Ni animales ni máquinas: conocimiento humano y responsabilidad epistémica

CLARIDADES
REVISTA DE FILOSOFÍA
ISSN 1889-6855 · ISSN-E 1989-3787

Neither animals nor machines: human knowledge and epistemic responsibility

María Caamaño Alegre

Universidad de Valladolid (España)

Fecha de envío: 02/03/2024
Fecha de aceptación:13/09/2024

DOI: 10.24310/crf.16.2.2024.19186

Resumen

A pesar de las características compartidas con los animales y las máquinas, sólo los humanos poseemos una esfera de responsabilidad epistémica, encontrándose esta asociada a rasgos específicamente humanos, como la conciencia, el conocimiento reflexivo y la libertad. Asumiendo algunas de las tesis principales defendidas desde la epistemología de la virtud, se argumentará que hay una forma de conocimiento propiamente humana, emergida a partir del conocimiento animal, y que las máquinas sólo alcanzan a simular. Se reivindicará la importancia de la contribución de William Kingdon Clifford (1877)

acerca de la ética de la creencia, pues no sólo nos ayuda a reconocer el valor vital de las creencias verdaderas, sino que nos advierte además sobre nuestra capacidad para decidir el modo en el que nos formaremos las creencias.

Palabras claves

Conocimiento humano; responsabilidad epistémica; conciencia; epistemología de la virtud.

Abstract

Despite the characteristics shared with animals and machines, only humans possess a sphere of epistemic responsibility, this being associated with specifically human traits, such as consciousness and reflective knowledge. Taking up some of the main propositions of virtue epistemology, it will be argued that there is a form of knowledge that is unique to humans, that has emerged from animal knowledge, and that machines can only simulate. The importance of William Kingdon Clifford's (1877) contribution on the ethics of belief will be vindicated, since it not only helps us to recognize the vital value of true beliefs, but also warns us about our ability to decide the way in which we will form our beliefs.

Keywords

Human knowledge; epistemic responsibility; consciousness; virtue epistemology.

«*El imitador, por ende, no tendrá conocimiento ni opinión recta de las cosas que imita, en cuanto a su bondad o maldad*».

(Plat. Rep. 602a).

«*"But", says one, "I am a busy man; I have no time for the long course of study which would be necessary to make me in any degree a competent judge of certain questions, or even able to understand the nature of the arguments". Then he should have no time to believe*».

(William Kingdom Clifford «The Ethics of Belief», Contemporary Review, 29, May, 1876: Diciembre-1877: 289-309: 295).

I. Introducción

No cabe duda de que los humanos comparten características con los animales y las máquinas, que, a su vez, comparten propiedades entre sí. En un delicioso texto de Thomas Henry Huxley, «On the Hypothesis that Animals are Automata, and its History» (1874), el autor desgrana la idea de que los fenómenos vitales, al igual que todos los del mundo físico, son susceptibles de explicación mecánica, y que el estudio de la biología, a la larga, es una aplicación de las grandes ciencias de la física y la química. Describe la posición de Descartes desde la cual los animales son meras máquinas o autómatas, desprovistos no sólo de razón, sino de cualquier atisbo de conciencia.[1] Por supuesto, la noción de conciencia que adopta Descartes presenta a ésta como teniendo una naturaleza eminentemente reflexiva, dándose en la forma de autoconciencia. Como se verá en la sección 3 (Block 1995), la noción de conciencia puede aplicarse en un sentido mucho más amplio, que incluya formas de conciencia sólo consistentes en la captación sensible del mundo[2].

El debate del siglo XIX acerca del solapamiento o no entre el reino animal y el de las máquinas se ha transformado, en el siglo XXI, en una

1. Descartes expondría brevemente esta doctrina en el «Discours de la Méthode», y más detalladamente en las «Réponses aux Quatrièmes Objections», y en la correspondencia con Henry More (Huxley 1874: 363).

2. Existe cierto desacuerdo, entre los especialistas, con respecto a si Descartes aceptaba que los animales tuviesen sensibilidad. El minucioso estudio de Peter Harrison (1992) apunta de forma convincente a que esto no era el caso, enfatizando, no obstante, cómo la concepción restrictiva cartesiana de la conciencia (como autoconciencia) no era compatible con la atribución de una conciencia meramente sensible a los animales. En lo que atañe a la posición de Descartes, por tanto, cabe aclarar varias cuestiones importantes entrelazadas, por una parte, los animales poseen sensibilidad, por otra, no poseen conciencia (en el sentido cartesiano, esto es, autoconciencia), siendo, por ello, al contrario que los humanos, meros autómatas. No olvidemos que, como señala Carlos Moya (2006, capítulo 2), desde el enfoque cartesiano, lo mental, o el pensamiento, es todo aquello de lo que tenemos conciencia, y por conciencia hemos de entender un modo de conocimiento directo, inmediato e infalible, en el que no cabe el error. Lo mental se caracteriza por ciertos rasgos epistemológicos, como son, la inmediatez e infalibilidad con que es conocido por el sujeto. La mente de cada cual constituye un ámbito privado formado por sus contenidos o datos inmediatos de conciencia, a los que el sujeto accede directa e infaliblemente, abriéndose con ello la posibilidad de razonar libremente.

controversia sobre el solapamiento o no entre la inteligencia humana y la inteligencia artificial. Si diferenciamos los tres dominios (el animal, el humano, el de las máquinas), y dejando al margen la cuestión de sus posibles solapamientos, podríamos preguntarnos: ¿cómo conocen los humanos? ¿conocen los animales? ¿conocen las máquinas? El interrogante que en la actualidad suscita un mayor interés es este último, que suele adoptar la forma siguiente: ¿pueden conocer las máquinas como lo hacen los humanos?

Las respuestas posibles son, por supuesto, diversas. Una posibilidad es negar que las máquinas conozcan, y entender que su capacidad es exclusivamente imitativa. Las máquinas sólo simularían conocer, imitando para ello a los humanos. Otra opción es considerar que, efectivamente, las máquinas conocen, si bien no a la manera humana. Como tercera posibilidad, podría afirmarse que las máquinas pueden conocer del mismo modo como lo hacen los humanos.

Negar que las máquinas conozcan al modo humano puede argumentarse de distintas formas, de las cuales mencionaré dos, no sólo compatibles, sino incluso conectadas en el sentido de que la segunda forma de argumentar presupone la primera. Una posible línea argumentativa sería la de objetar que, para conocer en sentido humano (o para conocer en absoluto), hace falta conciencia (intencionalidad, sensibilidad, experiencia consciente). El sujeto de conocimiento sólo se constituye como sujeto consciente, lo cual requiere una orientación hacia algo distinto del propio sujeto, que el sujeto percibe en esos términos, es decir, como distinto de sí mismo. Este requisito de consciencia vinculado a la representación no se satisfaría en el caso de las máquinas, pues su detección de propiedades externas al sistema no sería consciente, sólo el humano que emplea las máquinas entiende lo que representan ciertos signos (outputs) obtenidos a partir de otros signos (inputs). El conocido ejemplo de ficción de John Searle (1980, 1983/2002, 1999) sobre la «habitación china» invoca el mismo tipo de razones para negar que las máquinas puedan atribuir significado al modo humano, esto es, interpretando los signos lingüísticos como refiriendo a entidades del mundo. El argumento de Searle suponía un intento de refutación de aquellos puntos de vista que, apelando a la indistinguibilidad entre el comportamiento humano y el de los ordenadores en la realización de ciertas prácticas o tareas, justificaban la atribución de la misma capacidad

en relación con ellas. Alan Turing (1950), uno de los teóricos pioneros de la informática, propuso lo que hoy se conoce como «la prueba de Turing», conforme a la cual, si un ordenador puede hacerse pasar por humano en una conversación en línea, debemos conceder que es inteligente. Searle considera que la prueba de Turing no es válida para detectar si humanos y ordenadores comparten las mismas capacidades cognitivas. Si nos imaginamos la situación propuesta por él, tendríamos a un hablante nativo de inglés que no sabe chino encerrado en una habitación llena de cajas de símbolos chinos (la base de datos) junto con un libro de instrucciones para manipular los símbolos (el programa). Personas ajenas a la habitación enviarían otros símbolos chinos, igualmente desconocidos para la persona que está en la habitación, con preguntas en chino (el input). Siguiendo las instrucciones del programa, la persona que está en la sala es capaz de enviar símbolos chinos que son respuestas correctas a las preguntas (el output). A pesar de que la persona en la habitación, gracias al manejo del programa, pasaría la prueba de Turing, no entendería ni una palabra de chino. La contraposición que Searle trata de establecer entre la capacidad cognitiva humana y la capacidad computacional de los ordenadores se refleja en la siguiente cita:

> But this feature of programs, that they are defined purely formally or syntactically, is fatal to the view that mental processes and program processes are identical. And the reason can be stated quite simply. There is more to having a mind than having formal or syntactical processes. Our internal mental states, by definition, have certain sorts of contents. If I am thinking about Kansas City or wishing that I had a cold beer to drink or wondering if there will be a fall in interest rates, in each case my mental state has a certain mental content in addition to whatever formal features it might have. That is, even if my thoughts occur to me in strings of symbols, there must be more to the thought than the abstract strings, because strings by themselves can't have any meaning. If my thoughts are to be *about* anything, then the strings must have a *meaning* which makes the thoughts about those things. In a word, the mind has more than a syntax, it has a semantics. The reason that no computer program can ever be a mind is simply that a computer program is only syntactical, and minds are more than syntactical. Minds are semantical, in the sense that they have more than a formal structure, they have a content (1983/2002: 670-671).

La conclusión a que nos conduciría el argumento de Searle resulta clara: incluso si los ordenadores resultan indistinguibles de los humanos en los resultados que obtienen resolviendo tareas cognitivas, su capacidad

radica exclusivamente en simular las facultades humanas, pues carecerían de conciencia o intencionalidad (orientación consciente hacia el mundo). Careciendo de un lenguaje con significado, los ordenadores no podrían pensar sobre nada. En palabras del autor:

> The whole point of the parable of the Chinese room is to remind us of a fact that we knew all along. Understanding a language, or indeed, having mental states at all, involves more than just having a bunch of formal symbols. It involves having an interpretation, or a meaning attached to those symbols. And a digital computer, as defined, cannot have more than just formal symbols because the operation of the computer, as I said earlier, is defined in terms of its ability to implement programs. And these programs are purely formally specifiable-that is, they have no semantic content (1983/2002: 671).

Una segunda línea argumentativa, en contra de la tesis de que las máquinas puedan conocer como lo hacen los humanos, se articula a partir del reconocimiento de rasgos irreductiblemente humanos, que se dan en forma de propiedades entrelazadas, como libertad-responsabilidad (epistémicas), volición-evaluación (epistémicas). El reconocimiento de valores epistémicos, asociados al deseo de conocer, de poder anticipar el futuro o comprender las causas de lo que acontece, va unido a la búsqueda de una estrategia para maximizar la satisfacción de valores epistémicos. Estas propiedades se suelen reconocer como constitutivas de lo humano en general, y del conocimiento humano en particular, a la vez que como no compartidas con las máquinas (ni los animales). El conocimiento humano no emergería sólo de la conciencia, sino también de nuestra capacidad para valorar y establecer fines en función del valor que les atribuyamos a la consecución de cierto estado de cosas. Todo ello requiere de una consciencia reflexiva, esto es, de una comprensión de los propios deseos y fines cognitivos, así como de una representación de las distintas opciones de que el sujeto dispone para la consecución de dichos fines. No parece que el establecimiento reflexivo de fines sea un rasgo compartido con los animales, ni con las máquinas. De lo contrario, sin duda, viviríamos en un mundo muy distinto del que habitamos, donde los humanos no seríamos los únicos sujetos de decisión. Ambas cosas, reconocimiento de opciones interesantes para el propio sujeto y decisión sobre qué opción escoger, presuponen un ejercicio coordinado de reflexión y valoración.

Todos estos interrogantes, así como los posibles argumentos en apoyo de respuestas afirmativas o negativas a ellos, remiten a importantes debates en la filosofía de la mente. Explorar esas remisiones excede con mucho los límites de la presente reflexión. Me ceñiré, por tanto, a un tratamiento netamente epistemológico de la cuestión, sin apoyarme en ningún enfoque en particular acerca de la naturaleza de lo mental. Ciertamente, ese tratamiento hará más plausibles algunas aproximaciones a los fenómenos mentales frente a otras, pero de ello no pretende seguirse ninguna conclusión acerca de cuál pueda ser la aproximación correcta. Como mucho, quedará patente la relevancia que poseen ciertas tesis sobre lo mental para poder explicar aquellos rasgos epistemológicos que consideraré específicamente humanos. Por ejemplo, el teleofuncionalismo (Lycan 1987, 1996; Dennett 1991) permite esclarecer cuáles son los rasgos de lo mental y los límites del criterio de indistinguibilidad en la conducta para justificar la atribución de una misma propiedad mental. Lo idéntico en la conducta puede cumplir funciones muy distintas dependiendo de cómo se encuentren compuestos jerárquicamente los sistemas intencionales y de cuáles sean las funciones biológicas de las capacidades representacionales. El monismo anómalo (Davidson 1970/1980, 1973/1980, 1974/1980, Moya 2009, capítulo 6, 117-132), por otra parte, constituye un enfoque interesante al dar cuenta de manera comprehensiva de muchos de los rasgos reconocidos en el presente ensayo como específicos del conocimiento humano.

Los supuestos de los que partiré son los siguientes:

- Los seres humanos sabemos sobre qué obtenemos conocimiento, sobre qué es la información obtenida, conocemos el dominio objeto de conocimiento o sobre el que se quiere adquirir conocimiento. Este supuesto, de raíces cartesianas, implica el reconocimiento de la autoconciencia, del acceso a nuestros propios pensamientos. Posee igualmente una dimensión searleana, pues invoca nuestra capacidad para orientarnos hacia el objeto de nuestros estados mentales.

- Los sujetos humanos tenemos libertad para conocer: podemos fijar libremente objetivos epistémicos, escoger entre ellos, así como entre estrategias para obtenerlos. Es decir, tal y como coincidían en apuntar Descartes y Huxley, incluso desde posiciones muy alejadas, la cognición humana no se haya determinada mecánicamente.

Los sujetos humanos podemos representarnos distintas opciones epistémicas entre las que escoger.

- Los sujetos humanos reconocemos el conocimiento como un bien valioso. Se trata de un supuesto implícito en la tradición de reflexión epistemológica desde la antigüedad.

- Los sujetos humanos somos responsables de nuestras decisiones sobre qué y cómo conocer. Este supuesto se sigue del segundo, según el cual los humanos poseemos un alto grado de libertad en nuestra conducta cognitiva.

Defenderé que las máquinas no comparten con los humanos ninguna de las cuatro características arriba mencionadas (conciencia, libertad, normatividad, responsabilidad), si bien pueden simularlas.[3] Los animales, por su parte, hasta donde tenemos conocimiento, sólo comparten parcialmente el primero de los aspectos señalados. Me apoyaré principalmente en ideas de William Kingdom Clifford (1876-77) y Ernest Sosa (1991, 2007) para defender este punto de vista. El primero pone de relieve la importancia de la ética de la creencia, no sólo en lo que respecta a la vertiente moral de la actividad cognitiva, sino también en lo relativo a la significación cognitiva de la conducta moral, responsable, en relación con la formación de creencias. Sosa proporciona el marco general en el que la justificación de creencias emerge como indisolublemente ligada al ejercicio de la razón reflexiva y, en última instancia, a la responsabilidad del sujeto en lo que atañe al uso de dicha facultad.

Difícilmente se puede hablar de responsabilidad moral sin hablar de conciencia. Una cosa presupone la otra (aunque no a la inversa, es decir, la conciencia no presupone la responsabilidad moral). Ya que mi argumento respalda la tesis más fuerte, estaré asumiendo también la tesis más débil de que el conocimiento requiere de conciencia. Por último, resulta importante diferenciar dos tesis distintas:

a) Hay un conocimiento específicamente humano.

b) El conocimiento humano es el único que existe.

3. Tal vez las máquinas no estén todavía programadas para simular la conducta moral, sin embargo, podría llegarse rápidamente a ello a través de iniciativas como el instituto «Moralities of Intelligent Machines» (Universidad de Helsinki), cuyo objetivo es programar e implementar un código de conducta moral en las máquinas.

Argumentaré sólo en favor de a), de lo cual no se sigue lógicamente b). Con todo, defender a) sí implica negar que otros seres o entidades de nuestro entorno hayan llegado a conocer a la manera humana.

II. ¿Qué es conocer? Una respuesta desde la epistemología de la virtud

Suele atribuirse a Sócrates la primera formulación de la idea de que el conocimiento consiste en creencias verdaderas bien justificadas. En el *Teeteto de Platón*, se describe a un Sócrates convencido de que la mera creencia verdadera no constituye conocimiento, pues carece de una explicación sobre qué es lo que la hace verdadera. La capacidad para poder ofrecer dicha explicación no es otra cosa que la capacidad para justificar nuestra creencia en la verdad de cierta proposición. En palabras de Sócrates (quien describe lo oído en un sueño):

> Efectivamente, quien no puede dar y recibir una explicación de algo carece de saber respecto de ello. Sin embargo, si alcanza una explicación, todo esto le es posible hasta lograr la plena posesión del saber (Plat. *Teeteto*, 202c).

La explicación de la cosa conocida conlleva la posibilidad de explicar el juicio acerca de la cosa. Si puedo explicar la naturaleza de un objeto, puedo explicar por qué creo que cierta cosa es ese objeto. Pero, ¿qué es explicar? Sócrates y *Teeteto* repasan los distintos sentidos de «explicar» (Plat. *Teeteto*, 206d, 207a, 207d-208a, 208c, 209a-d). Diferencian entre tres sentidos: verbalizar el juicio acerca de la cosa, dar razón de cada cosa por los elementos que la componen y poder decir en qué una cosa difiere de todas las demás. Como es bien sabido, Sócrates no considera que estos tres sentidos de explicación sean suficientes para explicar lo que es el conocimiento (Plat. *Teeteto*, 210c). Sin embargo, todos ellos contribuirían a esclarecer la diferencia entre el conocimiento y la mera opinión verdadera. Dentro del marco socrático, el sujeto de conocimiento ha de poder dar las razones por las que cree que cierto juicio es verdadero y sólo tendrá conocimiento si sus razones son la correctas, es decir, si explican adecuadamente la verdad del juicio.

Gran parte del desarrollo de la teoría del conocimiento ha girado en torno a la cuestión de la justificación de las creencias, dando lugar a distintas posiciones internistas o externistas con respecto a ella (Grimaltos

& Iranzo, 2009). En relación con la justificación de creencias (o lo que Sócrates denominaba «explicación»), lo que se trata de dirimir es si debe o no estar siempre al alcance (cognitivo/epistémico) del sujeto. Las discrepancias surgen en relación con la cuestión de si el sujeto de conocimiento ha de tener un acceso consciente a aquello que justifica su creencia para que ésta pueda estar justificada. Según la posición internista, que es la adoptada por Sócrates, aquello que justifica nuestras creencias debe ser accesible epistémicamente por introspección o reflexión. Conforme a la posición externista, dicho acceso resulta insuficiente y, en muchos casos, también innecesario, para la justificación de creencias. A pesar del impulso cada vez mayor recibido por el externismo en las últimas décadas, muchas de nuestras intuiciones preanalíticas son de carácter internista, por ejemplo, la idea fuertemente arraigada de que la justificación depende en cierta medida de la capacidad para dar razones, y la expectativa de que la conducta cognitiva del sujeto en relación con la proposición creída ha de ser correcta (no ignorar evidencias disponibles, prestar atención a objeciones, etc…), lo que en parte dependerá de las decisiones que tome el propio sujeto. Desde el fiabilismo, por ejemplo, se considera que la justificación de una creencia depende de que haya sido generada a partir de un conocimiento fiable, entendiéndose la fiabilidad como la alta frecuencia de éxito en la generación de creencias verdaderas.

Desde la epistemología de la virtud se han combinado condiciones tanto internistas como externistas en lo referente a la justificación de creencias. Su principal impulsor, Ernest Sosa (2007), entiende las virtudes como facultades fiables, es decir, como habilidades que garantizan una alta frecuencia de éxito en la obtención de ciertos logros. No obstante, objeta al fiabilismo que subestime la relevancia de la evidencia aducida por los sujetos, pues Sosa considera que sí sería relevante tomar en cuenta la visión que el sujeto tiene sobre la fiabilidad de los procedimientos que aplica. Las dudas que un sujeto albergue acerca de la fiabilidad de los procedimientos seguidos interferirá con sus otras creencias. Son estas creencias de segundo orden, acerca de otras creencias, las que posibilitan que el sujeto desarrolle lo que el autor denomina una «perspectiva epistémica» (Sosa 1991). El desarrollo de dicha perspectiva supone el ejercicio de una razón reflexiva y es característica del conocimiento humano frente al animal. Los animales podrían desarrollar una conducta cognitiva apta, pero no una perspectiva

epistémica, al no poseer la virtud de la razón reflexiva. Los humanos poseerían ambas formas de conocimiento, el animal y el específicamente humano, de ahí que obtengan logros epistémicos, tanto en la forma de creencias «aptas», como en la de creencias justificadas, a través del conocimiento reflexivo.

En su intento de caracterizar la aptitud, Sosa ha utilizado a menudo (Sosa, 2007: capítulo 2) la analogía de un arquero experto disparando a un blanco. He aquí dos formas de evaluar el tiro de un arquero: 1) ¿tuvo éxito? o ¿ha dado en el blanco?; 2) ¿pone la ejecución del tiro de manifiesto la habilidad del arquero? o ¿se produjo de forma que tuviera probabilidades de éxito? Sosa denomina precisión al tipo de acierto y destreza al tipo de habilidad que se discute en (2). Un tiro es hábil si se realiza con destreza. Los tiros hábiles no tienen por qué ser precisos, ya que no todos los tiros hábiles tienen éxito. Y los tiros precisos no tienen por qué ser hábiles, ya que algunos tiros no hábiles son afortunados. Además de la precisión y la habilidad, Sosa sugiere que hay otro aspecto en el que se puede evaluar un tiro, relacionando ambos. Sosa lo denomina aptitud. Un tiro es apto si es preciso porque es hábil. La aptitud implica, pero requiere más que la conjunción de precisión y destreza, pues ha de definirse de modo que se excluya la posibilidad de que interfiera la suerte epistémica de distintas formas. Sosa sugiere que este modelo de evaluación es aplicable de forma bastante general a la evaluación de cualquier acción u objeto con un objetivo característico. En particular, es aplicable a la creencia con respecto a su objetivo de verdad:

- Una creencia es exacta si y sólo si es verdadera.
- Una creencia es hábil si se produce con destreza.
- Una creencia es apta si y sólo si es verdadera de un modo que manifieste, o sea atribuible, a la habilidad del creyente.

Es decir, se requiere que la habilidad explique el acierto, pues esta condición es la que permite excluir que la relación entre justificación y verdad se deba a una mera coincidencia. Desde este punto de vista, el conocimiento implica tanto la verdad (exactitud) como la justificación (destreza), pero no como meros componentes independientes de los que el conocimiento se compone funcionalmente. Las principales propiedades características de las creencias aptas se cumplen en el caso de los logros cognitivos de los animales y las máquinas. Los humanos compartiríamos

con animales y máquinas la aptitud en la realización de ciertas tareas cognitivas. Sin embargo, el paso al conocimiento reflexivo parece exclusivo de los humanos.

El conocimiento se vuelve reflexivo cuando, además de saber algo, nos formamos creencias de segundo orden sobre la fiabilidad de las fuentes de nuestro conocimiento. Las creencias, también las de primer orden, han de estar ubicadas en una perspectiva epistémica para que puedan estar justificadas. La capacidad del sujeto para evaluar la calidad de su aparato cognitivo se plasma en las creencias de su perspectiva epistémica o creencias de segundo orden. En los humanos no resulta suficiente el comportamiento virtuoso de primer orden, ha de haberlo también en el segundo orden. Ya no es únicamente la dotación del sujeto lo que determina su acceso al conocimiento, sino su uso de la razón reflexiva para formarse creencias adecuadas sobre su dotación cognitiva.

Al asumir el planteamiento de Sosa, estamos alejándonos de dos tendencias imperantes en la epistemología actual. Por una parte, la tendencia deflacionaria (Allen, 2017) alimentada por la creciente apreciación científica de las capacidades de los organismos (bacterias, plantas, etc.) y colectivos (enjambres de insectos, grupos humanos) que antes se consideraban demasiado diferentes de la cognición humana prototípica como para merecer la atención de los científicos cognitivos. En muchos casos, el comportamiento adaptativo se considera un comportamiento inteligente, en un sentido de «inteligencia» que prescinde de gran parte de los elementos arriba mencionados. Por otra parte, el enfoque asumido también nos distancia de otra opción muy extendida dentro del discurso epistemológico, como es la de dejar abierta la cuestión definicional acerca del conocimiento, recurriendo simplemente a convenciones fructíferas para la investigación interdisciplinar del comportamiento inteligente en humanos, animales y robots y favoreciendo la creación de conceptos racimo a fin de caracterizar aquello que llamamos «conocimiento» (Newen, 2017).

III. El papel de la conciencia

Nos orientamos hacia el mundo, pero también hacia estados de cosas posibles que nos representamos, hacia distintos fines y estrategias cognitivas. Es la conciencia de nosotros mismos como sujetos con estas capacidades lo que nos convierte en seres con responsabilidad epistémica. La autorrepresentación

constituye el núcleo de la conciencia reflexiva característica de la perspectiva epistémica y, por tanto, del conocimiento propiamente humano. De la misma forma que Sosa diferencia entre conocimiento animal y humano, podemos diferenciar entre conciencia animal y humana (participando la primera de la segunda, pero no a la inversa). No parece que las máquinas tengan conciencia en ninguno de los dos sentidos. De hecho, por más que las máquinas puedan mostrar cierta «sensibilidad» al contexto, no llegan a representarse un mundo exterior como tal, ni a sentirse como máquinas. Hay algo que consiste en sentirse humano, o algo que es sentirse un murciélago (Nagel, 1974), pero no hay nada que sea sentirse como una máquina. Los animales y humanos no compartimos la conciencia reflexiva, pero sí otro tipo de conciencia.

De nuevo, históricamente, no siempre se reconoció a los animales como seres conscientes. Descartes le atribuía conciencia sólo a los humanos. Huxley discrepaba en relación con esta cuestión, reconociéndoles también conciencia a los animales. Encontramos, en Huxley, una posición muy distinta, pues, no sólo emplea una noción más amplia de conciencia que la cartesiana, aceptando una conciencia sensible, sino que además considera la autoconciencia un rasgo gradual no exclusivamente humano. Resulta posible identificar, no obstante, un punto de coincidencia entre ambos autores, en particular, en lo referido a la noción de autómata, que ambos aplican a los animales en virtud de su explicabilidad puramente mecánica. Huxley explicita este aspecto de los animales de la siguiente manera:

> When we [people in general] talk of the lower animals being provided with instinct, and not with reason, what we really mean is, that although they are sensitive and although they are conscious, yet they act mechanically, and that their different states of consciousness, their sensations, their thoughts (if they have any), their volitions (if they have any), are the products and consequences of their mechanical arrangements (1874: 365).

A pesar su veneración hacia el genio filosófico y científico de Descartes, Huxley rechaza las implicaciones que la tesis mecanicista cartesiana posee en relación con la conciencia (su ausencia en los animales) y despliega un argumento evolutivo continuista o gradualista, según el cual, la conciencia no puede ser exclusiva de los humanos, pues ha debido evolucionar a partir de estadios anteriores de la evolución en los que ya habría surgido en una forma menos compleja. Recuperemos el pasaje de Huxley al respecto:

> But though I do not think that Descartes' hypothesis can be positively refuted, I am not disposed to accept it. The doctrine of continuity is too well established for it to be permissible to me to suppose that any complex natural phenomenon comes into existence suddenly, and without being preceded by simpler modifications; and very strong arguments would be needed to prove that such complex phenomena as those of consciousness, first make their appearance in man. We know, that, in the individual man, consciousness grows from a dim glimmer to its full light, whether we consider the infant advancing in years, or the adult emerging from slumber and swoon. We know, further, that the lower animals possess, though less developed, that part of the brain which we have every reason to believe to be the organ of consciousness in man; and as, in other cases, function and organ are proportional, so we have a right to conclude it is with the brain; and that the brutes, though they may not possess our intensity of consciousness, and though, from the absence of language, they can have no trains of thoughts, but only trains of feelings, yet have a consciousness which, more or less distinctly, foreshadows our own (1874: 363).

Es interesante advertir que, el argumento de Descartes para negar que los animales tengan conciencia (sólo simularía tenerla) se parece al argumento actual, y recuperado aquí, para negarle conciencia a las máquinas. Sin embargo, una diferencia crucial entre ambos argumentos tiene que ver con el hecho de que el argumento de tipo gradualista de Huxley no se aplica en el caso de las máquinas y, por otra parte, sólo en el caso de los humanos se da el sentirse como algo. Hay algo que es sentirse humano o como un humano, pero no hay nada que sea sentirse un ordenador (o sentirse como un ordenador). Ello va parejo a la posesión de lo que Giulio Tononi denomina formas intrínsecas (o constitutivas de experiencia) frente a formas atribuidas (no constitutivas de experiencia en las entidades a las que se les atribuyen). Si no hubiese eso, entiende el autor, no habría conceptos, ni significado, ni propósitos. En consecuencia, si no hay experiencia, no hay tampoco conocimiento. Una vez más, ha de enfatizarse la insuficiencia de la indistinguibilidad en lo que se hace para atribuir las mismas propiedades, entre ellas, la conciencia. Debe prestarse atención a cómo se hace algo, si se hace experimentando el mundo, sintiéndose como de cierta forma. A juicio de Tononi (2008), el cerebro humano consigue integrar información de una forma altamente sofisticada y *sui generis.* Por ello, los humanos no sólo reaccionamos de cierta forma, sino también ante cierta experiencia del mundo. No sólo suministramos outputs antes ciertos inputs, sino que

escogemos el modo de suministrar outputs dados ciertos inputs a los que atribuimos mayor o menor valor representacional del mundo.

Los humanos disponemos de las cuatro formas de conciencia diferenciadas por Ned Block (1995): 1) conciencia fenoménica (del mundo sensible), 2) conciencia de acceso (a los propios contenidos mentales para operar con ellos racionalmente), 3) autoconciencia (posesión del concepto de uno mismo), 4) conciencia de monitorización (control de nuestra propia conducta). La distinción de Block suscita muchos interrogantes que él mismo aborda. Concede que es posible poseer conciencia fenoménica sin conciencia de acceso, pero, a su parecer, resulta más complicado elucidar si resulta posible posee conciencia de acceso sin conciencia fenoménica. Tal vez ese sería el caso de los robots (Block, 1995: 211-12). Un argumento contrario a dicha posibilidad podría plantearse siguiendo las líneas del argumento de la habitación china, ahora incluyendo los qualia, o el aspecto subjetivo de la conciencia fenoménica, como el ingrediente posibilitador de la conciencia del mundo sensible. Daniel Dennett (1988) rechaza frontalmente ambos puntales de la argumentación a la Searle, considera que los qualia no existen (al menos no han llegado a caracterizarse de forma objetiva) y que el punto de vista de Searle yerra al no reconocer que programar es lo que permite desarrollar una mente (Dennett 1987: 325–6). Dennett, con todo, considera poco probable que los robots hayan llegado a estar programados del mismo modo que los humanos, e incluso sospecha que el soporte físico de la mente humana puede resultar crucial para sus posibilidades de programación.

Desde el monismo anómalo propuesto por Davidson (1970/1980, 1973/1980, 1974/1980, Moya 2009, capítulo 6: 117-132), se ofrecen igualmente claves interesantes con las que poder dar cuenta de la especificidad de la mente humana en general y del conocimiento humano en particular, pues permite combinar la explicación causal de la conducta humana con la consideración de los seres humanos como seres racionales, libres y responsables de sus acciones. El principal argumento de Davidson en favor del anomalismo de lo mental apunta a que la adscripción de propiedades mentales y la adscripción de propiedades físicas se encuentran regidas por principios constitutivos diferentes. La adscripción de predicados mentales estaría regida por el principio constitutivo de la racionalidad, mientras que la adscripción de predicados físicos se halla gobernada por

el principio constitutivo de la causalidad. Al atribuir predicados mentales a un sujeto hemos de presuponer en él un alto grado de coherencia y racionalidad en lo que atañe a su conjunto de creencias, deseos, decisiones y acciones. La conclusión que extrae Davidson es que, debido a esa diferencia de principios constitutivos, de compromisos en el uso de los conceptos mentales y físicos, no puede haber conexiones sistemáticas entre las propiedades mentales y las propiedades físicas, por lo que no podemos esperar hallar leyes psicofísicas. Las propiedades mentales son aquello que atribuimos a un sujeto para hacernos inteligible su comportamiento. En cuanto a la imposibilidad de la existencia de leyes psicológicas estrictas, Davidson arguye que, para que hubiera tales leyes, sería necesario que lo mental fuese un sistema causalmente cerrado, como lo es el mundo físico. Que un sistema es causalmente cerrado significa que no necesitamos salir del sistema para explicar cualquier cambio que tiene lugar en él. Pero el ámbito de lo mental no es un sistema causalmente cerrado. El monismo anómalo puede reconciliar así los dos principios que Kant consideraba irrenunciables: la necesidad natural y la libertad (Moya, 2009: capítulo 6).

IV. Responsabilidad epistémica

Lo que da sentido a la tesis de que tenemos responsabilidad epistémica y, por tanto, de que existe una ética de la creencia (de la formación de creencias), es la idea de que la formación de creencias está sujeta, al menos en cierta medida, a nuestro control y a nuestra voluntad (entendida como la facultad para decidir y ordenar nuestra propia conducta conforme a lo que consideramos deseable). En el presente contexto, la voluntad se entiende como libre albedrío, esto es, como la elección de algo sin precepto o impulso externo que a ello obligue.

El debate sobre la ética de la creencia se inicia con el ensayo «The Ethics of Belief», publicado en 1877 por el matemático y filósofo de Cambridge William Kingdon Clifford. Al comienzo del ensayo, Clifford defiende el riguroso principio de que todos estamos obligados a tener siempre pruebas suficientes de cada una de nuestras creencias. De hecho, las primeras secciones de «The Ethics of Belief» son tan severas que acabaron provocando una encendida crítica por parte de William James («The Will to Believe» 1896). El ruido de las críticas, lamentablemente, ha llegado a eclipsar algunas de las ideas más brillantes de Clifford. Entre ellas, su tesis de

que tenemos obligaciones epistémicas, derivadas de nuestra capacidad para investigar, que son a su vez morales. Podemos escoger si nos permitimos o no creer basándonos en evidencias insuficientes. Si nos lo permitimos, las consecuencias son, a juicio de Clifford, catastróficas, pues las creencias son lo que guía nuestra acción:

> No real belief, however trifling and fragmentary it may seem, is ever truly insignificant; it prepares us to receive more of its like, confirms those which resembled it before, and weakens others; and so gradually it lays a stealthy train in our inmost thoughts, which may someday explode into overt action, and leave its stamp upon our character forever (Clifford, 1876-77: 292).

Nuestro mal comportamiento epistémico, además, puede «contagiarse» a otros, privándoles en el futuro de la posibilidad de conocer:

> The harm which is done by credulity in a man is not confined to the fostering of a credulous character in others, and consequent support of false beliefs. Habitual want of care about what I believe leads to habitual want of care in others about the truth of what is told to me. Men speak the truth of one another when each reveres the truth in his own mind and in the other's mind; but how shall my friend revere the truth in my mind when I myself am careless about it, when I believe thing because I want to believe them, and because they are comforting and pleasant? (Clifford, 1876-77: 294).

Nuestras facultades nos capacitan para ciertos logros cognitivos, vitalmente valiosos. De dichas capacidadess emergen ciertas obligaciones. Conocer la verdad, además de lo valioso en sí mismo que nos pueda parecer, es necesario para el éxito en la acción. Si nuestra conducta epistémica obstaculiza el conocimiento de la verdad, no sólo nos conducirá individual y colectivamente a la acción fallida, sino que instaurará, extenderá y perpetuará una conducta cognitiva oscurantista u obstruccionista con respecto a la verdad, cercenando ya no nuestro éxito en el descubrimiento la verdad, sino la preservación de las condiciones de posibilidad para el descubrimiento de la verdad (Clifford, 1876-77: 289-294). Una mala conducta epistémica, esto es, un proceder cognitivo irracional conduciría, en terminología de Sosa, a una falta de perspectiva epistémica. Nuestro grado de acceso reflexivo no es comparable al de los animales o las máquinas. Somos conscientes de cómo obtenemos conocimiento, de cómo podríamos hacerlo, así como del contexto en el que cobra importancia el conocimiento. Limitarnos a conocer como animales o como máquinas

supondría renunciar a nuestras capacidades específicas. No cabe duda de que parte de nuestras capacidades operan como las de otros seres vivos e, incluso, como las de una máquina. No obstante, el acceso reflexivo a nuestro propio conocer y a sus posibilidades, no resulta equiparable a ninguna capacidad identificable fuera de la esfera del conocimiento humano.

V. Conclusiones

Los humanos estamos dotados de razón reflexiva, conciencia, autoconocimiento, intencionalidad. Todo ello nos permite conocer, pero también nos obliga a asumir una responsabilidad de tipo epistémico a la vez que moral.

No deja de resultar chocante que, quienes con gran naturalidad niegan que las máquinas sean moralmente responsables, a la vez afirmen con contundencia que son sujetos de conocimiento (o sujetos inteligentes). Quienes adoptan esa posición no parecen en absoluto conscientes que la falta de responsabilidad moral, en este caso, presupone una falta de responsabilidad epistémica, y esta falta delata una carencia en la posibilidad de conocer reflexivamente y valorativamente, esto es, a, el sentido humano. ¿Podrían sujetos no humanos conocer en el sentido humano? Este interrogante podría reformularse del siguiente modo: ¿podrían las máquinas conocer en el sentido humano, pero no a la manera humana? Es posible, sin embargo, para conocer, en el sentido humano de «conocer», habría que compartir mucho de la manera humana de conocer. Hemos avanzado en la comprensión de nuestro concepto de conocimiento, pero todavía estamos muy lejos de resolver el enigma de la conciencia humana y, por tanto, de descifrarnos a nosotros mismos.

Referencias bibliográficas

Allen, C. (2017). «On (not) defining cognition». *Synthese*, 194 (11): pp. 4233-4249.

Block, N. (1995). «On a confusion about a function of consciousness». *Behavioral and Brain Sciences*. 18(2): pp. 227-247. doi:10.1017/S0140525X00038188

Chignell, A. [en línea]: «The Ethics of Belief», *The Stanford Encyclopedia of Philosophy* (Spring, 2018 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/spr2018/entries/ethics-belief/ [Consultado: 18/03/2024].

Clifford, W. K. (1877). «The Ethics of Belief», *Contemporary Review*, 29, mayo: pp. 289-309.

Davidson, D. (1970). «Mental Events», en D. Davidson (1980) *Essays on Actions and Events*, Oxford: Clarendon Press: pp. 207-228.

- (1973). «The Material Mind», en D. Davidson (1980) *Essays on Actions and Events*, Oxford: Clarendon Press: pp. 229-239.

- (1974). «Psychology as Philosophy», en D. Davidson (1980) *Essays on Actions and Events*, Oxford: Clarendon Press: pp. 245-259.

- (1980). *Essays on Actions and Events*, Oxford: Clarendon Press.

Dennett, Daniel C. (1987), «Fast Thinking», in *The Intentional Stance*, Cambridge, MA: MIT Press, 324–337.

- (1988). «Quining qualia», en Anthony J. Marcel & E. Bisiach (eds.), *Consciousness in Contemporary Science*. Oxford: Oxford University Press.

- (1991). *Consciousness Explained*. Boston: Little, Brown and Co.

Dougherty, T. (2014). «The ethics of belief is ethics (period): reassigning responsibilism», en Matheson, J. y Vitz, R., (2014), *The ethics of belief: individual and social*, New York: Oxford University Press: pp. 146–168.

Grimaltos, T., & Iranzo, V. «El debate externismo/internismo en la justificación epistémica», en Quesada, Daniel (comp.), (2009). *Cuestiones de Teoría del Conocimiento*, Madrid: Tecnos: pp. 33-76.

Harrison, Peter (1992). «Descartes on animals». Philosophical Quarterly. 42 (167): pp. 219-227.

Huxley, T. H. (1874) «On the Hypothesis that Animals are Automata, and its History», *Nature* 10, Septiembre 3: pp. 362–366 (1874).

Lycan, W. G. (1987). *Consciousness*. Cambridge, MA: The MIT Press.

- (1996). *Consciousness and Experience*. Cambridge, MA: The MIT Press.

Madigan, T. J. (2009). *W. K. Clifford and* «*The Ethics of Belief*», Newcastle: Cambridge Scholars Publishing.

Matheson, J., & Vitz, R. (2014). *The ethics of belief: individual and social*, New York: Oxford University Press.

Moya, C. (2006). *Filosofía de la mente, Valencia: Universidad de Valencia.*

Nagel, T. (1974). «What Is It Like to Be a Bat? ». *The Philosophical Review*. 83 (4): pp. 435–450.

Newen, A. (2017). «What are cognitive processes? An example-based approach». *Synthese, 194 (11): pp. 4251-4268.*

Platón (1988). *Diálogos. IV República*, [introducción, traducción y notas de Conrado Eggers Lan], Madrid: Editorial Gredos S.A.

Platón (1988). *Diálogos. V Parménides. Teeteto. Sofista. Político.* [introducciones, traducciones y notas de Mª Isabel Santa Cruz, Álvaro Vallejo Campos, Néstor Luis Cordero], Madrid: Editorial Gredos S.A.

Searle, J. (1980). «Minds, Brains and Programs», *Behavioral and Brain Sciences*, 3: pp. 417–57.

- (1983/2002). «Can Computers Think?», en David John Chalmers (ed.), (2002) *Philosophy of Mind: Classical and Contemporary Readings*, New York: Oxford University Press USA: pp. 669-675.

- (1999). «The Chinese Room», en R.A. Wilson and F. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press: pp. 115-116.

Sosa, E. (1991). *Knowledge in Perspective*, Cambridge: Cambridge University Press.

- (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge* (Volume I), New York: Oxford University Press.

Tononi, G. (2008). «Consciousness as integrated information: a provisional manifesto». *Biological Bulletin 215:* pp. 216–42.

Turing, A. (1950). «Computing Machinery and Intelligence», *Mind*, 59: pp. 433–460.

María Caamaño Alegre: Profesora Titular del Departamento de Filosofía de la Universidad de Valladolid.

Líneas de investigación:

– Filosofía y metodología de la ciencia, evaluación de teorías, validez experimental y peculiaridades metodológicas de las ciencias sociales.

Publicaciones recientes:

– «From Ontological Traits to Validity Challenges in Social Science: The Cases of Economic Experiments and Research Questionnaires»(en

coautoría con José Caamaño Alegre), *International Studies in the Philosophy of Science, vol. 32*, n. 2, 2019, 101-127.

– «On Glasses Half Full or Half Empty: Understanding Framing Effects in Terms of Default Implicatures», *Synthese* (2021). https://doi.org/10.1007/s11229-021-03282-6

– «The Role of Presuppositions and Default Implicatures in Framing Effects», en Tadeusz Ciecierski & Paweł Grabarczyk (eds.), Context Dependence in Language, Action, and Cognition, De Gruyter's Epistemic Studies. Philosophy of Science, Cognition and Mind series, Berlin, 2021, DOI: https://doi.org/10.1515/9783110702286-011, 181-208.

– «Efectos marco, implicaturas por defecto y racionalidad», *SCIO. Revista de Filosofía*, n.º 22, julio de 2022, 127-156. https://doi.org/10.46583/scio_2022.22.1052

– «Empirical underdetermination: A bigger problem for the social sciences?» (forthcoming in the Proceedings of the 17th International Congress of Logic, Methodology and Philosophy of Science and Technology, College Publications, en prensa).

– «La revuelta historicista frente a Popper», Herencia y actualidad de Popper, Publicacions de la Universitat de València – PUV (en prensa).

Correo-e: mariaconcepcion.caamano@uva.es