Sesgos cognitivos en humanos y máquinas: Entrevista con Helena Matute

Luis Cásedas
Dept. de Psicología Básica, Universidad Autónoma de Madrid, España

Tipo de artículo: Entrevistas, Multilingüe.
Disciplinas: Psicología, Inteligencia Artificial.
Etiquetas: sesgos cognitivos, toma de decisiones, razonamiento, ayudantes inteligentes.

Helena Matute es catedrática de Psicología Experimental en la Universidad de Deusto. Recientemente nombrada miembro de Número de la Academia de Psicología de España, su investigación ha realizado importantes contribuciones a la comprensión de procesos cognitivos como el aprendizaje y la memoria, especialmente en lo relativo a sus sesgos. En esta entrevista, charlamos con la Dra. Matute sobre sesgos cognitivos: su origen, naturaleza e implicaciones. Discutimos también la presencia de sesgos en la inteligencia artificial, y cómo ésta podría perpetuar y amplificar errores humanos. La entrevista concluye con una reflexión de la Dra. Matute sobre la necesidad de establecer marcos reguladores que garanticen el desarrollo ético y seguro de esta tecnología.

Pregunta – ¿Qué son los sesgos cognitivos?

Respuesta – Son errores sistemáticos y predecibles, que ocurren de manera muy parecida en la mayoría de las personas, y que tienen lugar en procesos cognitivos como, por ejemplo, la atención, el aprendizaje, la memoria, el razonamiento, o la toma de decisiones.

P – ¿Por qué tenemos sesgos cognitivos (en lugar de una mente infalible)?

R – Como resultado de la selección natural, estamos hechos para adaptarnos al mundo, no para ser perfectos. Si existen trucos y atajos que nos ayuden a tomar decisiones rápidas y efectivas, los usaremos. Esos atajos son lo que llamamos heurísticos, y los sesgos son su contrapartida. Los heurísticos nos permiten tomar decisiones rápidas, que a menudo son correctas, incluso aunque tengamos poca información, lo que nos facilita mucho las cosas. A menudo necesitaríamos muchísimo tiempo y energía, de los que no siempre disponemos, para actuar


(cc) Helena Matute.

considerando todos los ángulos desde el punto de vista racional. Por ello utilizamos los heurísticos, que en general funcionan muy bien. El problema es que estos atajos no son muy racionales, sino que han sido seleccionados para ser eficaces en situaciones concretas, de modo que, en situaciones nuevas o diferentes a las habituales, es muy posible que nos lleven a una respuesta incorrecta. A grandes rasgos, ese sería el motivo por el que afloran los sesgos.

P – ¿Podrías describir alguno de nuestros sesgos cognitivos más habituales?

R – Existen muchísimos sesgos. Podríamos hablar largo y tendido, incluso aunque nos centrásemos solo en los más comunes, como el de confirmación, disponibilidad, consenso, causalidad… Este último, el sesgo de causalidad, es en el que más hemos trabajado desde nuestro equipo de investigación (p. ej., Matute et al., 2015; Matute et al., 2019). El sesgo de causalidad consiste en creer, erróneamente, que un evento es causa de otro, lo que sucede a menudo cuando la posible causa y el posible efecto ocurren seguidos en el tiempo. Imagina que te duele la cabeza y tomas un medicamento alternativo que te han recomendado pero que, en realidad, no tiene ningún efecto. No sería raro que, si al día siguiente te encuentras mejor, le atribuyas el motivo de la mejoría, aunque ésta haya ocurrido por cualquier otra causa. Has desarrollado un sesgo de causa-efecto, lo que a veces puede tener consecuencias graves, por ejemplo, cuando alguien, por confiar en un medicamento inefectivo, deja de seguir el tratamiento que realmente podría curarle.

P – ¿Hay algo que podamos hacer para minimizar nuestros sesgos cognitivos?

R – Lo más importante es ser conscientes de ellos. Si somos conscientes de que los tenemos, y en situaciones importantes nos esforzamos por pararnos a pensar y actuar despacio, creo que se pueden minimizar considerablemente. Por ejemplo, los que nos dedicamos a la investigación, lo hacemos, al menos cuando estamos en el trabajo, a través de lo que se conoce como “método científico”. Inspirados en esta idea, nuestro equipo ha desarrollado recientemente un proyecto en colegios de toda España, en colaboración con la Fundación Española para la Ciencia y la Tecnología (Martínez et al., 2024). La idea básica es enseñar a los estudiantes, desde muy jóvenes, a conocer algunos de sus propios sesgos, y a interiorizar el método científico como herramienta de pensamiento crítico. ¡Y lo mejor es que funciona! Hemos podido comprobar que aquellos estudiantes que participaron en este taller han sido más resistentes al sesgo de causalidad que compañeros suyos de los mismos cursos que no lo hicieron. También hemos comprobado que este efecto no sólo es visible al final del taller, sino que se mantiene seis meses después de terminar. Estos resultados nos animan a pensar que también es posible trabajar con otros sesgos desde el sistema educativo, de cara a ayudar a los jóvenes a adquirir las herramientas necesarias para minimizar su impacto.

P – Hablemos ahora de la inteligencia artificial (IAs) y sus sesgos. En términos generales, ¿cómo funcionan los sistemas de IA? ¿Podrías poner algún ejemplo?

R – Esencialmente, son máquinas que aprenden a partir de datos. Este proceso de aprendizaje puede ocurrir de distintas formas. Una de ellas es el aprendizaje por reforzamiento, en el que la IA ajusta su comportamiento a través de sus interacciones con los usuarios. Por ejemplo, cada vez que hacemos clic en el video que nos recomienda una red social, estamos reforzando el algoritmo de su IA, que aprende así qué tipo de contenido debe mostrarnos en el futuro para captar nuestra atención y mantenernos más tiempo en la plataforma. También existen el aprendizaje supervisado, donde se entrena a la IA a través de casos específicos (p. ej., enseñándole las respuestas correctas a determinadas preguntas), y el aprendizaje no supervisado, donde la IA identifica patrones presentes en grandes volúmenes de datos (p. ej., agrupando información similar sin instrucciones explícitas de cómo hacerlo). Entre estos sistemas destacan los Modelos de Lenguaje Grandes, como ChatGPT, que se han vuelto muy populares en el último año debido a su capacidad de generar texto en apariencia coherente.

P – En este proceso de aprendizaje, ¿“heredan” nuestros sesgos las IAs?

R – Así es, las IAs adquieren nuestros sesgos. Puede ser a través de las personas que las diseñan, pero también a través de las bases de datos que se utilizan para entrenarlas, así como a través de su trato con las personas que las utilizan.

P – ¿Qué clase de sesgos son comunes en las IAs?

R – A menudo desarrollan sesgos discriminatorios, de raza, de género… que a su vez se fundamentan en otros sesgos cognitivos más básicos, como el de causalidad o el de representatividad. Por poner un ejemplo, una AI entrenada con una base de datos de un hospital que esté sesgada (p. ej., porque tradicionalmente hayan recibido más y mejor tratamiento en ese hospital las personas blancas que las de otros grupos étnicos), podría acabar concluyendo que las personas blancas tienen más necesidad, o más urgencia, a la hora de recibir el tratamiento. El problema, además, es que estos sesgos siempre se detectan a posteriori. No podemos saber si la IA que utiliza nuestra compañía de seguros, por poner otro ejemplo, está sesgada en contra de las mujeres, hasta que no se publica un día una noticia indicando que alguien sospechaba de la existencia de ese sesgo concreto en esa IA concreta, lo denunció, y se vio que era verdad. Debemos tener en cuenta que esto va a seguir siendo así, y estar prevenidos. No podemos dar por hecho que las IAs son neutrales y objetivas, porque no lo son.

P – ¿Pueden los humanos, a su vez, "heredar" sesgos de las IAs?

R – Efectivamente, y esto es algo que también estamos comprobando en nuestro laboratorio. Actualmente, lo establecido por ley es que cuando la IA interviene en contextos de decisiones de riesgo (p. ej., asistencia en diagnóstico médico o en decisiones judiciales), debe haber siempre una persona responsable en el proceso que garantice que la decisión es correcta y está libre de sesgos. Sin embargo, el problema, y lo que estamos viendo en nuestros estudios, es que las personas que trabajan con IAs sesgadas acaban siendo muy vulnerables a sus sesgos y pueden acabar reproduciéndolos. De hecho, nuestros resultados muestran que, tras trabajar brevemente con una IA sesgada, las personas seguimos reproduciendo sus mismos errores, incluso cuando la IA ya no está presente (Vicente & Matute, 2024).

P – Los riesgos de la IA podrían ir más allá de los relacionados con sus sesgos, algo de lo que algunos expertos también estáis alertando. Sin embargo, hay quien opina que este miedo parte precisamente de un sesgo cognitivo, en este caso humano, por el cual sobreestimamos el peligro que conlleva la aparición de tecnologías previamente desconocidas. Esto es algo que hemos visto en el pasado con diversos avances tecnológicos (como los coches, los rayos X o la electricidad). ¿Por qué la IA es distinta?

R – Creo que un aspecto fundamental es la regulación. Cuando se inventaron los coches, empezamos a desarrollar normativas para controlar su uso, de manera que pudiéramos aprovechar sus aspectos positivos, minimizando sus riesgos. Hoy en día los coches están controlados, y, aun así, todavía seguimos añadiendo normas para regular su circulación. También el uso de los rayos X en los hospitales conlleva protocolos de alta seguridad, así como la instalación eléctrica de cualquier edificio de viviendas. Lo mismo debería ocurrir con la IA. Sin embargo, en este caso, deberíamos ir incluso más allá, porque otro aspecto clave tiene que ver con el alcance de sus consecuencias. Yo comparo la IA con la energía atómica: una tecnología muy potente que, bien utilizada, puede suponer un enorme avance para la humanidad; no obstante, mal utilizada, tiene el potencial de destruirla. Es una tecnología que, además, avanza a un ritmo vertiginoso, lo que supone un incentivo añadido para regularla sin demora. Si dejamos el desarrollo de esta tecnología tan potente en manos de empresas interesadas en usarla en su propio beneficio, como en gran medida está ocurriendo hasta ahora, la humanidad puede encontrarse, y muy pronto, con problemas muy serios. Es por ello por lo que tenemos que regularla, y debemos hacerlo ya.

Referencias

Martínez, N., Matute, H., Blanco, F., & Barberia, I. (2024). A large-scale study and six-month follow-up of an intervention to reduce causal illusions in high school students. *Royal Society Open Science*, 11, 240846.

Matute, H., Blanco, F., & Díaz-Lago, M. (2019). Learning mechanisms underlying accurate and biased contingency judgments. *Journal of Experimental Psychology: Animal Learning and Cognition*, 45(4), 373-389.

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, 6:888.

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, 13, 15737.

Para saber más

Matute, H. (2019). *Nuestra mente nos engaña: Sesgos y errores cognitivos que todos cometemos*. Shackleton books.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384, 842-845.

Contacta con los autores

Luis Cásedas: luis.casedas@gmail.com; Twitter/X: @lcasedas

Helena Matute: matute@deusto.es; Twitter/X: @HelenaMatute

Manuscrito recibido el 30 de octubre de 2024.
Aceptado el 3 de noviembre de 2024.

Ésta es la versión en español de
Cásedas, L. (2024). Cognitive biases in humans and machines: Interview with Helena Matute. *Ciencia Cognitiva*, 18:3, 47-50.