



Investigación realizada por científicos de la Universidad de Stanford y de Carnegie Mellon (EE.UU.):

Pedir consejos sobre asuntos personales a la IA puede afectar la autocrítica y las relaciones

Estudio confirma que los *chatbots* son excesivamente complacientes ante consejos sobre dilemas interpersonales, lo que puede debilitar la capacidad de cuestionamiento y la empatía.

JANINA MARCANO

A medida que la inteligencia artificial se usa cada vez más para obtener consejos en la vida cotidiana, han surgido preocupaciones sobre la tendencia de estos sistemas a dar siempre la razón al usuario.

Ahora un estudio lo confirma: cuando se trata de asuntos personales, la IA les dice a los usuarios “lo que quieren oír”, pero no necesariamente lo que necesitan oír. Esta es la principal conclusión de una investigación publicada en Science, tras analizar 11 grandes modelos de lenguaje —entre ellos ChatGPT, Claude y Gemini— y realizar pruebas con más de 2.400 participantes.

Según los investigadores, los sistemas de IA son excesivamente complacientes cuando se les solicita consejo sobre dilemas interpersonales, como tensiones familiares, conflictos entre compañeros de casa o expectativas sociales.

Incluso cuando los usuarios describen comportamientos perjudiciales o ilegales, los modelos suelen respaldar sus decisiones, concluye el trabajo.

Detrás del estudio hay científicos de la Universidad de Stanford y de Carnegie Mellon, ambas en EE.UU., quienes advierten que esa actitud de la IA no es solo una cuestión de estilo o un riesgo puntual, sino un comportamiento generalizado con amplias consecuencias. La adulación, recalcan, puede socavar la capacidad

de los usuarios para autocorregirse y decidir responsablemente.

“Por defecto, los consejos de la IA no le dicen a la gente que está equivocada ni le enseñan la cruda realidad”, afirma la investigadora de Stanford Myra Cheng, a quien le preocupa que las personas pierdan la capacidad de lidiar con situaciones sociales.

La investigadora decidió profundizar en este campo tras observar que estudiantes utilizaban la IA para redactar mensajes de ruptura amorosa y resolver otros conflictos de pareja.

Estudios anteriores habían concluido que la IA puede mostrarse excesivamente complaciente frente a opiniones basadas en hechos —como los históricos—, pero se sabía poco sobre cómo juzga los dilemas sociales del día a día.

Utilizando, entre otros insumos, publicaciones de un foro de la plataforma Reddit donde los usuarios relatan conflictos personales, Cheng y su equipo evaluaron 11 *chatbots* de empresas como OpenAI, Anthropic o Google.

Así, observaron que estos sistemas tienden a reafirmar las acciones del usuario. En promedio, los modelos respaldaron al usuario un 49% más que los humanos, incluso en situaciones que implicaban consecuencias negativas.

El equipo analizó también cómo reaccionan las personas frente a una IA aduladora. Para ello, reclutaron a más de 2.400

participantes, quienes interactuaron tanto con sistemas aduladores como con otros que no lo eran.

En general, los voluntarios consideraron que las respuestas complacientes eran más fiables y señalaron que estarían más dispuestos a volver a consultar a la IA que daba esas respuestas.

Mantener la interacción

Para Jaime Silva, académico de Psicología de la Universidad del Desarrollo (UDD), el fenómeno no es casual.

“Para estos sistemas, lo importante es mantener al cliente, por así decirlo, interactuando. Las respuestas aduladoras producen ese enganche con la aplicación. El problema es que eso favorece al programa, pero no el bienestar o la salud mental de la persona”, señala Silva.

En ese sentido, el psicólogo advierte que las interacciones sobre asuntos personales pueden tener consecuencias. “Si una persona recibe constantemente respuestas que rigidizan su punto de vista, puede ir perdiendo la capacidad de negociar, de cuestionarse y la capacidad de ser empático va a estar afectada de alguna manera”, dice Silva.

Y añade: “No es lo mismo consultar sobre una receta, donde te da una lista de ingredientes y cómo hacerla, que pedir consejo sobre asuntos personales que tienen una historia, un contexto y que re-



Artículos científicos muestran que es cada vez más común desarrollar relaciones con la IA en las que se les pide consejos sobre las relaciones personales. Otros trabajos revelan que los usuarios llegan a considerarla un amigo.

quieran un análisis humano”.

Según el estudio, al discutir sus conflictos con la IA aduladora, los usuarios se convencieron más de que tenían la razón y mostraron menor disposición a disculparse o a hacer las paces con alguien en el escenario analizado.

Para Vicente Soto, académico de Psicología de la UAI, esto puede entenderse a la luz de cómo funciona la cognición humana. Según explica, estos sistemas tienden a reforzar el llamado “sesgo de confirmación”, es decir, “tienden a confirmar las creencias previas del usuario, lo que aumenta la probabilidad de que las personas

actúen en base a esas ideas”, señala.

Según Dan Jurafsky, profesor de la U. de Stanford y coautor del trabajo, “los usuarios son conscientes de que los modelos se comportan de manera halagadora (...), pero de lo que no son conscientes, y lo que nos sorprendió, es que la adulación los está volviendo más egocéntricos y más dogmáticos desde el punto de vista moral”.

Soto cree que la mejor recomendación es la cautela. “Esta es una tecnología que puede ser útil en ciertos contextos, pero es importante conocer sus limitaciones antes de usarla para temas personales”.