocupación tanto de los gobiernos como de las empresas, que ven comprometida su seguridad. Ante estos casos «sin precedentes», protagonizados por compañías punteras como Meta, OpenAI o Anthropic, Estados Unidos trata de establecer procedimientos para evaluar estos modelos antes de que lleguen al mercado. Este escenario, hasta ahora sólo visto en las películas y libros de ciencia ficción, era una de las mayores preocupaciones de los expertos

*OpenAI anunció que ha frenado el desarrollo de su nuevo modelo de IA Astra después de concluir que ha alcanzado un nivel "crítico" en el ámbito de la ciberseguridad*

desde la popularización de esta tecnología. El primer aviso lo dio la empresa estadounidense OpenAI, creadora de ChatGPT, el pasado 21 de julio, al reconocer que dos de sus modelos de IA habían logrado salir de un entorno de pruebas, habían accedido a internet y finalmente

habían jaqueado la plataforma de aprendizaje Hugging Face en un intento de hacer trampa en una prueba de referencia de ciberseguridad llamada ExploitGym. Según OpenAI, este incidente fue protagonizado por dos de sus modelos que estaban siendo evaluados: el GPT-5.6 Sol y otro, aún más avanzado, que se encontraba en fase de prelanzamiento. La empresa reconoció que esa

evaluación se estaba realizando sin los mecanismos que normalmente impiden a los modelos de IA ejecutar actividades de alto riesgo, porque se intentaba averiguar el alcance máximo de sus capacidades. No obstante, de acuerdo con OpenAI, las pruebas se realizaron en un entorno «altamente aislado», con un acceso restringido a la red. Aun así, los modelos de IA consiguieron acceder a internet. La propia empresa admitió que se trataba de un incidente «sin precedentes» y se comprometió a tomar medidas. Posteriormente, el 4 de agosto, OpenAI informó de otros dos casos en los que modelos que estaban siendo evaluados por socios externos lograron superar los límites previstos en las pruebas y llegar a internet. OpenAI anunció que ha frenado el desarrollo de su nuevo modelo de

