¿PUEDEN SER PERSONAS LOS SISTEMAS INTELIGENTES?

Adela Cortina

1. La conferencia de Dartmouth.

Como es sabido, el sintagma *inteligencia* artificial nació el siglo pasado, concretamente, en la célebre conferencia de Dartmouth de 1955 sobre máquinas pensantes, impulsada por John McCarthy, quien introdujo la expresión "inteligencia artificial" en 1956, refiriéndose con ella a la creación de máquinas que pueden tenerse por inteligentes porque interactúan con los seres humanos hasta el punto de que una persona ya no sabe si está hablando con una máquina o con otra persona humana; una prueba que recibe el nombre de "test de Turing". Al proyecto se unen Minsky, Rochester y Shannon, y se entiende que la IA puede llegar a constituir un nuevo tipo de inteligencia.

En la Declaración Fundacional de la "Propuesta para un proyecto de investigación en verano sobre inteligencia artificial en Dartmouth" (1955) se dice expresamente que "el estudio procede sobre la base de la conjetura de que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, ser descrito con tanta precisión que puede fabricarse una máquina para simularlo. Se intentará averiguar cómo fabricar máquinas que utilicen el lenguaje, formen abstracciones y conceptos, resuelvan las clases de problemas ahora reservados para los seres humanos y mejoren por sí mismas"[1]. Como se echa de ver, se trata de construir máquinas que se comporten de un modo que se llamarían inteligentes si un ser humano se comportara de esa forma[2]. A mi juicio, conviene subrayar desde el comienzo que se trata de simulaciones, de que actúen "como si" fueran inteligentes realmente. La filosofía del "como si" planea

[1] J. McCarthy et alii, A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955", AI magazine 27 (4), 12.

[2] Es un aspecto que también subraya, entre otros, Luciano Floridi en *Ética de la inteligencia artificial. Principios, retos y oportunidades*, Barcelona, Herder, 2024, 63.

desde el comienzo y, sin embargo, se utiliza continuamente un lenguaje antropomórfico que no hace sino confundir los perfiles. Sorprendentemente, se habla de Ingeniería Ontológica como un campo de estudio, una materia que no trata del ser o el ente en tanto que ente, sino de un campo de las ciencias de la computación y ciencias de la información. Puede ser apasionante, pero el uso del rótulo "Ingeniería Ontológica" siembra confusión.

En cualquier caso, conviene recordar que es preciso hablar de "inteligencias", y no sólo de "inteligencia", tanto en el caso de los humanos, en el de los animales, en el de los vegetales, e incluso en el caso de la tierra si atendemos a la célebre hipótesis Gaia.

Ciertamente, desde su nacimiento la IA ha suscitado una gran cantidad de preguntas, pero en esta intervención voy a limitarme a una de ellas, que tiene un gran interés filosófico, tanto desde el punto de vista de la filosofía teórica como desde la perspectiva de la filosofía práctica: ¿los sistemas inteligentes (algoritmos, robots, máquinas) son sólo instrumentos que los seres humanos utilizamos y podremos utilizar en el futuro para los fines que nos propongamos, o podrán sustituir a los seres humanos porque irán cobrando paulatinamente el mismo estatus que ellos, incluso los superarán en cualidades morales, y no sólo en capacidad computacional?

Ésta no es una pregunta de ciencia ficción. Por el momento a un amplio conjunto de robots se les atribuye ya el calificativo de "autónomos", un adjetivo que, si pudiera ser aplicado en serio, exigiría que se les considerara como personas, con todas las consecuencias que esa caracterización comporta. El universo de las personas se ampliaría entonces más allá del mundo del *homo sapiens*, incluso más allá de los seres biológicamente constituidos, como han sugerido desde hace tiempo un buen número de autores. Las propuestas de transhumanistas y posthumanistas caminan en este sentido, pero no sólo las suyas, sino también

algunas intervenciones que forman ya parte de nuestra realidad social.

2. El fallido intento de reconocer personas electrónicas.

En efecto, en mayo de 2016 el Comité de Asuntos Legales del Parlamento Europeo lanzó la extraña propuesta de crear *personas electrónicas* en el marco de la legislación civil para regular la robótica. Se trataba con ella –se decía- de dar un estatus legal a los robots autónomos de vanguardia, para que sea posible reparar los daños que puedan causar cuando tomen decisiones autónomas o independientes de terceros. ¿Es lo mismo una decisión autónoma que una decisión que se toma con independencia de terceros?

Como es sabido, la Unión Europea se ha caracterizado por su preocupación normativa en relación con la IA, de la que dejaron constancia en 2019 las *Ethical Guidelines for Trustworthy AI*, a las que siguieron un buen número de documentos hasta culminar por el momento en la ley de 24 de marzo de 2024. Hasta el punto de que se ha denominado "el efecto Bruselas" a este afán de introducir normas para el diseño, producción y uso de la IA y de intentar que este afán se universalice[3]. Como apunta Anu Bradford, en el marco de los modelos de política digital, Estados Unidos se basaría en un mercado abierto; China, en un centralismo estatal; y la Unión Europea, en la regulación[4]. Y es verdad que ante las innovaciones la Unión Europea siempre responde teniendo en el trasfondo la apuesta por el Principio de Precaución, por fomentar la confianza, aun perdiendo por ello competitividad. Un asunto totalmente relevante, en el que ahora no podemos entrar.

Este afán normativo de la Unión Europea ha merecido la repulsa más expresiva de Donald Trump, quien insiste en mantener la ventaja de su país en IA y, por lo tanto, en

[3] Anu Bradford, "The Brussels Efect", Northwestern University Law Review, vol, 107, nº 1, 2012.

[4] Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology*, Oxford University Press, Oxford, 2023.

eludir una "regulación excesiva del sector", de modo que evitará participar en los planes de regulación y gobernanza que propone la UE.

Pero, regresando a la propuesta del Comité de Asuntos Legales del Parlamento Europeo de crear *personas electrónicas*, afortunadamente, no prosperó, entre otras cosas, porque un colectivo de más de 200 expertos en inteligencia artificial y robótica, líderes industriales y expertos en ética, medicina y derecho la criticó duramente, con razones que, a mi juicio, son discutibles. De criticarlas me he ocupado en otro lugar[5] y en un artículo, publicado en "El País" con el título "Personas electrónicas", que nuestro compañero Diego Gracia tuvo la deferencia de mencionar en su penúltima intervención en esta Academia. En esta ocasión las recordaré brevemente para entrar en el punto que, a mi juicio, es central para lo que aquí nos ocupa.

En principio, entiende el grupo de expertos que la *inteligencia artificial y los robots han de estar al servicio de la humanidad, por encima de cualquier otra consideración*. Los trata, pues, como instrumentos.

Esta afirmación no hace sino atenerse a las propuestas de la UE, que son humanocéntricas desde las primeras *Guidelines*. Precisamente este humanocentrismo calmó las inquietudes de la frankenfobia, el miedo a generar monstruos que se había extendido entre la ciudadanía. Más adelante se añadió que la IA debe estar también al servicio de la sostenibilidad de la naturaleza. Pero si algunos robots fueran realmente autónomos, no sería así, también ellos tendrían derecho a que todos los seres no autónomos estuvieran a su servicio.

En segundo lugar, el grupo de expertos critica que se *esté atribuyendo a la inteligencia artificial y a los robots unas capacidades situadas muy por encima de lo que se puede*

[5] Adela Cortina, *¿Ética o ideología de la inteligencia artificial? El eclipse de la razón comunicativa en una sociedad tecnologizada*, Paidós, Barcelona, 2024, cap. 7.

reconocer hoy, una atribución distorsionada por la ciencia ficción y por la prensa sensacionalista.

Evidentemente así es y, aunque los expertos no lo mencionen, ese sensacionalismo no es casual, sino motivado por el afán de atraer la atención en un mundo marcado no sólo por la extracción de datos a los usuarios de las plataformas, sino también por la Economía de la Atención. Atraer la atención para que los usuarios queden cautivos de la red es el objetivo. Como mencioné en el libro citado, se trata entonces de una ideología.

Sin embargo, la razón que se aduce en último lugar para negar que los robots puedan considerarse personas no puede ser más discutible, *en tanto que razón*. Según los expertos, no se puede tomar como modelo la persona natural, porque entonces tendrían derechos humanos, como el derecho a la dignidad y a la integridad, el derecho a la remuneración o a la ciudadanía. Todo esto estaría en contradicción con la Declaración de Derechos humanos de 1948, con la Carta de Derechos Fundamentales de la Unión Europea y con la Convención para la Protección de los Derechos Humanos y las Libertades Fundamentales, que sólo se refieren a seres humanos.

En efecto, si reconociéramos como personas a los robots llamados "autónomos", deberíamos reconocerles también estos derechos. Pero, a mi modo de ver, ésa no es una objeción para reconocerles como personas, sino la consecuencia ineludible que habríamos de aceptar si las reconociéramos como tales, y que llevaría a ampliar la Carta de Derechos a todas las personas, sean o no humanas, o bien a elaborar Cartas sobre los Derechos de las personas no humanas (animales, plantas, sistemas inteligentes).

De hecho, aunque no en este contexto, autores como Martha Nussbaum defienden la existencia de "personas en sentido amplio" y "personas en sentido estricto". Las personas en sentido amplio serían todas aquellas que son fines en sí mismas, porque gozan de unas capacidades propias,

que deben ser empoderadas, como es el caso de los animales con capacidad de sentir[6]. Por su parte, David DeGrazia afirma que hay animales no humanos, que son "personas limítrofes" entre lo que son personas de forma paradigmática y lo que claramente no son personas. Con ello se refiere a los grandes simios y a los delfines.

Pero yendo más allá del mundo biológico, autores peculiares como Mo Gawdat, ex director ejecutivo de Google, proponen extender la Declaración de Derechos Humanos y elaborar una Declaración de todos los seres inteligentes y autónomos, lo que él llama de modo impreciso una Declaración Universal de los Derechos Globales. No hacerlo –afirma- sería discriminatorio, incurrir en especismo, al atribuir un estatus superior a la especie humana y a los seres vivos frente a las IAs cuyo sustrato no es biológico, sino de silicio[7]. La discriminación no daría lugar sólo a "especismo", sino a "biologicismo". Y es verdad que existen ya publicaciones sobre derechos de los robots.

Como hemos mencionado, el texto fue criticado ampliamente, pero suscitó interesantes debates sobre la cuestión "¿pueden ser responsables las máquinas?", análoga a la que planteó Alan Turing "¿puede pensar una máquina?". La respuesta generalizada es que no pueden serlo. Los robots no tienen propiedad, ni, por tanto, libertad y responsabilidad, de ahí que los expertos críticos aconsejen que los robots cuenten con seguros contra los daños que puedan causar y responder de ellos. De donde se infiere que la propuesta de la comisión de convertir a los robots en persona electrónicas sólo sirve para confundir, porque no se pedirá responsabilidades a los robots, ya que no la tienen, sino a los dueños de los robots, que son personas con responsabilidades. Pero entonces la pregunta clave es, a mi juicio, la siguiente: ¿pueden ser realmente autónomos los robots, al menos los que actúan independientemente de

[6] Adela Cortina, *Las fronteras de la persona. El valor de los animales, la dignidad de los humanos*, Taurus, Madrid, 2009, p. 131.
[7] Mo Gawdat, *La inteligencia que asusta*, Paidós, Barcelona, 2024, 290-291; Cortina, 2024, 290 y 291.

terceros y autoaprenden? Ése es el nombre que se les da, ¿pero es el nombre acertado o no hace sino generar confusión?

El mundo de los robots es inmenso y las variedades, incontables. Los más conocidos son los robots domésticos, como Roomba, Pepper o Alexa, los robots cuidadores en residencias de ancianos. Y los que están más cerca –a mi juicio- de merecer el calificativo de "autónomos", que son los robots militares, los vehículos autónomos y los robots sexuales, que van a contar en esta intervención con un apartado dedicado expresamente a ellos.

La pregunta será, pues: ¿hay robots que pueden recibir el calificativo de autónomos, o es un antropomorfismo más de los muchos que pueblan el mundo de la IA? Un antropomorfismo que tiene la perversa consecuencia de desviar hacia otro lugar la responsabilidad de los verdaderos autores, que son personas humanas. Una responsabilidad que puede descubrirse siguiendo la trazabilidad de los actos de los robots presuntamente autónomos.

Ocurre, sin embargo, que la dignidad, por la que atribuimos un estatus especial a los seres a los que se la reconocemos no es un predicado *natural*, comprobable empíricamente, sino un predicado *biocultural* que no pude verificarse empíricamente, sino que tiene un fundamento metafísico, enraizado en una base biológica.

3. "Autonomía" es un predicado biocultural.

El asunto es de la mayor relevancia, porque la Declaración Universal de Derechos Humanos de 1948, aunque no sea un texto filosófico, hace descansar la exigencia de respetar esos derechos en la dignidad de los seres que pertenecen a la familia humana. Precisamente, porque esos derechos no son empíricamente comprobables, decía Bentham que los derechos humanos son disparates con zancos y MacIntyre, más suavemente, los considera "ficciones útiles". Es, decir, que no pueden verificarse empíricamente, pero

ayudan a avanzar en humanidad. En *Tras la virtud* dirá abiertamente: "No existen tales derechos (naturales o del hombre) y creer en ellos es como creer en brujas y unicornios"[8]. Y es verdad que no existen los derechos naturales como entidades empíricamente comprobables, ni tampoco la dignidad como cualidad verificable. Pero sí que la reconocemos a los seres que gozan de autonomía, porque son capaces de proponerse fines por sí mismos y de actuar por leyes que no atienden al interés egoísta, sino que las universalizaríamos.

En este punto la propuesta kantiana de considerar la autonomía como un predicado que forma parte de nuestro bagaje biocultural es difícilmente rebatible, al ligar autonomía a moralidad.

"En el Reino de los Fines –dirá en el bien conocido texto de *La Fundamentación de la Metafísica de las Costumbres*- todo tiene un *precio* o una *dignidad*. Aquello que tiene un precio puede ser sustituido por algo equivalente; en cambio lo que se halla por encima de todo precio y, por tanto, no admite equivalente, eso tiene dignidad"[9]

Y continúa algo más adelante:

"La moralidad es la condición bajo la cual un ser racional puede ser fin en sí mismo; porque sólo por ella es posible ser miembro legislador de un reino de los fines. Así pues, la moralidad y la humanidad, en cuanto que ésta es capaz de moralidad, es lo único que posee dignidad"[10]

Digno es, pues, el ser que no tiene equivalente en el mundo del intercambio, en el que todo es mercancía, y por eso no se le debe intercambiar por un precio. En los siglos en que está gestándose el capitalismo, cuando todo se convierte en mercancía, descubrimos que el núcleo de la vida

---

[8] Alasdair MacIntyre, *Tras la Virtud*, Crítica, Barcelona, 1981, 95.

[9] I, Kant, *Fundamentación de la metafísica de las costumbres*, Espasa-Calpe. Madrid, 1946, 92.

[10] Ibid, 93.

social es justamente algo que no debe intercambiarse por un equivalente, porque no tiene precio, sino dignidad.

La pregunta es ineludible: ¿puede atribuirse autonomía a los robots llamados "autónomos", de modo que debamos reconocerles dignidad y derechos, considerarles como personas morales, y no sólo legales, con todos los derechos que en ese caso les corresponderían? Para hacerlo no deberíamos esperar a que se cumpla la expectativa de la aparición de una superinteligencia, porque los robots supuestamente autónomos existen y funcionan ya. Un caso bien conocido sería el de los vehículos autónomos y las armas autónomas, pero más interesante todavía para esta intervención es el caso de los robots sexuales.

4. Matrimonios entre seres humanos y robots.

En el año 2017 se celebró el primer matrimonio entre un ser humano y un robot. Un ingeniero chino, llamado Zheng Jiajia, experto en IA, había pasado años buscando, sin éxito, una mujer que colmara sus aspiraciones amorosas, y decidió por fin crear para sí mismo una pareja robótica valiéndose de su pericia en ese ámbito. Se casó con "ella" en una ceremonia sencilla, teniendo por testigos a su madre y algunos amigos. Al parecer, el robot ginoide de Jiajia es una muñeca de dimensiones humanas, con una habilidad limitada para reconocer caracteres chinos y para hablar unas frases básicas, pero Jiajia proyectaba perfeccionarla en un futuro próximo[11].

Leyendo el artículo de John Danaher sobre "Sexuality" en que relata esta historia viene a la mente el comienzo de la excelente novela de Ian McEwan, *Máquinas como yo.* El protagonista, Charlie, compra por 86.000 libras uno de los primeros humanos manufacturados, verdaderamente viables, de un lote de 25 individuos, trece Evas y doce Adanes. Las Evas se acaban en seguida y debe contentarse con un Adán. Pero lo

[11] John Danaher, "Sexuality", en Markus D. Dubber, Frank Pasquale and Das Sumit (Comps.), *The Oxford Hanbook of Ethics of IA*, Oxford University Press, 2020, 403-420.

interesante es que, a diferencia de los seres humanos, debe programarlo, y Charlie recibe el robot con un manual de instrucciones de 470 páginas. A diferencia de los hijos biológicos, que vienen con la dote de la lotería natural, e incluso la social –por decirlo con Rawls-, el robot ha de ser programado, aunque más adelante irá evolucionando, gracias al aprendizaje de máquinas[12].

Regresando a la creación de Jiajia, es preciso subrayar que ésta no es una cuestión de ciencia-ficción ni tampoco una leyenda, como tantas otras que envuelven a la IA, sino un eco de sociedad, relatado en la prensa. Jiajia fabrica su propia pareja en la vida real y se propone perfeccionarla en el futuro, prolongando el ancestral sueño de Pigmalión, hecho realidad efectiva en el siglo XXI y no relatado sólo como una fábula. Como recordamos, Pigmalión, famoso como escultor, no encontró la mujer perfecta que buscaba y decidió esculpir estatuas de mujeres hermosas, pero un buen día se enamoró de una de ellas, Galatea, y Afrodita le concedió el don de que se convirtiera en una mujer de carne y hueso. En el caso de Jiajia, que yo sepa, su esposa no se ha convertido en mujer humana.

Algo más tarde Akihiko Kondo, un japonés de 35 años, se casó con Hatsune Miku, una cantante holográfica virtual. Había sido desgraciado en amor humano, buscó una pareja artificial y quería ser reconocido como miembro de una minoría de personas que no están interesadas en amantes humanos. Kondo se caracteriza como fictosexual, es decir, que siente una atracción amorosa o sexual+ dirigida a personajes de ficción. Hoy, según Danaher, ya hay varios miembros de la comunidad *on line* de "iDollators", que prefieren el amor a las muñecas al amor a seres humanos.

En el mundo de los robots, que abarca una amplia gama de actividades, los robots sexuales forman un capítulo ineludible, que recoge posiciones diversas a favor y en contra de su fabricación y uso. A favor se sitúan quienes consideran, por ejemplo, que, al contentarse con los robots,

[12] Ian McEwan, *Máquinas como yo*, Anagrama, Barcelona, 2019.

sería posible acabar con la prostitución y explotación de seres humanos, como también, reducir la transmisión de enfermedades venéreas; o que las personas tímidas en estas lides podrían perder su complejo y entablar relaciones fluidas con muñecas robóticas. En contra de estas relaciones se encuentran quienes entienden que estas prácticas son degradantes y además acostumbran a romper vínculos con los demás seres humanos. Sin embargo, no es en este tipo de consideraciones en el que deseo entrar en esta intervención, sino en una reflexión, a mi juicio, verdaderamente radical sobre la naturaleza de la IA en su conjunto y que consiste en tratar de responder a la tercera de las preguntas que formula Danaher en el artículo mencionado. A partir de la anécdota, pasaremos a la categoría.

Danaher se propone responder a tres preguntas: 1) Si hemos de añadir una "sexualidad digital" a quienes expresan su sexualidad con parejas digitales. Por su parte, considera que no debe hacerse porque de este modo se estigmatizaría a quienes fueran señalados con esa identidad. 2) Si este apoyo digital es algo reprobable. Cree que no, aunque hay que prevenir riesgos, por ejemplo, la pérdida de privacidad. 3) Si una forma sofisticada de IA puede ser objeto de amor. Y aquí se presenta –a mi juicio- la pregunta esencial: ¿podemos mantener una relación amorosa con algo que ha sido programado para amarnos, teniendo en cuenta que la elección amorosa supone una elección libre? Danaher considera que sí. Pero no cabe duda de que la cuestión es bien compleja, y, a mi juicio, nos conduce al título de esta intervención: ¿pueden *ser* personas los llamados sistemas inteligentes, en este caso, los robots, y no sólo comportarse *como si* lo fueran? ¿Pueden ser autónomos *realmente* y no sólo *funcionalmente*?

Sólo si es posible responder afirmativamente a estas dos preguntas, pueden ser los robots responsables de sus acciones, pueden ser personas electrónicas. Llamar "personas electrónicas" a entidades a las que se otorga una personalidad jurídica para que pueden responder de los daños causados por los robots es realmente desacertado y no es extraño que haya causado un enorme rechazo. Si el derecho

quiere ayudar a resolver problemas y no crear otros nuevos más vale que empiece por manejar los términos adecuadamente. Las cuestiones de palabras son solemnes cuestiones de cosas.

Sigue, pues, en pie la pregunta: ¿pueden ser personas los sistemas inteligentes si tenemos que reconocerles autonomía en ocasiones?

5. ¿Puede haber autonomía sin conciencia?

Los intentos de hominizar a seres no humanos viene de lejos y se refuerzan en el siglo pasado con el debate animalista. Parece que lecciones noveladas como *La Isla del Doctor Moreau* de G.H. Wells deberían zanjar esas disputas, porque la obra plantea de forma palmaria las dificultades de intentar transformar de forma artificial animales no humanos en humanos, tomando como una de las claves esenciales la de inculcarles una mente moral[13] A mi juicio, algo semejante se intenta hoy de nuevo, pero dando un paso más allá, al tratar de hominizar las máquinas inteligentes programando en ellas una suerte de mente moral. Pasaríamos entonces del sustrato biológico, propio de animales y seres humanos, a un sustrato de silicio.

Moreau se propone transformar animales no humanos en humanos y para ello cuenta con dos claves: la vivisección, practicada a los animales, que permite transformarlos fisiológicamente en "humanimales", y la mentalización, la conversión de las mentes animales en mentes humanas, tratando de inculcarles una "Ley de la Humanidad". Un "recitador de la ley" recuerda diariamente a los humanimales unas leyes que son propias de la conducta humana. Y en cada ocasión, al acabar su cantinela, lanza el recitador la pregunta persuasiva "¿acaso no somos Hombres?".

---

[13] Adela Cortina, *Ciudadanos del mundo*, Alianza, Madrid, 1997, 11-19; *¿Ética o ideología de la inteligencia artificial?*, 18-21. Moralización y hominización –no sólo humanización- irían entonces de la mano.

Pero no, no eran seres humanos, y de ahí que la novela acabe en el rotundo fracaso del experimento. "¿Por qué fracasó el intento de hominización?".

Tal vez la respuesta venga de algo tan sencillo, pero a la vez tan poco reflexionado, como es el hecho de que comprender realmente una ley de la humanidad, una orientación moral, exige contar con un *cuerpo humano* y, estrechamente ligado a él, con ese sentido que permite sintonizar con la ley, que ha recibido el nombre de *sentido moral* o *conciencia* y que lleva entrañados en su seno un conjunto de elementos.

En efecto, recordemos cómo Kant en *La Metafísica de las costumbres* (en la *Tugendlehre*) afirmaba que en los seres racionales hay unas *"Prenociones estéticas de la receptividad del ánimo para los conceptos del deber en general"*, que son de la mayor importancia, porque se refieren a disposiciones morales que, si no se tienen, tampoco hay un deber de adquirirlas[14]. Forman parte de la naturaleza humana y, aunque son estéticas, se les puede calificar a la vez de morales, porque preparan a la persona para escuchar la voz del deber, para comprenderla y para poder obedecerla o desoírla. Esas predisposiciones son, según él, el sentimiento moral, la conciencia moral, el amor al prójimo y el respeto por sí mismo, es decir, la autoestima. Constituyen una especie de antenas de la naturaleza, que permiten percibir el deber.

Si un ser careciera de estas disposiciones no contaría con la base natural indispensable para escuchar la voz de la ley moral y seguirla o rechazarla. Podemos decir, recurriendo a la distinción indispensable en el mundo filosófico entre base y fundamento, que la ley moral sería el fundamento de la obligación moral, pero las disposiciones naturales estéticas serían la base sin la que resulta imposible escuchar la voz de la ley y seguirla[15]. Se trataría, como

[14] Immanuel Kant, *La metafísica de las costumbres*, Tecnos, Madrid, 1989, 253-259.
[15] Adela Cortina, *Neuroética y neuropolítica*, Tecnos, Madrid, 2011, 94-96.

dice Jesús Conill, de una "estética de la libertad"[16] (2006, 41 ss.): es una predisposición del ánimo, estética, pero natural, que consiste en la razón práctica que muestra al hombre su deber en el caso concreto, absolviéndolo o condenándolo. De modo que forma parte de su naturaleza biológica moral la predisposición a juzgarse a sí mismo.

Justamente esta conciencia, como herencia biológica, podría ser el punto de contacto entre la naturaleza y la libertad, al igual que las otras tres prenociones estéticas. ¿Contaban los humanimales con esas bases naturales que hacen posible comprender las exigencias y las invitaciones de la moralidad y que constituyen un requisito imprescindible no sólo para la humanización, sino también para la hominización? Algo semejante venía a decir Charles Darwin cuando afirmaba en *El Origen del hombre*:

> "Suscribo totalmente la opinión de aquellos autores que sostienen que de entre todas las diferencias entre el hombre y los animales inferiores, el sentido moral o conciencia es, con mucho, la más importante. Este sentido, tal como señala Mackintosh, 'tiene una supremacía legítima sobre cualquier otro principio de acción humana'"[17].

La conciencia o sentido moral no sería, pues, según Darwin, una dimensión intrascendente de la vida humana, sino que, a su juicio, marca la diferencia más relevante entre el hombre y el animal y tiene supremacía sobre cualquier otro principio humano; una afirmación en la que insiste reiteradamente a lo largo de su obra.

6. No existen robots autónomos. "Autonomía" e "independencia de terceros" no se identifican.

En un interesante artículo de 2018 la revista *Nature* recogía opiniones de dos millones de personas, enfrentadas a diversos dilemas relacionados con los llamados "coches

[16] Jesús Conill, *Ética hermenéutica*, Tecnos, Madrid, 2006, 41 ss.
[17] Charles Darwin, *El origen del hombre*, Crítica, Barcelona, 2009, 125.

autónomos”[18]. Si en una situación en la que sólo caben dos alternativas un coche no tuviera más remedio que matar a algún ser vivo, ¿a cuál debería sacrificar: animal o ser humano, ocupantes del vehículo o viandantes, persona joven o anciana?

Es éste un tipo de dilema que, aunque en un contexto distinto, planteó ya Philippa Foot en 1967 y que se ha reproducido en múltiples versiones, hasta el punto de recibir el nombre de “dilema del trolley”[19].

Estos dilemas se plantean en relación con las máquinas llamadas “autónomas”, porque actúan con independencia de terceros, es decir, no necesitan constantemente de un operador humano para funcionar. Y en esta mala comprensión del adjetivo “autónoma” radica, a mi juicio, uno de los errores que conlleva confusiones de envergadura en el ámbito de la IA desde su surgimiento. Naturalmente, son seres humanos los que programan esas máquinas, pero después pueden actuar sin necesidad de ser accionadas constantemente. Es el caso de los llamados “vehículos autónomos”, pero también, entre otros, de los “Sistemas de Armas Autónomos” (SAA), propios de los sistemas de defensa, de los que se dice que pueden identificar, seleccionar y atacar objetivos terrestres sin intervención de un operador humano.

Por el momento todos los códigos de ética de la IA exigen que nunca pueda tomarse una decisión sin que la supervise un operador humano. Pero también sería necesario dejar de hablar de Sistemas de Armas Autónomos y de los “killer robots”, que desataron una encendida polémica. De estos “killer robots” se ha dicho que constituyen un atentado contra la dignidad humana, porque son sistemas no humanos los que distinguen entre combatientes y población civil, los que atacan sin discernimiento hospitales, asilos, campamentos ya arrasados. Pero esta polémica pierde su

---

[18] E.Awad, S. Dsouza, R. Kim et alii, “The Moral Machine Experiment”, Nature, vol. 563, nº 7729 (2018),59-64.

[19] Philippa Foot, “The problem of abortion and the doctrine of double effect”, *Oxford Review*, 1967, nº 5, 5-15. Ver también Adela Cortina, 2011, 65-72.

mordiente cuando nos percatamos de que son armas tan poco autónomas como puedan serlo las minas antipersonas. Los robots carecen de conciencia y de autonomía, de ahí que sea necesario averiguar la trazabilidad de la actuación, porque los responsables son, en último término, personas humanas.

La respuesta que ha generado mayor consenso entre los especialistas es la de que esos robots supuestamente autónomos, gozan de *autonomía funcional*, *pero no de autonomía ontológica y ética*, funcionan *como si* fueran autónomos, y con eso –se dice- nos basta para fabricarlos y utilizarlos. La cuestión sería entonces *ingenieril*, *no filosófica*, y con esta respuesta nos bastaría para funcionar.

Pero este acuerdo tan razonable tiene sus limitaciones porque, velis nolis, nos incita a ponernos en manos de la *Philosophie des Als Ob*, la *Filosofía del Como si* de Hans Vaihinger (1911), según la cual, los seres humanos nunca podemos conocer la realidad subyacente del mundo. Construimos sistemas de pensamiento y asumimos que encajan con la realidad. Nos comportamos como si el mundo encajara en nuestros modelos.

Aceptar la autonomía funcional para explicar la agencia de los robots tiene unas consecuencias que es preciso considerar en profundidad. En efecto, Danaher propone que en el caso de todos los sujetos agentes, también los humanos, se aplique un conductismo metodológico, como el que aplican los psicólogos y los científicos de la computación, por la siguiente razón: los robots amorosos actúan como si sintieran emociones, y no podemos traspasar la capa de la conducta llegando a la intimidad, pero igual ocurre con los seres humanos. No podemos llegar a esas propiedades metafísicas que solemos aducir para atribuirnos una conducta moral: la conciencia, las capacidades cognitivas, los intereses. Lo que hacemos es inferir esas propiedades a partir de su conducta. Podríamos hacer lo mismo con los sistemas inteligentes.

¿Basta con esta respuesta, o es necesario ir más allá para ajustarse a la realidad, que es la forma de evitar problemas indeseables?

7. Es imposible extender a las personas humanas el concepto de autonomía funcional.

Ciertamente, a la capacidad de cambiar de estado sin un estímulo externo se denomina a menudo "autonomía", pero, a mi juicio, es más adecuado denominarle "independencia externa" porque entiendo que la *autonomía* es una capacidad más exigente que la *independencia externa* y está ya relacionada con el mundo moral.

En cuanto a la capacidad de actuar por razones, se dice que un sistema inteligente puede deliberar sobre planes y ejecutarlos, contando con las informaciones que se le han proporcionado, tanto de hechos como de finalidades. Pero lo bien cierto es que no delibera en el sentido humano, sino de nuevo que actúa por razones en *un sentido funcional*, que actúa con representaciones funcionalmente equivalentes a las de los agentes humanos[20]. Se toman, pues, estos vocablos en un sentido ingenieril, no ontológico: "como si" fueran autónomos. Las diferencias entre la agencia moral de los seres humanos, que es una *agencia moral plena*, y la *artificial* serían las siguientes:

1) Los agentes morales plenos están dotados de creencias, preferencias, emociones morales (empatía, simpatía, culpabilidad, vergüenza), voluntad libre y la capacidad de reflexionar sobre sus razones morales, modificarlas, justificar las bases normativas de sus decisiones morales y tener en cuenta otros cursos de acción.

[20] Wendell Wallach y Allen Colin, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, New York 2008, 39; Mark Coeckelbergh, "How I learned to Love the Robot", en I,Oostertaken y J. van den Hooven (comps.), *The Capability Approach, Technology and Design*, Springer, Dordrech, 2012, 77-86; Mark Coeckelbergh, *Ética de la inteligencia artificial*, Cátedra, Madrid, 2021; Adela Cortina, *Neuroética y neuropolítica*, caps. 6 y 7; *¿Ética o ideología de la inteligencia artificial?*, 2024, cap. 7.

Esto es clave para la libertad de la voluntad. La agencia moral plena se aplica potencialmente a cualquier contexto, como es propio de la inteligencia general de los humanos.

Los sistemas inteligentes, por su parte, limitan su competencia a un área específica de actuación, dado que gozan de una inteligencia especial. Lo cual tiene sus ventajas para su uso, porque necesitan aplicar principios morales a los casos concretos para los que pueden emplearse, como puede ser el cuidado de ancianos, los vehículos autónomos, el trato con mascotas, con robots sexuales, o los conflictos bélicos. Pero son incapaces de comprender el contexto en el que han de aplicar los principios y sin comprensión no hay actuación moral responsable.

2) Los seres humanos pueden reflexionar sobre su conducta a través de ese "pensar despacio" del que habla Kahneman, y son capaces de entablar un diálogo con gentes cercanas, lo que les permite descubrir los móviles de su acción en el intercambio de argumentos. Con las máquinas es imposible, porque no se comunican. Y en este punto no ayuda en absoluto la inteligencia artificial generativa, incapaz de sacar a la luz una intención de sus "acciones". La atribución de responsabilidad exigiría que pudieran actuar según el deber y por mor del deber, cosa que no pueden hacer porque carecen de autonomía, voluntad libre y conciencia.

3) Los agentes morales plenos, a la hora de abordar el mundo, pueden adoptar dos perspectivas: la perspectiva del observador y la del participante. En el primer caso se expresan en un lenguaje formulado en tercera persona, la del observador de objetos y acontecimientos; en el segundo caso, se expresan en un lenguaje formulado en primera persona, adoptan la perspectiva del participante en acciones comunicativas con otros sujetos, que son a su vez participantes de modo actual o virtual y se reconocen mutuamente como interlocutores válidos. Estos dos tipos de lenguaje han surgido en el proceso de la evolución en comunidades lingüísticas y cooperativas y han resultado

indispensables para la vida humana[21]. El participante en las acciones comunicativas revela la existencia de la intersubjetividad, del vínculo con los demás sujetos. Es desde esta perspectiva desde la que los sujetos se reclaman mutuamente razones de sus actos y dan razones de ellos, con lo cual están presuponiendo que son responsables de sus actos, que podían haber actuado de forma distinta. La conciencia de la libertad y, por tanto, de la responsabilidad aflora a través del lenguaje expresado en primera persona en las acciones comunicativas. Los sistemas de IA son incapaces comunicarse desde un trasfondo de intersubjetividad. No pueden ser personas.

El lenguaje en que se intercambian razones es indispensable para que exista el lenguaje referido a la observación. El trasfondo de un mundo vital compartido intersubjetivamente es necesario para que haya objetividad del saber. Como bien decía Hölderlin, "somos un diálogo".

Llevaba, pues, razón Rudyard Kipling en su escrito sobre "El secreto de las máquinas" cuando se dirige a ellas con las siguientes palabras:

"Mas recuerda la ley que rige nuestras vidas;

No estamos construidas para entender mentiras,

No podemos amar, llorar ni perdonar.

Si nos manejas mal, encontrarás la muerte.

Somos mayores que el Hombre o los Reyes.

Sé humilde cuando te arrastres por debajo de nuestras bielas,

Nuestro toque puede cambiar todas las cosas creadas.

Somos todo en el mundo, salvo los dioses.

Aunque nuestro humo oculte los cielos de sus ojos,

[21] En este sentido dirá Habermas "el dualismo epistémico tiene que haber surgido de un proceso de aprendizaje evolutivo y haberse acreditado en la confrontación cognitiva del homo sapiens con un mundo lleno de riesgos" (Entre naturalismo y religión, Paidós, Barcelona, 2006, 174).

Desaparecerá y las estrellas volverán a brillar;

Porque pese a nuestra fuerza, peso y magnitud,

No somos más que el producto de su intelecto”.

También esta advertencia es certera para las máquinas de la revolución 4.0.