Óscar Quejido Alonso, profesor del departamento de Filosofía y Sociedad de la Universidad Complutense, analiza desde un planteamiento filosófico nuestra relación con la IA, producto del «sueño occidental de replicar nuestra propia forma de racionalidad». Las respuestas a las preguntas sobre sus efectos en el medio y largo plazo las podemos encontrar en una rama de la filosofía contemporánea: la filosofía de la mente.

ALGUNAS CUESTIONES FILOSÓFICAS EN TORNO A LA INTELIGENCIA ARTIFICIAL

ÓSCAR QUEJIDO ALONSO

Desde que a mediados del siglo pasado los primeros estudios sobre computación hicieran factible la posibilidad de crear una inteligencia artificial, se han producido significativos avances tecnológicos en este sentido. En concreto, en los últimos años, el cambio en el planteamiento del aprendizaje de las máquinas ha permitido, finalmente, poner en juego una revolucionaria tecnología de computación, denominada Chatbot, exponencialmente muy superior a la que hasta ahora manejábamos, que busca replicar la capacidad multitarea de la mente humana, especialmente cuando estamos manteniendo un diálogo sobre prácticamente cualquier cosa.

Sin embargo, llama la atención comprobar cómo, más allá de los innegables avances técnicos y de los indudables beneficios que la IA representa en esferas fundamentales de nuestra vida cotidiana, en otro orden de cosas que no remiten a lo inmediato, seguimos lejos de encontrar respuestas satisfactorias a preguntas que tienen que ver con la comprensión de los efectos que, a medio y largo plazo, puede producir introducir esta nueva tecnología en nuestras vidas. La falta de respuestas a preguntas más fundamentales, más filosóficas, si se prefiere, tiene que ver, en primer lugar, con la falta de una reflexión crítica sobre el rumbo que deben tomar las sociedades contemporáneas. Algo que, por otra parte, no es tampoco exclusivo de nuestro presente; la irrupción de otras revoluciones anteriores, como la Revolución Industrial, a comienzos del siglo XIX, aunque cuestionada, fue impuesta por intereses que sobrepasaban el bienestar de la sociedad y en nombre de un progreso que, como sucede en la actualidad, no es cuestionado. De hecho, en nuestros días, como en otras ocasiones, esta falta de reflexión no conlleva en ningún caso la modificación de la agenda de las grandes empresas tecnológicas, que multiplican a diario sus inversiones en investigación, en un negocio con mucha rentabilidad para el que aún no contamos con una legislación específica. Ni siquiera cuando una suma de expertos, muchos de ellos vinculados directamente al desarrollo directo de estas tecnologías, firman una carta, como sucedió hace poco, pidiendo una ralentización en la comercialización y el uso de la misma, con el fin de comprobar, mínimamente, sus efectos.

En un nivel de cuestionamiento todavía más radical, muchas de las complicadas preguntas sobre nuestra relación con la IA tendrán que ver con el viejo sueño occidental de replicar nuestra propia forma de racionalidad, nuestra propia forma de inteligencia, a la que, por otra parte, se ha considerado como la más propia de la humanidad. Desde una perspectiva filosófica o crítica, este sueño implica dimensiones epistemológicas, ontológicas y éticas, las cuales arrastran, como digo, a su vez, tanto nuestra comprensión de lo que significa ser «inteligente» como la compleja relación entre una inteligencia que podría ser llamada «artificial», por contraposición a una que sea «natural». Es muy significativo que uno de los debates actuales gire en torno a la denominada singularidad, es decir, al momento en el que una máquina sea capaz de mejorarse a sí misma, sin la intervención de un programador humano, cuando en absoluto estamos interesados en entender las motivaciones que nos llevan a intentar replicar nuestra propia inteligencia, en la forma de una IA, con el fin de mejorarnos. ¿Por qué iba a querer una máquina mejorarse a sí misma? Parece que, una vez más, proyectamos en esta tecnología nuestros más íntimos anhelos.

En último término, en la relación con la IA aparecen, en muchas ocasiones de manera no evidente, nuestras valoraciones filosófico-antropológicas a la pregunta ¿qué es lo más propiamente humano? O, dicho de manera más clara, ¿en qué consiste eso de ser «racional»?

Las respuestas a estas preguntas remiten, en la cultura occidental, a una suerte de humanismo que, desde la Antigua Grecia, nos describe a nosotros mismos como el animal racional, diferenciándonos, con ello, por medio de la razón de todos los demás animales que, supuestamente, carecerían de esta cualidad. Este tema de la diferenciación cualitativa de lo humano se habría extendido ahora también a las máquinas, aunque la solución pasa por responder antes a complicadas cuestiones como la de entender el funcionamiento del conocimiento, tanto teórico como práctico, así como a la posibilidad de dar una buena descripción de la mente y de su funcionamiento, ese lugar en el que tradicionalmente reside nuestra razón.

Muchas de las cuestiones que estamos planteando y que, como digo, implican al núcleo de nuestra relación con la IA—cuestiones

La filosofía de la mente nace, coincidiendo con la emergencia de la IA, en torno a los años cincuenta del pasado siglo, con el objetivo de ofrecer nuevas respuestas al que muchos consideran uno de los problemas filosóficos por excelencia: el problema de las relaciones de la mente con el cuerpo o con el cerebro.

Autómata, Museu Marès, Barcelona

que deberíamos si no resolver, sí al menos tener claras, en la medida de lo posible, por sus implicaciones– han sido abordadas por una rama de la filosofía contemporánea, denominada filosofía de la mente.

La filosofía de la mente nace, como campo de reflexión contemporáneo, coincidiendo con la emergencia de la IA, en torno a los años cincuenta del pasado siglo, aunque separada de esta, y con el objetivo de ofrecer nuevas respuestas al que muchos consideran uno de los problemas filosóficos por excelencia: el problema de las relaciones de la mente con el cuerpo o con el cerebro. Un viejo problema filosófico que, en la Modernidad, trató de resolver el filósofo francés René Descartes, afirmando, al menos desde una perspectiva teórica, que mente y cuerpo son dos realidades totalmente diferentes y diferenciadas. En el planteamiento cartesiano, la *res cogitans* (sustancia pensante) se identifica con la mente, y esta con la conciencia, en tanto que centro de la racionalidad; en este sentido, la actividad de la mente, el pensamiento, es caracterizado como lo contrario de la *res extensa* (sustancia extensa), es decir, de la materia y su comportamiento.

En el primer cuarto del siglo XVII estaban de moda en Europa[1] los primeros autómatas, mecanismos que por medio de engranajes y ruedas dentadas permitían el movimiento, aparentemente autónomo, de figuras articuladas con forma humana o animal, por ejemplo, en los relojes de cuco. El mecanicismo cartesiano que vendría a sustituir en física al todavía vigente hilemorfismo aristotélico, supuso una revolución profunda que implicaba la redefinición tanto de la noción de «lo natural» como aquello que lleva el principio del movimiento en sí mismo, frente al de «lo artificial», que implicaba una acción externa en su transformación. Los autómatas vendrían a desdibujar esta caracterización tradicional entre lo natural y lo artificial, procedente de Aristóteles.

Descartes consideraba que el pensamiento era una propiedad exclusiva de los humanos y que tanto los animales como estos nuevos autómatas estaban constituidos exclusivamente por materia, por *res extensa*. Esto no les impediría, a su juicio, moverse o incluso hablar, pero nunca realizarían estas tareas como lo haría cualquier humano porque, precisamente, carecerían de inteligencia o razón. En la parte sexta del *Discurso del método*, Descartes escribe:

> Si bien se puede concebir que una máquina [o un animal] esté de tal modo hecha que profiera palabras, y hasta que las profiera a propósito de acciones corporales que causen alguna alteración en sus órganos, como, v. g., si se la toca en una parte, que pregunte lo que se quiere decirle, y si en otra, que grite que se le hace daño, y otras cosas por el mismo estilo, sin embargo, no se concibe que ordene en varios modos las palabras para contestar al sentido de todo lo que en su presencia se diga, como pueden hacerlo aun los más estúpidos de entre los hombres.

Sorprende ver la anticipación con la que Descartes detecta, y trata de resolver por medio de su caracterización de la *res cogitans*, el problema de la delimitación de las fronteras entre lo animal, lo artificial y lo propiamente humano. En esta misma sección señala, en la misma línea:

1 S. Turró, *Descartes. Del hermetismo a la nueva ciencia*, Barcelona, Anthropos, 1985.

> Aun cuando [*los autómatas*] hicieran varias cosas tan bien y acaso mejor que ninguno de nosotros, no dejarían de fallar en otras, por donde se descubriría que no obran por conocimiento, sino solo por la disposición de sus órganos, pues mientras que la razón es un instrumento universal, que puede servir en todas las coyunturas, esos órganos, en cambio, necesitan una particular disposición para cada acción particular; por donde sucede que es moralmente imposible que haya tantas y tan varias disposiciones en una máquina que puedan hacerla obrar en todas las ocurrencias de la vida de la manera como la razón nos hace obrar a nosotros.

La flexibilidad de la mente humana, su inteligencia, reside en su capacidad para *comprender* el sentido de los términos empleados al hablar o de las acciones en un determinado contexto. Este será, en opinión de Descartes, el elemento propio de la razón humana que permite diferenciarnos de máquinas y animales. Es cierto que este planteamiento, sustentado a su vez en su radical dualismo de sustancias, por muy perspicaz que fuera, no deja de generar algunas objeciones importantes. El tema, sin ir más lejos, de la unión y de la conversión entre lo mental y lo corporal en el seno de este dualismo de sustancias obligó a Descartes —muy consciente de estos problemas— a proponer la famosa glándula pineal, como una suerte de «traductor» entre la esfera de lo mental (inmaterial y temporal) y la de lo material, caracterizada por su espacialidad. El propio Descartes reconocía no haber visto nunca un cuerpo sin alma ni un alma sin cuerpo.

Si bien es cierto que, ya en el siglo xviii, David Hume cuestionaría en profundidad el modelo de la mente propuesto por Descartes, las objeciones más sistemáticas y mejor planteadas no llegarían hasta el siglo xx de la mano del conductismo de Gilbert Ryle, extendiéndose posteriormente por las diferentes propuestas fisicalistas, hasta llegar a los primeros planteamientos funcionalistas que, desde los años setenta, equipararían, de una manera aún muy simplificada bajo la metáfora computacional, el funcionamiento de la mente con el de las computadores, dentro del paradigma del procesamiento de la información, por el que, tanto la mente natural como la artificial, funcionarían procesando representaciones simbólicas.

La historia de las diferentes propuestas a la hora de entender tanto qué es la mente como su funcionamiento son muy variadas, desde la configuración contemporánea de la filosofía de la mente en los últimos 75 años. Sin duda, merece la pena repasar detenidamente sus principales hitos para comprender algo mejor las dificultades y las reticencias que nos puede provocar un desarrollo incontrolado de la IA, y algunos de ellos son fundamentales para entender la perspectiva filosófico-antropológica que estamos dibujando.

En la década de los cincuenta, en el momento en el que Gilbert Ryle publica *The concept of mind* (1949), un texto considerado por muchos como fundacional para la filosofía de la mente contemporánea, uno de los puntos compartidos entre los diferentes planteamientos de la época tenía que ver con el desplazamiento respecto a las condiciones de verificabilidad de las afirmaciones hechas en referencia a la mente. A partir de este momento, cualquier afirmación con sentido sobre la mente tenía que poder ser verificada por un observador externo. De esta manera, la introspección, la forma de acceso a lo mental implícita en el modelo cartesiano, quedaba devaluada en favor de la perspectiva de la tercera persona, a partir de criterios objetivos de verificación. Estos mismos criterios se encontraban presentes en la propuesta del lógico Alan Turing, cuando en 1950 se preguntaba en un famoso artículo, «Computing machinery and intelligence», si las máquinas podían pensar, dando una respuesta afirmativa a la pregunta, basándose en la idea de que un observador externo, presente en una conversación, no pudiese diferenciar entre las respuestas dadas por otro humano y las dadas por una máquina. Criterios similares a este famoso «test de Turing» son los que se esgrimen en la actualidad cuando, por ejemplo, se cuestiona la efectividad de los Chatbots cuando estos dan, en ciertas ocasiones, respuestas totalmente erróneas, inverosímiles o inadecuadas a preguntas que, en principio, cualquiera podría responder con facilidad.

¿Por qué iba a querer una máquina mejorarse a sí misma? Parece que, una vez más, proyectamos en esta tecnología nuestros más íntimos anhelos.

Las cosas no resultan nada fáciles tampoco cuando tratamos de dar cuenta de las conductas inteligentes a partir de la conciencia. La filosofía de la mente habría detectado un campo problemático en el que las experiencias subjetivas, eminentemente cualitativas, podrían ser un elemento fundamental en los procesos de conocimiento y de toma de decisiones racionales. Sin duda, hoy es posible localizar las partes de nuestro cerebro que se activan ante la experiencia visual del color rojo, por ejemplo, pero como se ha señalado en muchas ocasiones, otra cosa muy distinta es dar cuenta de la *experiencia subjetiva* de la rojez que cada individuo vive y experimenta de una manera intransferible e incomunicable en términos objetivos y cuantificables. En la actualidad resulta imposible que las neurociencias expliquen esta experiencia, esta forma de *sentirnos nosotros mismos* por medio de nuestro estar experimentando el mundo; algo que, por otra parte, no está claro qué función podría cumplir en términos evolutivos.

Si la conciencia, entendida como esta forma de experiencia cualitativa y subjetiva del mundo, aporta algo fundamental para la inteligencia propiamente humana es algo que está por ver. Del mismo modo que está por ver si seremos capaces de replicar artificialmente algo que hoy en día desconocemos. Los más optimistas señalarán que, por muy complicado que pueda ser en su funcionamiento, la conciencia reside en el cerebro, por lo que, tarde o temprano las neurociencias conseguirán comprender sus mecanismos más profundos, siendo, desde ese día, reproducible como tantas otras funciones. Si, por otra parte, la mente, en tanto que conciencia, pertenece a la experiencia propia de cada individuo, entonces, un fondo de indeterminación permanecerá siempre ajeno a la ciencia. Al menos a la ciencia como la entendemos hoy.

Sin duda, nos encontramos en un momento crucial en lo que respecta a la implementación de nuevas tecnologías vinculadas a los desarrollos de la IA. Más allá de los problemas éticos, cuando esta reproduce los sesgos de género, capacitistas o raciales, o más allá también de las cuestiones ambientales que hacen prácticamente insostenible los consumos de energía necesarios para el funcionamiento de los procesadores por medio de los cuales los chatbots *aprenden*, hay un territorio especialmente sensible: aquel donde estas tecnologías se combinan con los avances biotecnológicos buscando la mejora de la condición humana. Las propuestas filosóficas transhumanitas vendrían a complicar la falta actual de reflexión crítica, filosófica, cuando defienden la implementación de la IA como una forma de perfeccionamiento de la inteligencia, sin asumir antes un cuestionamiento de lo que significa «mejorar». El humanismo tradicional que señalábamos más arriba como hilo conductor de una caracterización adecuada de lo más propiamente humano, se ve *acríticamente* reforzado por un ámbito de acción muy marcado ideológica y económicamente, impidiendo, una vez más, aquello que ya Nietzsche señalaba como el fin de la humanidad: hacernos dueños de nuestra propia historia.