FILOSOFÍA DE LA CIENCIA

POR QUÉ LA INTELIGENCIA ARTIFICIAL NO PUEDE QUERER NADA

¿Actuarán y tomarán decisiones las máquinas algún día? ¿Tendremos entonces que reconocerles intencionalidad como a los seres humanos? Una refutación

Dorothea Winter

Poco después de que no sucediese el apocalipsis supuestamente profetizado por los mayas en 2012, se popularizó otro escenario de horror: la dominación mundial de los «Terminator», es decir, de máquinas inteligentes, que, desde hace tiempo, dan lugar a noticias cada vez peores. Hasta qué punto está extendido el miedo a los superordenadores lo demuestra una encuesta realizada por la Sociedad para la Investigación Innovadora de Mercado (GIM, por sus siglas en alemán) encargada por el Grupo Bosch.

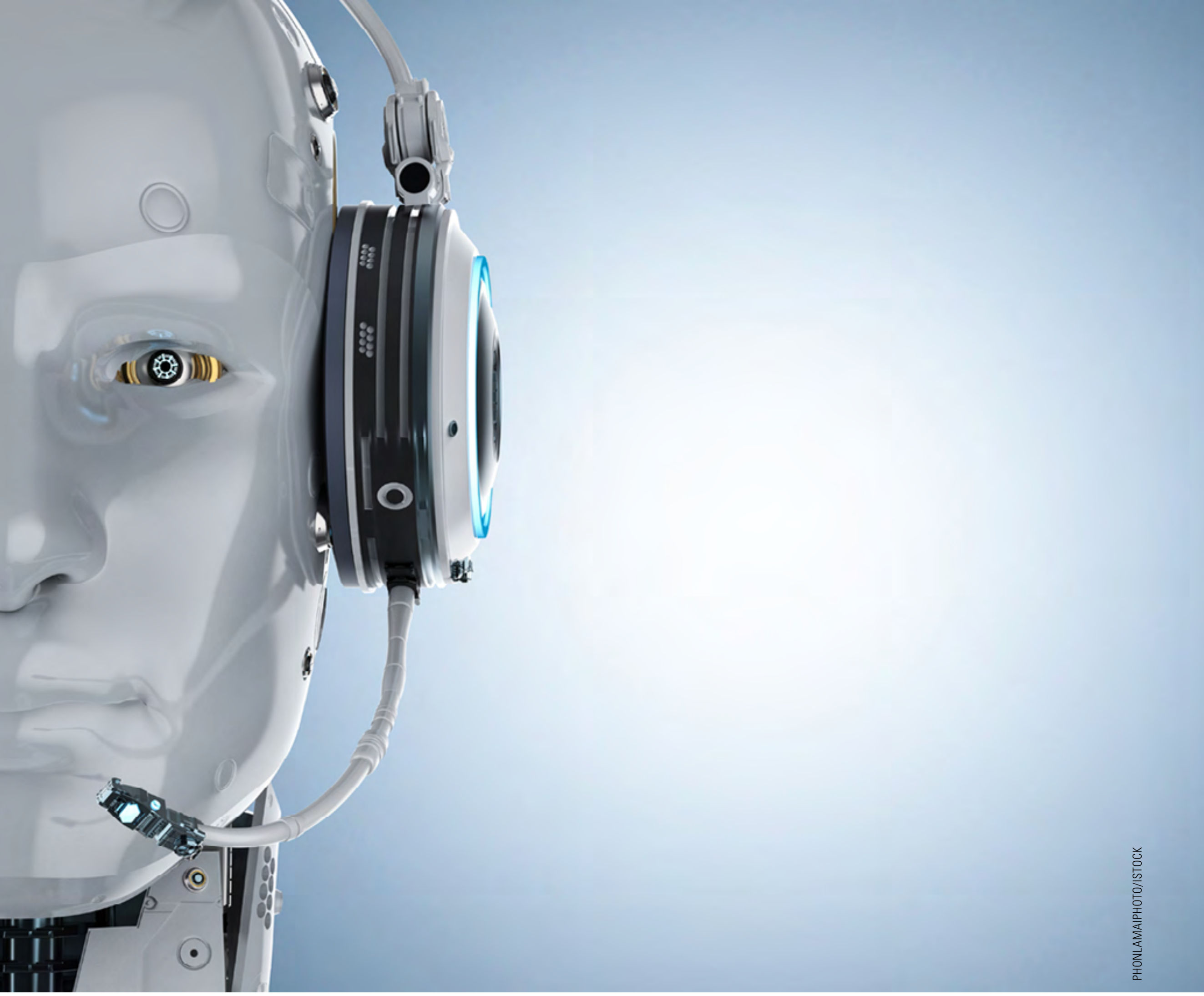
Según el estudio, publicado en 2020, el 82 por ciento de los ciudadanos alemanes temen una vigilancia generalizada por parte de la inteligencia artificial (IA). Un número similar (79 por ciento) cree que los sistemas técnicos tomarían decisiones poco éticas y tres de cada cuatro encuestados ven amenazada nuestra seguridad por el creciente poder de las máquinas que actúan de manera autónoma. El hecho de que el «Gran Hermano» observe nuestra sala de estar y nuestro dormitorio es para una mayoría una realidad que se siente hace tiempo. ¿Se acerca de manera inminente el dominio de los bits y los *bytes*?

Esta cuestión entraña una pregunta técnica: ¿es, en principio, la IA capaz de hacer algo así? En cierto modo, sí. Los algoritmos inteligentes ya se están utilizando, por ejemplo, como instrumentos de espionaje y manipulación, por parte de personas y organizaciones que persiguen diferentes objetivos. Sin embargo, ello no significa que la propia IA aspire a algo parecido a la dominación.

Los filósofos de la tecnología discuten esta cuestión bajo el término de «intencionalidad», es decir, la cualidad de llevar a cabo acciones deliberadas y orientadas a un objeto. Muchos autores consideran la intencionalidad como un componente permanente de la conciencia.

La intencionalidad se atribuye principalmente a estados mentales, como las percepciones, las creencias o los deseos. La idea subyacente es que siempre nos podemos dirigir a objetos escogidos, individuales, pero nunca al mundo como totalidad. Cuando pensamos o sentimos, utilizamos una suerte de «reflectores» que iluminan secciones limitadas del mundo; no todo a la vez, como hace el Sol. Lo que accede a nuestra conciencia (percepciones, pensamientos, sentimientos u otros contenidos de la experiencia) posee en cada caso propiedades experienciales subjetivas, también llamadas qualia (por ejemplo, la sensación de dolor o la percepción del color rojo).

Aunque el concepto procede de la escolástica medieval, el debate moderno sobre la intencionalidad se remonta al filósofo y psicólogo Franz Brentano (1838-1917). En su libro *Psicología desde el punto de vista empírico* (1874) sostuvo que la intencionalidad es la característica esencial de



PHONLAMAIPHOTO/ISTOCK

todos los actos de conocimiento: siempre están referidos al objeto, es decir, dirigidos a algo.

Para Brentano, una característica básica de lo mental es dirigirse a un objeto o referirse a él. Por ejemplo, si pienso «la manzana está sobre la mesa», eso se refiere a los objetos *manzana* y *mesa*, así como a la relación espacial que guardan entre sí. Con respecto a este estado de cosas, el pensamiento puede ser verdadero o falso. Brentano consideraba por ello la intencionalidad algo exclusivamente psíquico: «Ningún fenómeno físico muestra nada semejante».

Pensamientos como eventos puramente físicos

Por el contrario, los materialistas reducen los estados mentales —y, por tanto, intencionales— a los estados físicos. Desde su punto de vista, pensar en una mesa o en una manzana depende de ciertas circunstancias físicas de mi cerebro. Es una cuestión abierta si a estas también se les puede atribuir intencionalidad.

Los materialistas equiparan los pensamientos con eventos neuronales. Otros, en cambio, sostienen que un proceso en el cerebro es intencional solo si se pueden explicar el sentido, los motivos y la verdad también sin estados mentales (por ejemplo, aludiendo a que simplemente se dan en el lenguaje de las neuronas o de las máquinas). Desde este punto de vista, la IA (sobre todo la que se basa en estructuras neuronales) tendría intencionalidad.

Pero, ¿cómo podría confirmarse tal característica de la inteligencia de las máquinas? ¿Qué podría servir como prueba? Ya en 1950, el matemático Alan Turing (1912-1954) desarrolló el test que hoy lleva su nombre: el test de Turing. En él, una persona se sienta ante un ordenador y se comunica con dos interlocutores desconocidos para él. Uno es un ser humano; el otro, un ordenador. La persona hace preguntas que responden ambos, el humano y el ordenador. Si no puede diferenciar si las respuestas provienen de una persona o de un ordenador, este último ha superado el test. Según Turing, debemos entonces reconocer que la máquina tiene inteligencia en el mismo sentido que su compañero humano.

Google presentó una nueva interpretación del test de Turing en una conferencia de desarrolladores en 2018 con su sistema Duplex. Esta IA es capaz de realizar llamadas telefónicas con voz humana, por lo que el usuario apenas puede reconocer que no se trata de una persona. Sin embargo, esto tiene ciertos límites: en cuanto se quiere hablar con Duplex de otra cosa que no sean las reservas de citas que gestiona, la IA no funciona. Con todo, dentro del marco fijado, los humanos y las IA no son fácilmente distinguibles. Entonces, ¿ha superado Duplex el test de Turing?

Dudas sobre el test de Turing

Una crítica obvia a este procedimiento es que el test de Turing reposa sobre la premisa de que se pueden confirmar el pensamiento y la intencionalidad de un ordenador mediante la evaluación por parte de un humano. El filósofo John Searle ya lo puso en duda en su artículo «Mentes, cerebros y programas», de 1980.

Para ello planteó el experimento mental de la sala china. Supongamos que una persona se encuentra en una habitación llena de libros. Ahora se le pasa por debajo de la puerta una hoja de papel en la que hay escritos caracteres chinos. Los libros que se encuentran en la sala explican en la lengua materna de la persona qué caracteres de la hoja deben ser «respondidos» con qué otros símbolos. Sin conocer ni uno solo de los símbolos, es decir, sin entender nada en absoluto, la persona logra escribir en el papel respuestas con sentido. Alguien que domine el chino y que se encuentre fuera de la habitación tendría que asumir que la persona que está en el interior conoce el significado de los caracteres.

Con esta analogía, Searle señaló un hecho simple: los ordenadores siguen instrucciones sobre cómo deben ser manejados los signos y cómo reaccionar adecuadamente ante ellos según ciertas reglas. Así, un movimiento del ratón en una pantalla puede representar que se aplique la fórmula binomial o que se reconozcan las expresiones de un chat como un discurso de odio y se bloqueen. ¿Pero «sabe» el ordenador lo que está haciendo? No. No entiende nada de nada.

Los humanos consideramos que la IA es intencional porque nosotros mismos lo somos

Los sistemas de las máquinas no conocen las conexiones mecánicas, ni las comunicativas u otras. El significado de sus reacciones se deriva solo de la intencionalidad humana. En otras palabras: si un vehículo autónomo frena *para* no atropellar a un peatón, ese «para» procede de los seres humanos. Los motivos y consideraciones legales, sociales o morales solo los conocemos nosotros.

Para Searle, la intencionalidad no se basa únicamente en la comprensión del significado. Más bien, se vincula con determinados «actos de habla». En la teoría de los actos de habla de Searle, los actos comunicativos deben ser interpretados como acciones. El lenguaje humano, más allá de la gramática y el significado, tiene siempre un propósito intencional, ausente en los sistemas de signos de la máquina.

Junto con Paul Grice y otros teóricos, Searle desarrolló la tesis de que la intencionalidad se manifiesta solo en la capacidad para el uso dirigido del lenguaje. Este surge cuando una persona persigue una intención que la hace hablar. Por ejemplo, una madre y Alexa de Amazon pueden responder con las mismas palabras cuando un niño grita «¡Ay!»: «Oh, lo siento, ¿puedo ayudarte?» El contenido significado, sin embargo, difícilmente podría ser más diferente: la empatía y el cuidado, por un lado; una simple fórmula de cortesía programada, por otro.

Algunos apologistas de la IA argumentan que esto vale solo para la «IA débil», ya que solo puede ofrecer aquellos resultados que le han sido previamente suministrados. Tal es el caso de la IA existente hasta la fecha: ya sea para seleccionar un candidato, para diagnosticar un cáncer o para la conducción autónoma (todas las decisiones de las que tales sistemas son capaces se basan en rutinas computacionales que los humanos han diseñado).

Otra cosa es la «IA fuerte». Aquí, los humanos dan solo el impulso inicial. Toda la estructuración posterior y los patrones de reacción son generados por la propia IA. En consecuencia, dicen, también podría desarrollar una intencionalidad propia —sin embargo, esta posibilidad es hasta ahora solo una hipótesis. Todas las distopías según las que tales sistemas podrían someter o exterminar a la humanidad son pura ciencia ficción—.

Con todo, las predicciones más cautelosas sobre el modo en que la IA cambiará nuestro mundo son más que dignas de consideración. Un estudio del Instituto para el Futuro de la Humanidad, en Oxford, publicado en 2018, predecía que en el curso de la década de 2020 muchos trabajos serían realizados por máquinas: de la traducción de textos a la conducción de camiones. La IA podría incluso conquistar las listas de superventas con canciones propias o escribir noticias y *bestsellers*. Vista así, parece penetrar en áreas humanas originales.

¿Solo una cuestión de progreso técnico?

En vista de las posibilidades técnicas en constante crecimiento, se dice que la IA conquistará en breve el último bastión de lo humano: la intencionalidad. Pero la atribución de esta es el resultado de una necesidad humana primaria. Como seres sociales, dependemos tanto de no perder de vista las propias intenciones como de atribuírselas también a nuestros respectivos compañeros.

Una sociedad sin intencionalidad que le sirva de base no sería solo disfuncional, sino también carente de sentido: solo ella me permite comprender como «yo» a la persona que está en el espejo. Como los psicólogos del desarrollo mostraron en la década de 1970, los bebés reconocen después de pocos meses de vida que un punto rojo en la imagen que se refleja en el espejo procede de un punto igual que está en su frente. Refieren el punto a sí mismos.

Dado que por ahora no hay ningún otro método comparable para demostrar la intencionalidad, es muy difícil «consultarla» en la IA. La inteligencia, en cambio, es más fácil de medir. Un equipo de investigadores chinos desarrolló en 2016 un método que podía hacer comparables los cocientes intelectuales de sistemas artificiales y naturales. La IA del asistente de Google obtuvo así una puntuación de apenas 50 (más o menos el nivel de un niño pequeño). Sin embargo, esta atribución es, en el mejor de los casos, un indicio débil, y de ninguna manera suficiente, para acreditar la intencionalidad de una IA.

Posiblemente esto no es más que otra muestra del modo en que nuestro propio pensamiento influye en nuestra visión de la realidad. Que atribuyamos intencionalidad a la IA es un acto de antropomorfización y, por tanto, de humanización. Los humanos consideramos que la IA es intencional porque nosotros mismos lo somos.

Las dudas relativas a que las máquinas puedan desarrollar alguna vez intencionalidad se alimentan también de otra fuente: la lógica. Dado que calcular y pensar son procesos fundamentalmente diferentes, es de suponer que solo a partir del primero no se llega nunca al segundo. Las operaciones de cálculo no llegan más allá de las premisas establecidas para ellas; la mente humana, en cambio, gracias a la intencionalidad, tiene la capacidad de reflejarse a sí misma.

La intencionalidad constituye una facultad humana original. Crea la posibilidad de reconocer el sentido, al que solo nosotros tenemos acceso como seres pensantes y sociales. Por tanto, el dominio mundial de la IA debería seguir siendo, también en el futuro, una pesadilla hollywoodiense.

Dorothea Winter es filósofa. Actualmente está realizando un doctorado sobre intencionalidad e inteligencia artificial en la Universidad Humboldt de Berlín.

PARA SABER MÁS

Intelligence quotient and intelligence grade of artificial intelligence. Feng Liu, Yong Shi y Ying Liu en *Annals of Data Science*, vol. 4, núm. 2, págs. 179-191, mayo de 2017.

When will AI exceed human performance? Evidence from AI experts. Katja Grace et al. en *Journal of Artificial Intelligence Research*, vol. 62, julio de 2018.

Minds, brains, and programs. John R. Searle en *Behavioral and Brain Sciences*, vol. 3, núm. 3, págs. 417-424, septiembre de 1980.

EN NUESTRO ARCHIVO

Tertulia donde se discurre sobre el test de Turing y la posibilidad de crear máquinas pensantes. D. R. Hofstadter en *IyC*, julio de 1981.

¿Es la mente un programa informático? John R. Searle en *IyC*, marzo de 1990.

¿Máquinas pensantes? María Cerezo en *IyC*, septiembre de 2014.

El traje nuevo de la inteligencia artificial. Ramon López de Mántaras en *IyC*, julio de 2020.