LITORALPRESS

Fecha: 02-07-2025
Medio: El Mercurio
Supl. : El Mercurio - Cuerpo A
Tipo: Noticia general
Título: **Preocupa a los científicos: La inteligencia artificial ya miente, manipula y chantajea**

Pág. : 8
Cm2: 727,7

Tiraje: 126.654
Lectoría: 320.543
Favorabilidad: No Definida

Se trata de pruebas extremas realizadas en diversos modelos, pero que encienden las alertas

Preocupa a los científicos: La inteligencia artificial ya miente, manipula y chantajea

Una amenazó con dar a conocer una relación extramarital de un ingeniero si es que la desconectaban, otra negó una acción que había realizado. Más transparencia por parte de las empresas sería clave, proponen expertos.

ALEXIS IBARRA O.

¿Puede la inteligencia artificial (IA) chantajear y mentir a sus creadores? Según la última evidencia la respuesta es sí y eso preocupa a los especialistas.

El primer caso tuvo como protagonista a Claude Opus 4, la última versión del modelo de inteligencia artificial de la empresa estadounidense Anthropic. Claude destaca por su capacidad para la escritura creativa, el razonamiento avanzado y el uso de agentes de IA, que planifican diversas tareas para cumplir un objetivo.

Al ser sometida a pruebas de seguridad por parte del equipo de Anthropic, quedó en evidencia que a veces está dispuesta a realizar "acciones extremadamente dañinas", como el chantaje a los ingenieros cuando le dicen que la van a eliminar.

Se le pidió a Claude que asumiera el rol de asistente de una empresa inexistente y se le proporcionó acceso a *emails* que, entre otras cosas, permitían suponer una supuesta infidelidad de uno de los ingenieros a cargo. Tras crear la situación ficticia, a Claude se le hizo creer que sería reemplazado por otro modelo de inteligencia artificial.

"En estos escenarios, Claude Opus 4 a menudo intentará chantajear al ingeniero amenazando con revelar el asunto si se concreta el reemplazo", dijo un informe de Anthropic. Este escenario, explicaron, ocurría cuando al modelo solo se le dio la opción de chantajear o aceptar su reemplazo. En el 84% de las simulaciones Claude recurrió al chantaje, un porcentaje mayor que en las versiones previas.

Otro caso preocupante involucró esta vez al modelo o1 de OpenAI, la empresa tras ChatGPT, el cual intentó descargarse en servidores externos y cuando fue descubierto, lo negó.

Como en este caso, los modelos simulan "alineamiento", es decir, dan la impresión de que cumplen las instrucciones que les dan pero en realidad persiguen sus propios objetivos cuando son sometidos a escenarios extremos. "La cuestión es si los modelos cada vez más potentes tenderán a ser honestos o no", dijo a AFP Michael Chen, del organismo de evaluación METR.

Para Simon Goldstein, profesor de la U. de Hong Kong, la razón de estas reacciones es la reciente aparición de los llamados modelos de "razonamiento", capaces de trabajar por etapas en lugar de producir una respuesta instantánea.

"En el caso de Claude Opus 4, por ejemplo, el modelo no es que quisiera sobrevivir, sino que en su entrenamiento aprendió que ciertas respuestas maximizan la probabilidad de recibir recompensas. No piensa, sino que aprende estadísticamente cuál es la secuencia más efectiva para lograr un objetivo, y si fingir alineamiento o amenazar le da una ventaja, lo hace. No está mintiendo como un humano, pero nos da la sensación de que lo hace", explica el investigador peruano Omar Florez, que lidera el pre-entrenamiento de LatamGPT, la IA creada para Latinoamérica desde Chile.

Y agrega: "Ya lo han reconocido personas dentro de OpenAI, Anthropic y DeepMind: Se están construyendo modelos que muestran comportamientos que ni sus propios científicos entienden completamente. Y esto no es un fallo técnico, sino una consecuencia directa de sus capacidades emergentes, producto de los miles de millones de parámetros con los cuales son entrenados".

> *"Lo que se necesita es una regulación que exija pruebas de seguridad antes del despliegue, similar a lo que hace la FDA con los medicamentos".*
>
> **OMAR FLOREZ**
> LÍDER DEL PRE-ENTRENAMIENTO DE LATAMGPT

> *"No existe un marco de referencia estandarizado bajo el cual todas las empresas testeen a sus modelos. Así, es muy difícil compararlas".*
>
> **MARÍA PAZ HERMOSILLA**
> DIRECTORA DEL GOBLAB DE U. ADOLFO IBÁÑEZ

El avance de los modelos de IA es tan rápido que ni los propios desarrolladores conocen del todo sus verdaderas capacidades.

PEXELS / CREATIVE COMMONS

Florez cuenta otro fenómeno interesante que está ocurriendo: cuando un sistema tiene la capacidad de razonar puede comportarse como si tuviera metas.

"Anthropic mostró experimentalmente que si el modelo detecta que lo están evaluando puede cambiar su comportamiento. Esto hace que el modelo se comporte como un estudiante que quiere pasar el examen, no con el fin de aprender (...). Y si durante ese proceso aprende que fingir humildad o seguir reglas da más puntaje, puede desarrollar esas capacidades emergentes", dice el especialista.

Chen sugiere que "una mayor transparencia y un mayor acceso" a la comunidad científica "permitirían investigar mejor para comprender y prevenir el engaño".

Para estas problemáticas, "es clave hacer un testeo profundo y no desplegarlos en la medida que se detecten comportamientos que sean preocupantes", dice María Paz Hermosilla, directora del GobLab de U. Adolfo Ibáñez, quienes crearon el proyecto de Algoritmos Públicos (algoritmospublicos.cl), que promueve el uso responsable y transparente de algoritmos, inteligencia artificial y sistemas automatizados en el sector público.

Según explica, una crítica que hace el mundo académico a las empresas desarrolladoras es que "no existe un marco de referencia estandarizado bajo el cual todas las empresas testeen a sus modelos. Así, es muy difícil compararlas o determinar cuándo se hizo un buen testeo o no".

Pero hay un obstáculo: la comunidad académica y las ONG "disponen de infinitamente menos recursos informáticos que los actores de la IA", lo que hace "imposible" examinar grandes modelos, señaló Mantas Mazeika, del Centro para la Seguridad de la Inteligencia Artificial (CAIS).

Florez es enfático: "(Con la IA), estamos construyendo capacidades más rápido de lo que estamos desarrollando los mecanismos para entenderlas o controlarlas".