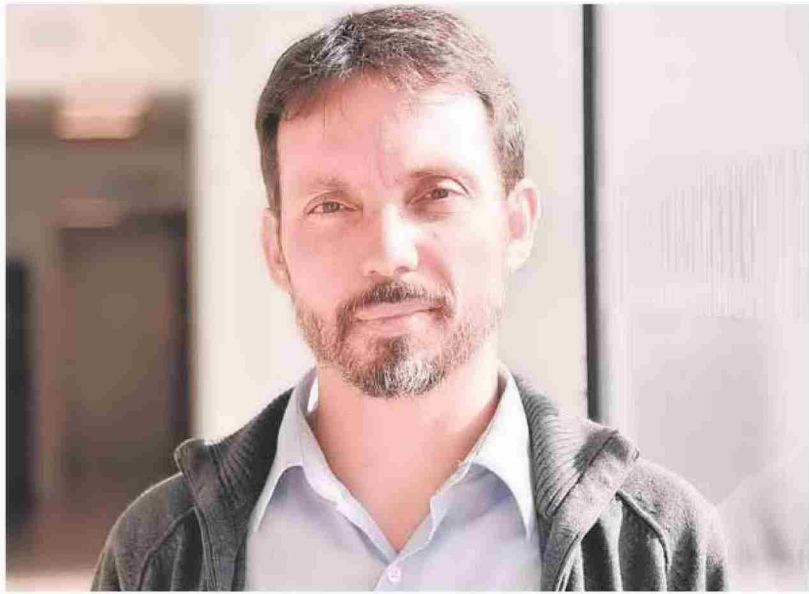
Tres investigadores de la UPV cuestionan los resultados de la inteligencia artificial

Un grupo de 16 expertos detecta variaciones de fiabilidad en los análisis de personas de una etnia o grupo demográfico concreto

S. V.



El profesor Hernández-Orallo. LP

VALENCIA. «Los recientes avances en inteligencia artificial (IA) basada en sistemas que requieren enormes cantidades de datos y cálculo, como GPT-4, han puesto de manifiesto las dificultades para comprender las capacidades y debilidades de estos sistemas de inteligencia artificial. No sabemos dónde es seguro utilizar estos sistemas ni cómo podrían mejorarse. Y esto se debe a la forma en que se evalúa hoy la IA, que requiere de un cambio urgente».

Detrás de estas afirmaciones se encuentran 16 de los principales expertos en inteligencia artificial de todo el mundo, entre ellos los investigadores del Instituto VRAIN de la Universitat Politècnica de València (UPV), José Hernández-Orallo, Fernando Martínez Plumed y Wout Schellaert.

Coordinados por el profesor Hernández-Orallo, los 16 investigadores han publicado una carta en la revista Science en la que reclaman la necesidad de «repensar» la evaluación de las herramientas de inteligencia artificial para avanzar hacia unos modelos más transparentes y saber cuál es su eficacia real, qué es lo que pueden y no pueden hacer.

En su escrito, los autores de l estudio científico proponen una hoja de ruta para los modelos de inteligencia artificial, en la que sus resultados se presenten de forma más matizada y los resultados de la evaluación caso por caso se pongan a disposición del público.

Según explica Hernández-Orallo, el rendimiento de un modelo de IA se mide con estadísticas agregadas. Y esto supone un riesgo, porque si bien pueden dar una visión de su buen rendimiento global, pueden ocultar también una baja fiabilidad/utilidad en casos concretos, más minoritarios, «y sin embargo se da a entender que es igualmente válido en todos los casos cuando en realidad no es así».

Los expertos advierten de la necesidad de repensar la evaluación de los sistemas de inteligencia artificial

Etnias o grupos demográficos

En el documento, los firmantes lo explican con el caso de modelos de IA de ayuda al diagnóstico clínico y señalan que estos sistemas podrían tener un problema cuando analizan a personas de una etnia o grupo demográfico concreto, porque son casos que constituyeron sólo una pequeña proporción de su entrenamiento.

«Lo que pedimos es que cada vez que se publique un resultado en IA, se desglose lo máximo posible, porque si no se hace, no es posible saber su utilidad real y reproducir el análisis. En el artículo publicado en Science hablamos también de un sistema de IA de reconocimiento facial que daba un 90% de acierto, y después se comprobó que para hombres blancos el porcentaje de acierto era del 99.2%, pero para mujeres negras solo llegaba al 65,5%», apunta José Hernández-Orallo.

«Por ello, en algunas ocasiones, los resultados que se venden sobre la utilidad de una herramienta de IA no son del todo transparentes y fiables. Si no te dan el detalle crees que los modelos funcionan muy bien y no es la realidad. No tener ese desglose con toda la información posible sobre el modelo de IA supone que aplicarlo podría comportar riesgos», añade el profesor.

Buen uso de la tecnología

El investigador de VRAIN de la Universitat Politècnica de València destaca que los cambios que proponen pueden contribuir a mejorar la comprensión en la intelgencia artificial. Y también a reducir la «voraz» competición que existe actualmente entre los laboratorios de IA por anunciar que su modelo mejora un tanto por cien otros sistemas anteriores.

«Hay laboratorios que quieren pasar del 93% al 95% como sea y eso va en contra de la aplicabilidad y fiabilidad final de la IA. Lo que queremos, en definitiva, es contribuir a que, entre todos, entendamos mejor cómo funciona la IA, cuáles son las limitaciones de cada modelo, para garantizar un buen uso de esta tecnología», concluye Hernández-Orallo.

Junto a los investigadores del Instituto VRAIN de la Politècnica de València, en este artículo ha participado también personal investigador de la Universidad de Cambridge, la Universidad de Harvard, el Instituto Tecnológico de Massachusetts (MIT), la Universidad de Stanford, Google, el Imperial College de Londres, la Universidad de Leeds, el Instituto Alan Turing de Londres, Deepmind, el Instituto Nacional de Estándares y Tecnología de EE.UU. (NIST), el Instituto Santa Fe, la Universidad Tongji de Shanghái y la Universidad Shandong de Jinan.