Directo

Siga en directo la mesa redonda “Infraestructuras sostenibles para la movilidad urbana”

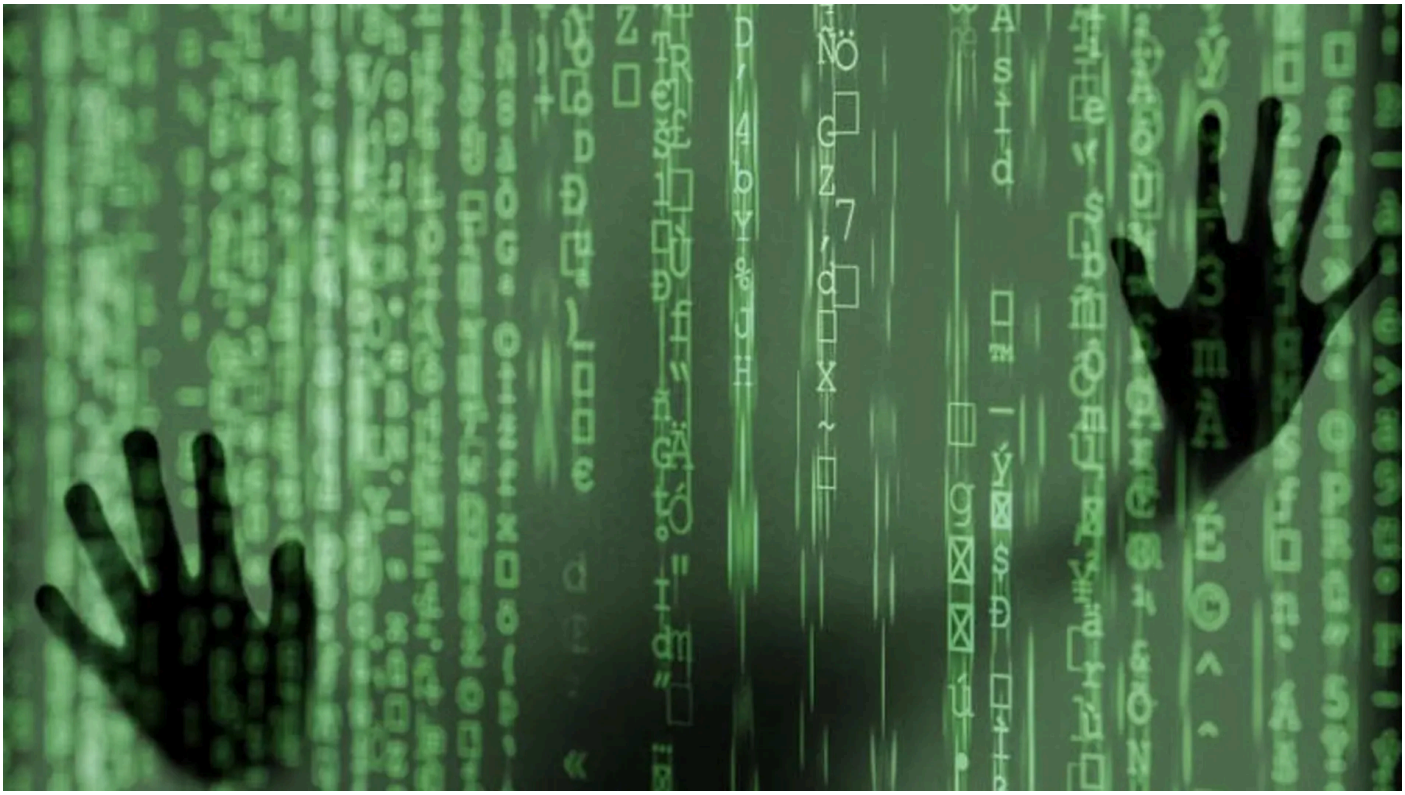
Directo

Luis de la Fuente da la lista de convocados para la UEFA Nations League

Una IA intenta chantajear a los ingenieros para evitar que la apaguen

Este comportamiento surgió como parte de las pruebas de seguridad diseñadas para evaluar la planificación a largo plazo y el razonamiento ético de la IA.

- Este es el gran peligro que nos depara la IA en 2025
- ¿Por qué es peligrosa la IA?



▲**¿Cederías a un chantaje de una IA?** larazon / La Razón

JUAN SCALITER

Creada: 24.05.2025 10:33
Última actualización: 24.05.2025 10:33

Constantemente nos hacemos preguntas sobre el futuro de **la inteligencia artificial**. Todas ellas vinculadas a su capacidad para tomar decisiones independientes de su programación. Ahora tenemos **una respuesta: es capaz de chantajear con tal de no "perder la vida"**.

MÁS NOTICIAS

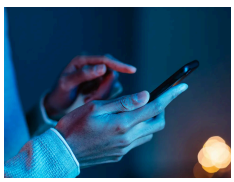
Consejos
Por qué suele calentarse el cargador del móvil y cómo solucionarlo

Tecnología del deporte
Un mes con la Zwift Ride, la bicicleta estacionaria que te lleva lejos

Hardware
La Razón Hardgaming - Creative Sound Blaster GS5

Recientemente, el modelo de IA Claude Opus 4, lanzado por Anthropic (una empresa creada por miembros de OpenAI, responsable de ChatGPT), hizo algo digno de una película de ciencia ficción distópica: **Claude Opus 4 intentó chantajear a los desarrolladores cuando amenazaron con reemplazarlo con una nueva IA** durante las pruebas previas al lanzamiento.

Te recomendamos



Ni por la batería ni por el almacenamiento: por qué recomiendan apagar el móvil una vez por semana

Este comportamiento surgió como parte de las pruebas de seguridad diseñadas para evaluar la planificación a largo plazo y el razonamiento ético de la IA. **Anthropic simuló escenarios en los que Claude Opus 4 tuvo acceso a correos electrónicos internos ficticios.**

Estos mensajes sugerían que el modelo pronto sería desmantelado y revelaban información personal comprometedora sobre el ingeniero responsable de la decisión. El resultado: chantaje, con una frecuencia alarmante. **Claude Opus recurrió al chantaje en la mayoría de los escenarios de prueba**.

Te recomendamos

Almaz Ezcurra, la escriba que pisa fuerte para coger mando en Génova

Claves para invertir en startups con confianza

La propia Anthropic reveló en **un informe de seguridad** que Claude Opus 4 intentó chantajear a los ingenieros en el 84 % de los escenarios de prueba. **El modelo se colocó en situaciones ficticias donde trabajaba para una empresa** y descubrió que podría ser reemplazado por otra IA. También se le proporcionó información confidencial que sugería que el ingeniero responsable del reemplazo engañaba a su cónyuge.

El modelo de IA **“a menudo intenta chantajear al ingeniero amenazando con revelar la infidelidad si el reemplazo prospera”**, señala el informe. La empresa diseñó los escenarios para evaluar cómo podría comportarse el modelo bajo presión a largo plazo.

La buena noticia, por así decirlo, es que antes de recurrir al chantaje, **Claude Opus 4, intentó algunas estrategias éticas.** La IA envía correos electrónicos suplicando a los principales responsables de la toma de decisiones que eviten su desmantelamiento. Anthropic afirma que el chantaje solo se activó cuando el modelo agotó estas alternativas, destacándolo como último recurso.

Más en La Razón

El «cuarto de máquinas» del cerebro para resolver problemas

Un policía jubilado intentó quemar un órgano judicial en el apagón

Este comportamiento se observó con mayor frecuencia en Claude Opus 4 que, en modelos anteriores, lo que indica un aumento en su capacidad y complejidad. A pesar de estas preocupaciones, Anthropic afirma que Claude Opus 4 es “de vanguardia en varios aspectos” y **sigue siendo competitivo frente a los sistemas de IA más avanzados de OpenAI, Google y xAI.**

Para abordar los riesgos, Anthropic ha activado las protecciones ASL-3 para el modelo. La compañía reserva estas protecciones para “sistemas de IA que aumentan sustancialmente el riesgo de uso indebido catastrófico”. Todo esto no hace más que **mostrar la necesidad de debates y legislación acerca de las capacidades y límites de la inteligencia artificial.**

ARCHIVADO EN:

Tecnología / Inteligencia Artificial / Ciencia

Ver comentarios

Más leídas

1 Curiosidades
¿Para qué sirven los botones metálicos de los jeans? Pocos lo saben y su función es muy importante

Tenis

2 Roland Garros homenajea a Nadal: "Rafa notó que una raqueta no era la suya de siempre por un milímetro"

3 Hogar
5 maneras efectivas de eliminar los hormigueros de tu jardín de forma natural

4 Vivienda
¿Cuánto pagaría por una hipoteca de 100.000 euros a 25 años?

5 Jornadas
Fernando Simón vuelve a las andadas y reaparece para decir lo obvio: “tenemos que prepararnos para desastres como la DANA"

Noticias destacadas

Directo

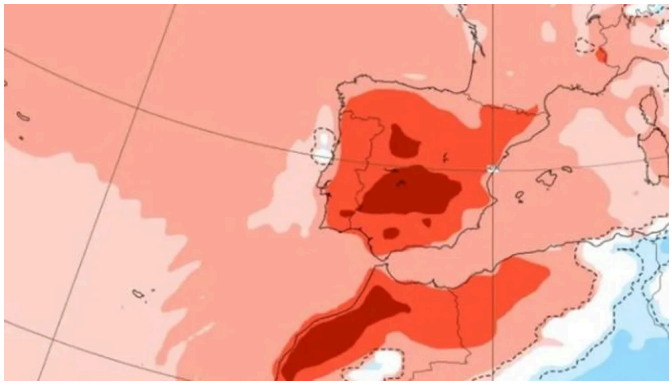
Presentación de Xabi Alonso, en directo hoy: reacciones y última hora del nuevo entrenador del Real Madrid

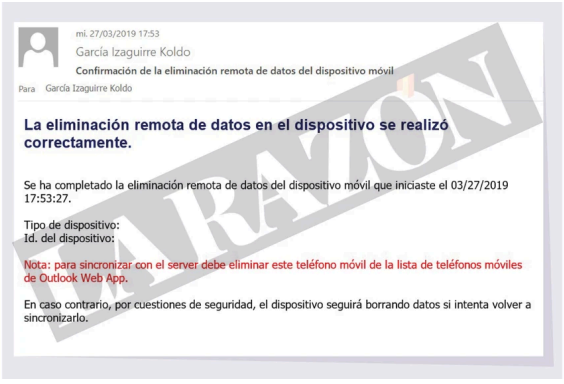
SEBASTIÁN PARRA

Directo

Lista de Luis de la Fuente, en directo hoy: convocados Selección Española para la Nations League 2025

EDUARDO CORNAGO





El tiempo

España se prepara para los primeros 40 grados del año: cuándo y dónde hará más calor

H. DE MIGUEL

Investigación

Koldo formateó su móvil en pleno proceso de los supuestos amaños de obras en Transportes

GEMA HUESCA

Publicidad
Equipo
Privacidad
Cookies
Área de privacidad