El estatus moral de las entidades de inteligencia artificial

JOAN LLORCA ALBAREDA

UNO DE LOS GRANDES TEMAS FILOSÓFICO-MORALES de la segunda mitad del siglo XX fue el concepto de estatus moral. Su aparición tuvo lugar en el seno de tres debates: el aborto, los derechos de los animales y el ecologismo (Hursthouse 2013). La extensión de los derechos de las mujeres, su entrada en el mercado laboral y los avances tecnológicos en el campo de la medicina fueron los causantes de que se comenzara a cuestionar el rechazo político, legal y moral del aborto. La ética animal y la ética ecológica también plantearon desafíos importantes. Los animales han sido entendidos históricamente como propiedades e instrumentos de los seres humanos. No obstante, sufren y pueden ser dañados (Campos & Lara 2015; Singer 2011). La estructura y funcionamiento del mercado mundial conlleva una enorme cantidad de sufrimiento para muchos animales, por ejemplo, para aquéllos que son sacrificados para el consumo de carne o para aquéllos dañados por la contaminación de los espacios naturales. Esto último también invita a repensar los daños producidos a los ecosistemas (Callicott 1980; Leopold 2000). La diversidad de especies decrece con gran rapidez, así como muchos espacios naturales de gran antigüedad. Tampoco parece que los efectos de la acción humana sobre los ecosistemas sean moralmente neutros.

Lo que estas dos últimas reivindicaciones tienen en común es una innovación de enorme importancia para la filosofía moral: el ser humano ya no es el único foco de respeto y consideración moral, pues encontramos otro tipo de entidades que también merecen respeto y consideración *por sí mismas* (Kamm 2008). Los animales y ecosistemas habían jugado ya previamente un papel moral, pero de forma derivada. Immanuel Kant (2005) expresó esta idea a través de la noción de los *deberes indirectos*: los animales deben ser moralmente considerados porque, de no ser así, la calidad moral de las acciones humanas se

J. Llorca Albareda (✉)
Universidad de Granada, España
e-mail: joanllorca@ugr.es

Disputatio. Philosophical Research Bulletin
Vol. 12, No. 24, Mar. 2023, pp. 241–249
ISSN: 2254-0601 | [SP] | **ARTÍCULO**

vería gravemente perjudicada. Toda consideración moral hacia otro tipo de entidades dependería de que los seres humanos se vieran afectados de un modo u otro. El concepto de estatus moral, por el contrario, remarca que entidades distintas a los seres humanos poseen propiedades, características y capacidades que las hacen merecedoras de respeto y consideración moral. La idea de que el ser humano es el único que posee propiedades que cualifican para la consideración moral depende íntegramente de un prejuicio antropocéntrico (Singer 2011). El ser humano es merecedor de respeto no por pertenecer a la especie humana, sino por una o varias propiedades, características o capacidades que lo hacen merecedor de tal cosa.

La historia del concepto de estatus moral ha consistido en la búsqueda de la propiedad fundamental de la consideración moral, ya sea ésta la (auto)consciencia, la sintiencia o la racionalidad. Se abandona la vinculación exclusiva entre el ser humano y estas propiedades y se pregunta qué otras entidades pueden poseer estas propiedades y en qué grado (DeGrazia 2008). Tras décadas de intenso debate en bioética, ética animal y ecoética parecía que el círculo de consideración moral se había ampliado tanto, su inclusividad era tal, que no se había dejado ninguna entidad moralmente relevante fuera de él (Singer 1983). No obstante, la emergencia de la inteligencia artificial (IA) está planteando objeciones de gran calado a estas visiones optimistas. Por un lado, las tecnologías de la información y la comunicación están transformando profundamente nuestra realidad. Los datos y flujos de información plantean desafíos éticos enormemente importantes en términos de privacidad y seguridad; además de que, como es habitual en la historia del concepto de estatus moral, se comienza a examinar si estas unidades de información merecen ser moralmente consideradas por sí mismas (Floridi 2013). Por otro lado, los nuevos desarrollos en IA, derivados principalmente de los avances en aprendizaje automático, redes neuronales artificiales y algoritmos genéticos, están permitiendo vislumbrar un futuro en el que las entidades artificiales sean capaces de formas cognitivamente complejas de deliberación moral y trato interpersonal. Hasta el momento la IA había resultado exitosa a nivel local, lo que se denomina IA débil o estrecha, superando a la inteligencia humana en la realización de ciertas actividades concretas (Meseguer & López de Mántaras 2017). Las críticas más relevantes a las expectativas desmedidas de la IA se habían fundado precisamente en esta idea: la IA funciona excelentemente a nivel parcial, en ámbitos predefinidos y bien demarcados, pero es incapaz de una inteligencia general similar a la de los seres humanos (Dreyfus 1979; Russell & Norvig 2010; Searle 1980; Véliz 2021). Los nuevos avances en este tipo de tecnologías están alterando radicalmente estos presupuestos (Bostrom

2014).

Si son posibles las entidades artificiales que poseen determinadas propiedades susceptibles de ser consideradas moralmente, ¿puede esperarse que la IA entre dentro de los márgenes del estatus moral? El interrogante por el estatus moral de la IA plantea un nuevo desafío para los modos tradicionales de entender la ética (Gunkel 2012). Las propiedades definitorias de la consideración moral, presuntamente imparciales tras las críticas del animalismo y el ecologismo, deben ser repensadas en el marco de nuevos tipos de entidades. Y no pueden ser rechazados, al igual que se hacía sobre la base de la especie humana, por su sustrato inorgánico, so pena de incurrir en discriminaciones, sino que debe mostrarse si poseen o carecen de estas propiedades (Bostrom & Yudkowsky 2018). La complejidad del análisis hace que las investigaciones llevadas a cabo hasta el momento hayan adoptado tres tipos de aproximaciones: ontológica, epistemológica y político-legal.

La *aproximación ontológica* refiere al cuestionamiento filosófico de las propiedades fundamentales del estatus moral. La IA puede cumplir muchas de las propiedades que tradicionalmente se han identificado como fundamentales, por lo que resulta imprescindible su revisión crítica. Esto ha llevado a dos estrategias argumentativas distintas. En primer lugar, muchos autores vuelven a la pregunta clásica por la propiedad central del estatus moral. Estas discusiones han planteado características como la consciencia (Mosakas 2021), la deliberación moral (Himma 2009) o la vida interna (Nyholm 2020). Las discusiones retornan, de este modo, a la metodología seguida por la ética animal y la ética ecológica: cuestionar aquello que nos hace considerar moralmente a una entidad, mostrar sus inconsistencias y proponer una nueva propiedad más ajustada a la realidad moral. En segundo lugar, un grupo reducido de autores pone en tela de juicio que la forma en la que discurre la vida moral consista en identificar en un primer momento un grupo de propiedades en una serie de entidades y posteriormente considerar moralmente a estas entidades de acuerdo con este análisis. La vida moral, por el contrario, se basa en las relaciones que tenemos con diferentes tipos de entidades y las obligaciones morales para con ellas sólo son posibles en el marco de estas relaciones (Alonso 2023; Coeckelbergh 2014; Gunkel 2012). El estatus moral emerge de las relaciones entre las entidades y no está predefinido en la naturaleza individual de cada una de ellas. A saber, la consideración moral fracasa cuando se funda en la ontología, en la búsqueda de una serie de propiedades fundamentales, por lo que se requiere de un enfoque relacional que complejice el modo en que comprendemos la filosofía moral (Gunkel 2018; Llorca-Albareda & Díaz-Cobacho 2023).

La *aproximación epistemológica* se pregunta por el conocimiento que tenemos y que podemos tener de las propiedades definitorias del estatus moral. Su significado es en muchas ocasiones inasible y solemos tener un difícil acceso a las mismas, por lo que resulta imprescindible clarificar qué tipo de conocimiento podemos tener de ellas y si este conocimiento cambia el modo en que entendemos las propiedades. Desde este enfoque se han seguido dos tipos de estrategias. En primer lugar, la dificultad de acceder a propiedades como la consciencia o la racionalidad exige algún tipo de método que nos permita diferenciar aquellos seres que poseen la propiedad de aquéllos que no. El primero en tratar de abordar la cuestión fue Alan Turing (1950) a través del llamado test de Turing o *Imitation Game*. Esta propuesta diferenciaba entre seres inteligentes y seres no inteligentes a través de entrevistas anónimas en las que se medía la inteligencia mediante la calidad de las respuestas de los entrevistados. La anonimidad de los participantes eliminaba los potenciales prejuicios del entrevistador. La IA y el ser humano podían poseer la propiedad en cuestión en el mismo grado siempre y cuando respondieran lo mismo. En la actualidad se han propuesto otros métodos alternativos que miden la misma o diferentes propiedades (Sparrow 2004). En segundo lugar, se ha argumentado que la dificultad en el acceso a la vida interna de los seres humanos tiene consecuencias muy significativas para el modo en que debemos entender las propiedades definitorias del estatus moral. Este es el *problema de las otras mentes*, el cual refiere a la imposibilidad que cada individuo tiene para acceder a las mentes del resto de individuos que lo rodean. Tan sólo podemos acceder a nuestras propias mentes. John Danaher (2020) ha llevado este problema hasta sus últimas consecuencias, defendiendo una concepción behaviorista del estatus moral; esto es, ante la imposibilidad de saber a ciencia cierta si una entidad posee o no la propiedad moral fundamental, debemos determinar su estatus moral en función de si se comporta del mismo modo en que lo haría un individuo que poseyera dicha propiedad. Otra línea importante de propuestas ha defendido el argumento precautorio, aduciendo que, debido a que no podemos saber si se posee esa propiedad, el peligro de no considerar moralmente a una entidad que cualifica para ser considerada moralmente debe invitar a ser lo más inclusivos posible (Neely 2014).

La *aproximación político-legal* trata de pensar el problema de las obligaciones para con la IA en términos de las dinámicas de reconocimiento de las estructuras político-legales. Este enfoque se aleja de los análisis puramente normativos y se adapta a los modos de proceder legal y políticamente en contexto concretos. Se han seguido, al igual que en las dos anteriores, dos formas argumentativas. En primer lugar, se han desarrollado numerosos

argumentos por analogía, principalmente mediante dos figuras legales: las corporaciones y los derechos de los animales. Las corporaciones, por un lado, han mostrado que la personeidad legal de entidades artificiales es posible siempre y cuando se produzca una separación clara entre la responsabilidad civil y la responsabilidad penal (Laukyte 2017). Los derechos de los animales muestran, por otro lado, que el molde jurídico del derecho romano permite dar cuenta de entidades intermedias que no son ni personas legales plenas ni meros objetos sujetos únicamente a las leyes de propiedad (Gunkel 2018). En segundo lugar, se han llevado a cabo análisis que analizan teóricamente las bases de la personeidad político-legal y analizan los argumentos a favor y en contra de la inclusión de la IA en las estructuras político-legales en el marco de las tradiciones de pensamiento jurídico-político (Calverley 2008).

El progresivo desarrollo de la IA y los debates filosóficos que suscita hacen que sea sencillo anticipar la importancia que cobrará en un futuro cercano este tipo de discusiones. El estatus moral de la IA se erige como uno de los debates más atractivos que se desprenden de estas nuevas tecnologías, además de que la variedad de ámbitos en los que éstas se insertan —por ejemplo, el sector biomédico, el sector financiero o el sector educativo— y los soportes que utilizan —por ejemplo, las aplicaciones móviles, los cuerpos robóticos o los bots virtuales— requieren de una labor de especificación y aplicación que exigirá respuestas contextuales y específicas en los debates sobre el estatus moral. Cuestiones tales como la consideración moral de los robots sexuales (Levy 2007) o el papel moral jugado por los robots asistenciales con los que se establecen relaciones duraderas (Sorell & Draper 2014) dan cuenta de la actualidad y relevancia de estas discusiones.

REFERÊNCIAS

ALONSO, Marcos (2023). «Can Robots have Personal Identity?». *International Journal of Social Robotics* 15, pp. 1-10. DOI: https://doi.org/10.1007/s12369-022-00958-y

BOSTROM, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

BOSTROM, Nick y YUDKOWSKY, Eliezer (2018). «The ethics of artificial intelligence». En: *Artificial intelligence safety and security*, editado por Roman V. Yampolskiy. Oxfordshire: Routledge, pp. 57-69.

CALLICOTT, John Baird (1980). «Animal liberation: A triangular affair». *Environmental Ethics* 2, no. 4: pp. 311-338. DOI: https://doi.org/10.5840/enviroethics19802424

CAMPOS, Olga y LARA, Francisco (2015). *Sufre, luego importa. Reflexiones éticas sobre los animales*. Madrid: Plaza y Valdés.

CALVERLEY, David (2008). «Imagining a non-biological machine as a legal person». *AI & Society* 22, no. 4: pp. 523-537. DOI: https://doi.org/10.1007/s00146-007-0092-7

COECKELBERGH, Mark (2014). «The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics». *Philosophy & Technology* 27, no. 1: pp. 61-77. DOI: https://doi.org/10.1007/s13347-013-0133-8

DANAHER, John (2020). «Welcoming robots into the moral circle: a defence of ethical behaviourism». *Science and Engineering Ethics* 26, no. 4: pp. 2023-2049. DOI: https://doi.org/10.1007/s11948-019-00119-x

DEGRAZIA, David (2008). «Moral status as a matter of degree?». *The Southern Journal of Philosophy* 46, no. 2: pp. 181-198. DOI: https://doi.org/10.1111/j.2041-6962.2008.tb00075.x

DREYFUS, Hubert (1979). *What computers can´t do: the limits of artificial intelligence*. Cambridge: MIT Press.

FLORIDI, Luciano (2013). *The Ethics of Information*. Oxford: Oxford University Press.

GUNKEL, David (2012). *The machine question: critical perspectives on AI, robots, and ethics*. Cambridge: MIT Press.

GUNKEL, David (2018). *Robot rights*. Cambridge: MIT Press.

HIMMA, Kenneth Einar (2009). «Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?». *Ethics and Information Technology* 11, no. 1: pp. 19-29. DOI:

https://doi.org/10.1007/s10676-008-9167-5

HURSTHOUSE, Rosalind (2013). «Moral status». En: *International Encyclopedia of Ethics*, editado por Hugh LaFollette. John Wiley & Sons. Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/9781444367072.wbiee076

KAMM, Frances Myrna (2008). *Intricate ethics: rights, responsibilities, and permissible harm*. Oxford: Oxford University Press.

KANT, Immanuel (2005). *La metafísica de las costumbres*. Madrid: Tecnos.

LAUKYTE, Migle (2017). «Artificial agents among us: Should we recognize them as agents proper?». *Ethics and Information Technology* 19, no. 1: pp. 1-17. DOI: https://doi.org/10.1007/s10676-016-9411-3

LEOPOLD, Aldo (2000). *Una ética de la tierra*. Madrid: Los Libros de la Catarata.

LEVY, David (2007). *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper Perennial.

LLORCA-ALBAREDA, Joan y DÍAZ-COBACHO, Gonzalo (2023). «Contesting the Consciousness Criterion: A More Radical Approach to the Moral Status of Non-Humans». *AJOB Neuroscience* 14, no. 2: pp. 158-160. DOI: https://doi.org/10.1080/21507740.2023.188280

MESEGUER, Pedro & LÓPEZ DE MÁNTARAS, Ramón (2017). *Inteligencia artificial*. Madrid: Los Libros de la Catarata.

MOSAKAS, Kestusis (2021). «On the moral status of social robots: considering the consciousness criterion». *AI & Society* 36, no. 2: pp. 429-443. DOI: https://doi.org/10.1007/s00146-020-01002-1

NEELY, Erica (2014). «Machines and the moral community». *Philosophy & Technology* 27, no. 1: pp. 97-111. DOI: https://doi.org/10.1007/s13347-013-0114-y

NYHOLM, Sven (2020). *Humans and robots: ethics, agency, and anthropomorphism*. Londres: Rowman & Littlefield Publishers.

RUSSELL, Stuart y NORVIG, Peter (2010). *Artificial intelligence: a modern approach*. Upper Saddle River: Pearson Education.

SEARLE, John (1980). «Minds, brains, and programs». *Behavioral and Brain Sciences* 3, no. 3: pp. 417-424. DOI: https://doi.org/10.1017/S0140525X00005756

SINGER, Peter (1983). *The Expanding Circle: Ethics and Sociobiology*. Oxford: oxford University Press.

SINGER, Peter (2011). *Liberación animal*. Madrid: Santillana.

SORELL, Tom y DRAPER, Heather (2014). «Robot carers, ethics, and older

people». *Ethics and Information Technology* 16, no. 3: pp. 183-195. DOI: https://doi.org/10.1007/s10676-014-9344-7

SPARROW, Robert (2004). «The Turing Triage Test». *Ethics and Information Technology* 6, no. 4: pp. 203-213. DOI: https://doi.org/10.1007/s10676-004-6491-2

VÉLIZ, Carissa (2021). «Moral zombies: why algorithms are not moral agents». *AI & Society* 36: pp. 487-497. DOI: https://doi.org/10.1007/s00146-021-01189-x

TURING, Alan Mathison (1950). «Computing machinery and intelligence». *Mind* 236, no. 59: pp. 433-460. DOI: https://doi.org/10.1093/mind/LIX.236.433.

El estatus moral de las entidades de inteligencia artificial

El debate sobre el estatus moral de la inteligencia artificial (IA) está ganando cada vez más relevancia en los contextos sociales y académicos. Los funcionamientos internos de las nuevas tecnologías, así como los tipos de respuestas que despiertan en los usuarios humanos, están desafiando concepciones éticas ampliamente arraigadas en el imaginario común. Tras las discusiones en bioética, ética animal y ecoética, vuelve a tomar importancia el concepto de estatus moral en un nuevo sentido: entidades artificiales de reciente surgimiento aparecen como potenciales candidatas para ser moralmente consideradas en virtud de una posible posesión de la(s) propiedad(es) fundamental(es) del estatus moral. Es por ello por lo que se requiere de análisis que aborden teóricamente este fenómeno y den cuenta de su estructura y contenidos principales. Este artículo asume esta tarea y articula las líneas fundamentales del debate contemporáneo acerca del estatus moral de la IA. Con este propósito, el artículo conecta las discusiones contemporáneas con los orígenes conceptuales del estatus moral al mismo tiempo que propone como matriz explicativa tres aproximaciones desde las que se está desarrollando el debate en la literatura académica.

Palabras Clave: Consideración moral · Ética aplicada · Ética de la inteligencia artificial· Propiedad.

The moral status of artificial intelligence entities

The debate on the moral status of artificial intelligence (AI) is gaining increasing relevance in social and academic contexts. The inner workings of new technologies, as well as the types of responses they elicit in human users, are challenging widely rooted ethical conceptions in the social imaginary. After its predominant role in bioethics, animal ethics, and ecoethics, the concept of moral status status is again gaining importance in a different sense: newly emerging artificial entities appear as potential candidates for moral consideration by virtue of a possible possession of the fundamental properties of moral status. Therefore, theoretical analyses are required to address the phenomenon and account for its structure and main contents. This article takes up this task and articulates the fundamental lines of the contemporary debate about the moral status of AI. To this end, the article connects contemporary discussions with the conceptual origins of moral status while proposing as an explonatory matrix three approaches from which the debate is being addressed in the academic literature.

Keywords: Moral consideration · Applied ethics· Ethics of artificial intelligence · Property.

JOAN LLORCA ALBAREDA es candidato predoctoral en la Universidad de Granada y contratado con cargo al proyecto SocrAI+ (Mejora Moral e Inteligencia Artificial. Aspectos Éticos de un Asistente Virtual Socrático/ Ref: B-HUM-64-UGR20). Su tesis doctoral versa sobre el estatus moral de las nuevas entidades de inteligencia artificial, aunque también dedica su investigación a otras temáticas de filosofía y ética de la tecnología. Éstas son algunas de sus últimas publicaciones: Llorca-Albareda, Joan & Díaz-Cobacho, Gonzalo (2023). «Contesting the Consciousness Criterion: A More Radical Approach to the Moral Status of Non-Humans». *AJOB Neuroscience* 14, no. 2: pp. 158-160; Llorca Albareda, Joan & Rueda, Jon (2023) «Divide and Rule? Why Ethical Proliferation is not so Wrong for Technology Ethics». *Philosophy & Technology* 36, 10. **Contacto**: Departamento de Filosofía I, Facultad de Filosofía y Letras, Universidad de Granada, Calle del Profesor Clavera s/n, Campus de Cartuja, Granada, España. e–mail (✉): joanllorca@ugr.es · iD: http://orcid.org/0000-0003-4889-3859.

HISTORIA DEL ARTÍCULO | ARTICLE HISTORY
Recibido/Received: *26–October–2022*; Aceptado/Accepted: *07–March–2023*; Published Online: *31–March–2023*

COMO CITAR ESTE ARTÍCULO | HOW TO CITE THIS ARTICLE
Llorca Albareda, Joan (2023). «El estatus moral de las entidades de inteligencia artificial». *Disputatio. Philosophical Research Bulletin* 12, no. 24: pp. 241–249.