

Las empresas de IA defienden sus controles de seguridad ante el Concejo de Nueva York

Las principales empresas de inteligencia artificial (IA) defendieron este lunes ante el Concejo Municipal de Nueva York sus sistemas de seguridad, pero evitaron comprometerse de forma inequívoca a no desplegar modelos que no superen una evaluación independiente.

Durante la audiencia, la presidenta del Concejo, Julie Menin, pidió bajo juramento a los representantes de OpenAI, Anthropic, Meta y Google que se comprometieran a no lanzar un modelo si fallaba una prueba interna o una evaluación de terceros no seleccionados por la propia empresa.

La representante de OpenAI, Morgan Dwyer, dijo al respecto que la compañía «no va a lanzar modelos que no consideremos seguros» y defendió que sus evaluaciones cuentan con terceros y con la colaboración de los Gobiernos de Estados Unidos y Reino Unido.

Agregó que la empresa ha retrasado lanzamientos para reforzar las salvaguardias, algo que «volveremos a hacer» si es necesario.

Mientras, el representante de Anthropic, Logan Graham, explicó que su empresa lleva cuatro años trabajando en la cuantificación de riesgos, incluidos los relacionados con la bioseguridad y la ciberseguridad, pero reconoció que se trata de un proceso «intrincado y lleno de matices».

Graham afirmó que Anthropic publica informes en los que explica «qué probabilidad creemos que tienen determinados tipos de estas amenazas de llegar a producirse» y defendió las evaluaciones independientes como «uno de los componentes más importantes» del proceso.

Mientras, el representante de Meta, Shane Cahill, describió un sistema de seguridad «de varias capas» y, preguntado por la posibilidad de cuantificar los riesgos, dijo que no quería «ser impreciso».

Por su parte, la representante de Google, Alice Friend, explicó que la compañía utiliza marcos específicos para detectar capacidades peligrosas, pero admitió que predecir el riesgo catastrófico futuro «no es una ciencia perfecta».

Friend agregó que «no existe actualmente una base científica suficientemente rigurosa para asignar probabilidades precisas a determinados riesgos catastróficos».

En otro momento de la audiencia, Dwyer reconoció además que los compromisos voluntarios y la autorregulación no son suficientes para garantizar la seguridad de la IA y aseguró que OpenAI apoya una regulación específica para los llamados modelos fronterizos.

Menin, por su parte, advirtió de las «serias preocupaciones de seguridad existencial» asociadas al desarrollo de esta tecnología.

También defendió que la IA no puede quedar sometida únicamente a la autorregulación de las empresas y abogó por establecer medidas de seguridad. EFE

Alberto News