Daimon

Revista Internacional de Filosofía

Número 90. Septiembre-Diciembre 2023

INTELIGENCIA ARTIFICIAL

Sección 1ª: Inteligencia artificial, datos y objetividad. ¿El regreso del naturalismo dataísta?

Editores:

Ariel Guersenzvaig y David Casacuberta

Sección 2ª: Ética aplicada para una Inteligencia artificial confiable

Editores:

Elsa González-Esteban y Domingo García Marzá

UNIVERSIDAD DE MURCIA
DEPARTAMENTO DE FILOSOFÍA

Daimon. ***Revista Internacional de Filosofía*****,** fundada en 1989, es una publicación cuatrimestral del Departamento de Filosofía de la Universidad de Murcia (España). Desde entonces, *Daimon* ha abierto un espacio filosófico de reflexión, análisis y crítica de problemas referidos principalmente al ser humano, en todas las dimensiones de su existencia.

Tiene como objetivo la publicación de investigaciones originales: publica por lo tanto trabajos que abordan, desde una perspectiva filosófica, las múltiples dimensiones o esferas de la existencia humana. *Daimon* es, pues, una revista que se dirige a investigadores, pero también a todo el que se interesa por el pensamiento filosófico en sentido amplio, desde la frontera de la ciencia hasta la de la literatura. En las páginas de *Daimon,* el especialista puede encontrar nuevos enfoques de un determinado problema o autor; el investigador, un espacio en el que publicar, contrastar o confirmar sus ideas; y el lector aficionado, artículos, revisiones críticas y reseñas de libros que pueden alimentar su curiosidad y ampliar su formación.

Daimon combina, en fin, el rigor académico con la originalidad de las investigaciones, sin olvidar la apertura y pluralidad necesarias en una publicación filosófica que quiere interesar también al lector ilustrado en general.

Daimon figura en el *European Reference Index for the Humanities* (ERIH) en la categoría ERIH PLUS *(Philosophy, 2016-01-11)*; sus artículos son registrados en las bases filosóficas de datos nacionales e internacionales siguientes: *Base ISOC - Filosofía. CINDOC* (España); *Dialnet* (España); *Francis, Philosophie. INIST. CNRS* (France); *Philosopher´s Index* (Bowling Green, OH, USA); *Repertoire Bibliographique de Philosophie* (Louvain, Belgique); *Ulrich´s Internacional Periodicals Directory* (New York, USA), *Scopus* (Editora Elsevier, Ámsterdam, Holanda), *Web of Science* (Clarivate Analitycs, Estados Unidos).

Daimon obtuvo por vez primera en 2014 el Certificado de Revista Excelente de la Fundación Española para la Ciencia y la Tecnología (FECYT, http://calidadrevistas.fecyt.es/Paginas/Home.aspx).

Edición electrónica: *www.um.es/daimon*

Daimon. *Revista Internacional de Filosofía*

Daimon. Revista Internacional de Filosofía

Publicación cuatrimestral. Número 90. Septiembre-Diciembre 2023

Monográfico sobre
Inteligencia Artificial

Sección 1ª: Inteligencia artificial, datos y objetividad. ¿El regreso del naturalismo dataísta?

Editores:
Ariel Guersenzvaig y David Casacuberta

Sección 2ª: Ética aplicada para una Inteligencia artificial confiable

Editores:
Elsa González-Esteban y Domingo García Marzá

UNIVERSIDAD DE MURCIA
DEPARTAMENTO DE FILOSOFÍA

Daimon. Revista Internacional de Filosofía

Publicación cuatrimestral. Número 90. Septiembre-Diciembre 2023

Administración: *Daimon* es una revista cuatrimestral, editada y distribuida por el Servicio de Publicaciones de la Universidad de Murcia. Apartado 4021. 30080 Murcia (España). Tfno.: 868883012. Fax: 868883414.

Redacción e intercambios: ver *Normas de publicación*, al final de la revista.

ISSN de la edición en papel: 1130-0507.
ISSN de la edición digital (disponible en http://revistas.um.es/daimon): 1989-4651.
Depósito legal: V 2459-1989.

Maquetación, diseño de cubierta: Compobell, S.L. Murcia.

Daimon. Revista Internacional de Filosofía
Publicación cuatrimestral. Número 90. Septiembre-Diciembre 2023

Monográfico sobre Inteligencia Artificial

Artículos

Sección 1ª: Inteligencia artificial, datos y objetividad. ¿El regreso del naturalismo dataísta?
Editores: Ariel Guersenzvaig y David Casacuberta

Sección 2ª: Ética aplicada para una Inteligencia artificial confiable
Editores: Elsa González-Esteban y Domingo García Marzá

Reseñas

SECCIÓN 1

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 7-12
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.576781

Presentación de la sección sobre Inteligencia artificial, datos y objetividad. ¿El regreso del naturalismo dataísta?

Presentation of the section on Artificial Intelligence, Data, and Objectivity. The return of dataist naturalism?

ARIEL GUERSENZVAIG[*]
DAVID CASACUBERTA[**]

Es función de la filosofía analizar los problemas desde una perspectiva amplia, lo más genérica posible y transdisciplinar, versus a un acercamiento más detallado y específico de las ciencias experimentales. Sin embargo, esa generalidad puede hacernos perder de vista problemas importantes que se mueven en marcos más específicos y, al llevarse a un extremo, acabar produciendo un discurso filosófico vago y sin concreción.

Encontramos actualmente esta situación en buena parte del discurso divulgativo filosófico que se está generando en relación al impacto de la inteligencia artificial (IA) en la sociedad. En este, no se distingue entre las diferentes técnicas y metodologías usadas para desarrollar sistemas computacionales con IA, los diferentes campos de aplicación de esas tecnologías, ni los diferentes tipos de problemas éticos que esas tecnologías pueden presentar. Sin precisar los conceptos clave (ni siquiera la propia noción de «inteligencia» o qué significa «pensar»), se explora, por poner un ejemplo, la complejísima cuestión de si los

* Elisava Facultad de Diseño e Ingeniería de Barcelona, UVIC-UCC <aguersenzvaig@elisava.net>, Profesor Contratado Doctor. Sus principales áreas de investigación son, por un lado, el impacto ético de la inteligencia artificial en la sociedad y, por otro lado, la ética de la actividad profesional del diseño. Pertenece al Grupo de Investigación consolidado HIMTS (Human, Interaction, Materials, Technology, and Society). Es miembro del comité de ética de la investigación de la Universidad de Vic-UCC. Publicaciones recientes: Guersenzvaig, A., & Sangüesa, R. (2022). A critical reflection on the treatment of AI system's 'agency' in the (Spanish) media. *Avances en Interacción Humano-Computadora*, *1*(7), 1-4: Guersenzvaig, A. (2021). *The Goods of Design: Professional Ethics for Designers*. Rowman & Littlefield.

** Universidad Autónoma de Barcelona <david.casacuberta@uab.cat>. Profesor Contratado Doctor en el Departamento de Filosofía de la Universidad Autónoma de Barcelona. Su línea de investigación actual son los impactos sociales y cognitivos de las tecnologías digitales. Actualmente es miembro del Grupo de Trabajo de Ética, Seguridad y Regulación de bioinformática Barcelona, investigador del grupo consolidado GEHUCT (Grupo de Estudios Humanísticos en Ciencia y Tecnología). Publicaciones recientes: Casacuberta, D., Guersenzvaig, A., & Moyano-Fernández, C. (2022). Justificatory explanations in machine learning: for increased transparency through documenting how key concepts drive and underpin design and engineering decisions. *Ai & Society*, 1-15; Casacuberta, D., & Guersenzvaig, A. (2019). Using Dreyfus' legacy to understand justice in algorithm-based processes. AI & SOCIETY, 34(2), 313-319.

sistemas con IA «piensan mejor que nosotros». Se genera así un discurso que resulta ser, o bien de corte apocalíptico (como en el caso de los riesgos existenciales de la IA), o bien de un utilitarismo banal que no va más allá de la evaluación superficial de pros y contras (como cuando se reduce el debate en torno a la IA a la examinación de sus beneficios y perjuicios más directos y, de ser posible, de manera matemática).

Sea como fuere, en estos casos la discusión se centra en pseudoproblemas que no tienen ningún impacto en la actualidad y se dejan de lado problemas reales que necesitan *ya* de soluciones y respuestas filosóficas, quedando totalmente desapercibidos. Esta superficialidad la encontramos, por ejemplo, en la discusión filosófica en torno a los vehículos autónomos. La banalidad se hace patente cuando, de manera espuria, se utiliza como principal instrumento de exploración el experimento mental del *dilema del tranvía*, genialmente formulado por Philippa Foot (1967) en el marco de una reflexión mucho más amplia sobre la doctrina del Doble efecto y, más específicamente, sobre la diferencia entre actuar y dejar que algo ocurra. En el caso de los vehículos autónomos, tal como sucede en el famoso *Moral Machine Experiment* (Awad et al., 2018), se plantean escenarios dicotómicos y a veces implausibles (e.g., «¿El coche autónomo debe embestir y matar a un bebé, o a una persona sin hogar?»). Estos escenarios, en vez de permitir y propiciar una reflexión ética profunda como la desarrollada por Foot, se consideran literalmente y como destino filosófico final. Se eluden así cuestiones más relevantes, y también filosóficamente más ricas, como, por ejemplo, qué tipo de movilidad requiere una sociedad moderna y plural, cuál es la responsabilidad de los fabricantes de vehículos, qué rol deben tener el estado y el gobierno, cómo se navegan las tensiones entre los derechos y libertades de los automovilistas y los de otros usuarios de la vía pública, qué grupos se ven principalmente beneficiados o perjudicados por el uso de automóvil, o, de manera aún más abarcadora, qué modelos de ciudad son más conducentes al bienestar.

En paralelo, este uso espurio del *dilema del tranvía* demuestra una falta completa de imaginación y comprensión filosófica, interpretando de manera literal lo que en realidad es un experimento mental cuyo fin es mostrar la interacción y posible inconsistencia de algunas de nuestras intuiciones éticas, así como la centralidad de la intención de los agentes en la consideración de los daños que puedan ocasionar. Otra ingenuidad filosófica muy común en los escritos dataístas es cuando se habla de dejar la investigación científica o la discusión ética en manos de algoritmos.

En este número especial hemos intentado hacer una contribución para enmendar esta confusión y ayudar a elaborar un discurso filosófico más profundo y específico que trate problemas concretos en lugar de ofrecer un falso discurso totalizador que sea, en realidad, superficial y ambiguo. Por lo que hace a esta sección, pensamos que desde la filosofía no se ha hecho suficiente hincapié en la insostenibilidad de los presupuestos de la filosofía de la ciencia dataísta que cree que se acerca un fin del método científico, en el que algoritmos de aprendizaje automático, en base a grandes volúmenes de datos, podrán hacer predicciones útiles en todos los campos de la investigación científica, y en especial en aquellas ciencias aplicadas para el bienestar humano, como la biomedicina, sin necesidad de tener que construir hipótesis y teorías, produciendo así un conocimiento verdaderamente objetivo. En otras palabras, una ciencia *naturalista*, puramente determinada por hechos empíricos, y libre de valores y teorías previas.

Específicamente, entonces, queremos cuestionar el aparente naturalismo dataísta de estos sistemas algorítmicos y tratar su supuesta neutralidad matemática, presentada como garantía

de un recorrido riguroso e impersonal desde los datos al resultado. El enfoque cuantitativo y estadístico suele asociarse a la capacidad de ofrecer credibilidad y confianza. Durante los últimos dos siglos, las mediciones y análisis estadísticos que afectan todas las áreas de la vida pública y privada han permitido la creación y revisión de teorías y han vertebrado el debate público (Desrosières, 1993; Porter, 2020). Una frase como "esto está respaldado por datos" se convirtió en una frase común para legitimar afirmaciones y decisiones sobre pobreza, educación, empleo y prácticamente cualquier otro aspecto de la vida social. Esta mentalidad, conocida como «dataísmo» (Brooks, 2013), se vio acentuada por la amplia disponibilidad de las computadoras y las bases de datos, así como el surgimiento del *Big Data*, la utilización de grandes volúmenes de datos con fines computacionales.

Un influyente ensayo escrito por Chris Anderson *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* (Anderson, 2008) epitomiza la visión del dataísmo radical como ningún otro. Anderson argumenta que la disponibilidad de grandes volúmenes de datos (el llamado *Big Data*) y la inteligencia artificial harán que las teorías y las hipótesis *ex-ante* sean redundantes. Esta visión subvierte así los principios generales de la filosofía de la ciencia posterior a 1950 en la que el itinerario del descubrimiento comienza con la formulación de hipótesis tentativas (impulsadas por conjeturas fundamentadas en teorías y datos previos) que se validan empíricamente, generando así nuevos datos que sirven para aceptarlas, modificarlas o rechazarlas.

Con la siempre creciente adopción de sistemas computacionales para la automatización de todo tipo de ámbitos públicos y privados, los datos y los cálculos estadísticos van camino de consolidarse como el estándar de facto del conocimiento aplicado. A pesar de los serios problemas que genera su uso, mediante sistemas de IA, por ejemplo, se otorgan préstamos y créditos (Eticas Foundation, 2021), se gestionan presupuestos gubernamentales (Valle-Cruz et al., 2022), se decide quién debe ser investigado por posible fraude en la percepción de prestaciones por cuidado de menores (Hadwick & Lan, 2021), se gestiona y evalúan empleados (Tewari & Pant, 2020), se vigilan las fronteras (Sánchez-Monedero & Dencik, 2022), o se toman decisiones jurídico-penales (Casacuberta y Guersenzvaig, 2019; Morales Moreno, 2021). No resulta entonces exagerado sugerir que con la estadística evolucionando hacia la «ciencia de datos», la visión dataísta que domina buena parte del desarrollo tecnológico actual comienza, en la práctica, a gobernar el mundo.

La inteligencia artificial originalmente estuvo fundamentalmente vinculada a teorías cognitivistas y a un enfoque «simbólico». Sin embargo, la principal técnica de la inteligencia artificial actual es el llamado «aprendizaje automático» *(machine learning* en inglés*)*, que está basado de manera general en representaciones de conocimiento obtenidas mediante técnicas matemáticas y estadística aplicada en conjunción con el procesamiento computacional de enormes volúmenes de datos. El vínculo entre la IA y el enfoque simbólico sigue existiendo pero este se ha debilitado mucho debido a la preponderancia del aprendizaje automático que se enmarca dentro del enfoque denominado «conexionista».

Recientemente, la última evolución de los sistemas de este tipo como los generadores de texto e imágenes como ChatGPT o Stable Difussion (técnicamente basados en un modelo de aprendizaje profundo llamado «transformador») han cosechado amplia atención. También encontramos multitud de otros sistemas orientados a realizar predicciones, evaluaciones y clasificaciones en base a *Big Data*. Así, por ejemplo, mediante la utilización de imágenes

dermatológicas, un sistema de IA «aprende» a detectar melanomas procesando ingentes cantidades de fotos y detectando patrones y correlaciones estadísticas sin necesidad de modelos conceptuales o teóricos previos acerca de *qué es* un melanoma. Manteniéndonos en el ámbito de la salud y por ilustrar con otro ejemplo, estos grandes volúmenes de datos sirven también para crear «simulacros digitales», es decir modelos computacionales que sirven como representación de personas o grupos de personas (y también animales no humanos o plantas). Estos «gemelos digitales» pueden servir para generar predicciones de la evolución de una enfermedad o de la aplicación de un medicamento a lo largo del tiempo. Según informe*s recient*es, las pruebas *in silico* ya comienzan a reemplazar a algunas pruebas tradicionales en laboratorio (Moingeon et al., 2023).

Los métodos cuantitativos y estadísticos para investigaciones científicas no son nada nuevo. A partir del siglo XIX, basándose en el razonamiento inductivo promovido por filósofos como Bacon y los éxitos empíricos de Kepler, Newton y otros científicos naturales durante el siglo XVII, un grupo de pensadores ejecutó una verdadera revolución epistémica al considerar los patrones estadísticos como intrínsecamente explicativos (Hacking, 1990; Porter, 2020). Figuras como Quetelet y Galton establecieron las mediciones cuantitativas y el razonamiento estadístico como un modo legítimo de investigación incluso en las ciencias sociales. Esto se fortaleció con la aparición de la ciencia positiva, matematizada, que adoptó estos métodos como instrumentos tanto para la generación de conocimiento como para su demostración. Dichos métodos ubicuos y sus premisas de rigurosidad, neutralidad de valores y objetividad han sido frecuente y duramente criticados por varios autores (e.g., Hacking, 1990; Desrosières, 1993; Lewontin, 1993; Porter, 2020).

En línea con lo que comentábamos sobre los gemelos digitales en biomedicina, Cho et al., (2022, p.1) plantean que «los simulacros digitales marcan un hito importante en la trayectoria para abrazar la cultura epistémica de la ciencia de datos y un potencial abandono de los conceptos epistemológicos médicos de causalidad y representación». Vale la pena insistir que la perspectiva dataísta del *fin de la teoría* no es un corpus cohesionado de fuentes académicas o un marco teórico en un sentido estricto; más bien, es una mentalidad, e incluso una ideología (Blakely, 2020), que busca permear la generación de conocimiento, y sus implementaciones prácticas, en prácticamente todos los campos de la actividad humana.

¿Hasta qué punto la visión dataísta de la ciencia se ajusta a la realidad? ¿De qué forma esa visión oculta y distorsiona problemas epistémicos y éticos muy relevantes? Estas preguntas y otras similares son algunas de las que queremos explorar con esta selección de artículos. Así, esta Sección 1ª, *Inteligencia artificial, datos y objetividad. ¿El regreso del naturalismo Dataísta?* arranca con *¿Son las computadoras agentes inteligentes capaces de conocimiento?* de Gustavo Esparza y Daniel Martínez. El artículo explora los fundamentos epistémicos del dataismo analizando las posibilidades y límites de los algoritmos generados por aprendizaje automático a la hora de establecer conocimiento científico. Partiendo de un análisis de los fundamentos filosóficos de la arquitectura de programación en dos sistemas de Inteligencia Artificial (AlphaGo y Hide and Seek) se postula que tales algoritmos pueden llegar a ser recursos de conocimiento especiales para la comprobación de hipótesis, estableciendo así cómo los algoritmos de aprendizaje automático pueden realmente producir innovaciones epistémicas.

El resto de artículos se centra en tratamientos críticos de cuestiones asociadas al aprendizaje automático, en particular al mantenimiento y proliferación de sesgos epistémicos y éticos.

El artículo de Cristian Moyano, *La IA usada en biología de la conservación es una buena estrategia de justicia ambiental?,* aborda la cuestión de los sesgos y la analiza en el contexto específico del conservacionismo, aportando así un enfoque distintivo acerca de un tema ampliamente comentado en las investigaciones éticas acerca de la IA en la última década y en este mismo número. En dicho artículo se reconocen las posibilidades que ofrece la IA en el campo de la conservación de especies biológicas, pero apunta también a los problemas reales que un uso no controlado de estos algoritmos podría provocar al expandir sesgos epistémicos y éticos ya presentes en la actualidad en procesos de conservación. El artículo muestra así la importancia de ir caso por caso y analizar en cada disciplina específica que tipos de sesgos epistémicos y éticos son relevantes y cómo afrontarlos.

El tema de los sesgos se explora también en *Discurso influenciado: aprendizaje automático y discurso de odio* de Federico Javier Jaimes. El artículo analiza cómo la propagación sistémica de sesgos vía algoritmos de aprendizaje automático puede, a partir del concepto de «discurso influenciado», explicar la reproducción social de los discursos de odio al enmarcar teóricamente las formas en que algoritmos de aprendizaje automático expanden y normalizan discursos de odio.

Cerrando una brecha: una reflexión multidisciplinar sobre la discriminación algorítmica, de Pilar Dellunde, Oriol Pujol y Jordi Vitrià, plantea un acercamiento sistemático al concepto de discriminación algorítmica, sin duda uno de los principales y más relevantes problemas epistémicos y éticos a la hora de considerar la aplicación de algoritmos de aprendizaje automático en la esfera humana. El artículo plantea la necesidad de entender un algoritmo como una tecnología intencional y por tanto abierta a sesgos, imposibilitando así esa supuesta metodología radicalmente objetiva y libre de teorizaciones que postula el dataísmo.

Finalmente, *Más allá de los datos: la transformación digital del museo tradicional,* de Alger Sans y Vicent Costa investiga asimismo el problema de los sesgos y explora filosóficamente cómo la introducción de tecnologías digitales e inteligencia artificial en el museo tradicional podría transformar los procesos educativos. Los autores insisten en la insuficiencia de un acercamiento dataísta a la hora de detectar sesgos epistémicos y éticos, los cuales podrían llevar a una situación de injusticia epistémica en los museos, y a exacerbar así las discriminaciones y exclusiones ya existentes en los museos tradicionales

Referencias

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Blakeley, J. (2020). *We built reality: How social science infiltrated culture, politics, and power.* Oxford University Press.

Brooks, D. (2013, Feb 4) *The Philosophy of Data*. The New York Times. Consultado 01/07/2023 desde https://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html

Cho, M. K., & Martinez-Martin, N. (2022). Epistemic Rights and Responsibilities of Digital Simulacra for Biomedicine. *The American journal of bioethics : AJOB*, 1–12. Advance online publication. https://doi.org/10.1080/15265161.2022.2146785

Desrosières, A. (1993). *La politique des grands nombres: Historie de la raison statistique*. Éditions La Découverte.

Eticas Foundation. (2021) *Sesgo de calificación crediticia y reproducción de desigualdad en préstamos para vivienda*. Consultado 01/07/2023 https://eticasfoundation.org/es/sesgo-de-calificacion-crediticia-y-reproduccion-de-desigualdad-en-prestamos-para-vivienda/

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*, 5, 5-15.

Casacuberta, D., Guersenzvaig, A. Using Dreyfus' legacy to understand justice in algorithm-based processes. *AI & Soc* 34, 313–319 (2019). https://doi.org/10.1007/s00146-018-0803-2

Hacking, I. (1990). *The Taming of Chance*. Oxford University Press.

Hadwick, David & Lan, Shimeng. (2021) Lessons to Be Learned from the Dutch Childcare Allowance Scandal: A Comparative Review of Algorithmic Governance by Tax Administrations in the Netherlands, France and Germany. *World tax Journal*. 13(4), 609-645.

Lewontin, R. (1993). *Biology as Ideology: The Doctrine of DNA*. Harper Perennial.

Morales Moreno, A. M. (2021). Algoritmos en el estrado, ¿realmente los aceptamos? Percepciones del uso de la inteligencia artificial en la toma de decisiones jurídico-penales. *Ius et Scientia*, 7 (2), 57-87. https://doi.org/10.12795/IETSCIENTIA.2021.i02.05

Moingeon, P., Chenel, M., Rousseau, C., Voisin, E., & Guedj, M. (2023). Virtual patients, digital twins and causal disease models: Paving the ground for in silico clinical trials. *Drug discovery today*, 28(7), 103605. https://doi.org/10.1016/j.drudis.2023.103605

Porter, T. (2020). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life,* new edition. Princeton University Press.

Sánchez-Monedero, J., & Dencik, L. (2022). The politics of deceptive borders: 'biomarkers of deceit' and the case of iBorderCtrl. *Information, Communication & Society,* 25(3), 413-430. https://doi.org/10.1080/1369118X.2020.1792530

Tewari, I., & Pant, M. (2020). Artificial Intelligence Reshaping Human Resource Management : A Review. *2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation* (ICATMRI), Buldhana, India. https://doi.org/10.1109/ICATMRI51801.2020.9398420

Valle-Cruz, D., Fernandez-Cortez, V., & Gil-Garcia, J. R. (2022). From E-budgeting to smart budgeting: Exploring the potential of artificial intelligence in government decision-making for resource allocation. *Government Information Quarterly*, 39(2), 101644. https://doi.org/https://doi.org/10.1016/j.giq.2021.101644

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 13-28
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.557551

¿Son las computadoras agentes inteligentes capaces de conocimiento?*

Are computers intelligent agents capable of knowledge?

*GUSTAVO ESPARZA***
*DANIEL MARTÍNEZ****

Resumen. El objetivo del presente artículo es estudiar los fundamentos filosóficos de la arquitectura de programación en dos sistemas de Inteligencia Artificial (*AlphaGo* y *Hide and Seek*). El problema dilucida la distinción epistemológica de los conceptos "conocimiento", "intuición" y "abducción", para definir si el cumplimiento exitoso de una métrica programada, por parte de una computadora, es condición suficiente para atribuirle un comportamiento inteligente. A través del análisis de ambos ejemplos se muestran dos cuestiones: i) el cumplimiento exitoso de un objetivo programado ofrece nuevos recursos de conocimiento, ii) dichos conocimientos dependen de la ejecución de un programa cuyo procesamiento es desarrollado por una IA y, por tanto, las operaciones superan las capacidades intelectivas humanas. Las conclusiones apuntan a que las computadoras son recursos de conocimiento especiales de comprobación de hipótesis.
Palabras clave: Inteligencia Artificial, Inteligencia Humana, abducción, AlphaGo, Hide and Seek.

Abstract. The aim of this paper is to study the philosophical foundations of the programming architecture in two Artificial Intelligence systems (AlphaGo and Hide and Seek). The problem elucidates the epistemological distinction of the concepts "knowledge," "intuition" and "abduction," in order to define whether the successful fulfillment of a programmed metric, by a computer, is a sufficient condition to attribute intelligent behavior to it. Through the analysis of both examples, two issues are shown: i) the successful fulfillment of a programmed objective offers new knowledge resources, ii) such knowledge depends on the execution of a program whose processing is developed by an AI and, therefore, the operations exceed human intellectual capabilities. The conclusions point to the fact that computers are special hypothesis-testing knowledge resources.
Keywords: Artificial Intelligence, Intuitive Thinking, Creative Thinking, Alpha Go, Hide and Seek.

Recibido: 15/02/2023. Aceptado: 17/06/2023.

* Agradecemos la lectura y recomendaciones de dos lectores anónimos cuyas aportaciones y sugerencias permitieron una mejora sustancial al presente trabajo. Ambos autores fuimos responsables de: planteamiento del problema, revisión bibliográfica, redacción del artículo, discusión y revisión de los avances, presentación de resultados. El orden de los autores se eligió de acuerdo al siguiente criterio: primer autor se encargaría de las gestiones con la revista (autor de correspondencia), segundo autor, daría seguimiento a las comunicaciones derivadas de lo primero.

** Universidad Panamericana, Aguascalientes, México. Profesor investigador del Instituto de Humanidades. Línea de investigación: "Kant y la tradición kantiana. Las preguntas de la sabiduría humana", con especial énfasis en la filosofía de Cassirer. Su último libro en coautoría se titula: *The Bounds of Myth. The Logical Path from Action to Knowledge* (Brill, 2021). Miembro del Sistema Nacional de Investigadores, Nivel 1.

*** Centro de Investigaciones en Óptica, A.C., Aguascalientes, México. Estudiante de doctorado en el área de visión e inteligencia artificial. Su trabajo de doctorado se enfoca en el desarrollo de un sistema de navegación autónomo basado en aprendizaje reforzado profundo.

1. Introducción

El objetivo del presente artículo es estudiar los fundamentos filosóficos de la arquitectura de programación mediante la cual una máquina procesa información aleatoriamente con el fin de lograr un resultado que sobrepasa las capacidades humanas; actualmente existen computadoras cuyos resultados son considerados altamente especializados o incluso superiores a los alcanzados por cualquier ser humano. Como ejemplo de esto, consideramos la programación de *AlphaGo* (2021) como el primer agente (sistema inteligente) capaz de vencer a un campeón mundial de 'Go' y a las mejoras estratégicas alcanzadas en la interacción de equipos de agentes virtuales que juegan a las escondidas (en adelante *Hide and Seek*) (Baker, *et al*., 2020). En ambos casos se mostrará a la arquitectura como un modelo de resolución de problemas cuyos resultados podrían clasificarse como una actividad cognitiva altamente compleja. Lo relevante de esto, como se mostrará más adelante, es que el éxito del programa no puede explicarse recurriendo únicamente a la capacidad cognitiva del programador, pues sus propias habilidades no alcanzarían el mismo éxito que el programa ofrece, por lo que el logro puede atribuirse a la operación del programa. Sin embargo, el mismo resultado requiere del acto de programación a cargo de un agente humano para que tal finalidad se cumpla, generándose así una interdependencia.

Antes de comenzar a indagar en este problema, establecemos tres cuestiones a tomar en cuenta: primero, si bien las posibilidades de ejecución de un programa dependen de la métrica definida inicialmente por el algoritmo diseñado, así como del algoritmo de aprendizaje en lo general, los resultados alcanzados por la computadora no pueden atribuirse al agente programador porque en la definición inicial de métricas no se detallan los procedimientos para alcanzar los resultados. Segundo, el conjunto de resultados alcanzados por la computadora depende de la definición paramétrica inicial; existe evidencia de programas (*AlphaGo* y *Hide and Seek*) que permite sostener que alguna forma de procesamiento de datos no ha sido definida inicialmente, de tal modo que la diversificación de los comportamientos y procesamiento de la información depende de la arquitectura de programación. A pesar de lo anterior, no hablaremos de "libertad artificial" o "intención artificial" atribuible a la computadora, pero, según veremos, es notable que las estrategias de juego presentados en ciertos programas no se encuentre definida de facto en los parámetros de funcionamiento y ni siquiera dentro de las estrategias tradicionales de juego desarrollados por los profesionales de dichos juegos; consideramos que este fenómeno merece una atención filosófica para aproximar una respuesta a qué ocurre tanto en el juego mismo como en la arquitectura de programación. Tercero, derivado de lo anterior, consideramos que una línea de interpretación que puede explicar estos logros es la siguiente: la definición de los parámetros y su arquitectura, al tener como intención el éxito de la tarea designada, asume el proceso de aprendizaje profundo (*Deep Learning*) como una fase de interconectividad de los datos en la cual la combinación y permutación iterada de los datos sigue (a) un patrón predefinido, pero también (b) un esquema creativo cuyo único proceso de orden es, precisamente, el parámetro de éxito definido. Esto anterior (c) implica, por un lado, que el agente humano concibe un medio para resolver un problema inicialmente planteado, pero, por el otro, requiere de una IA para ejecutar la solución. Esta es la paradoja general a la que queremos referirnos, ya que el

resultado final si bien puede presentarse como una novedad estratégica del juego, es un logro que no puede atribuirse a un ser humano directamente, pero tampoco a una computadora únicamente. Consideramos, por ende, que esta interrelación merece un estudio detallado.

El plan de trabajo queda como sigue: el segundo apartado estudia los conceptos de conocimiento, intuición y abducción; la finalidad es ofrecer un marco epistemológico interpretativo que permita definir qué acciones cognoscitivas pueden ser consideradas para explicar casos de éxito logrados por un agente inteligente y si es que alguno de los escenarios puede aplicarse al programador o una computadora. El tercer apartado describe la arquitectura de programación de ambos agentes y, además, se reflexionan sus implicaciones filosóficas apoyados de los conceptos previamente propuestos. El cuarto apartado plantea que los sistemas inteligentes construyen nuevas formas de resolución de problemas a partir de las funciones generales propuestas, lo que puede ser considerado como una solución inteligente. Finalmente, se hará notar que los objetivos alcanzados por las computadoras ofrecen nuevos marcos estratégicos de solución de problemas que regularmente no pueden ser alcanzados mediante recursos humanos, lo que permite reinterpretar tanto nuestros alcances y capacidades intelectivas.

2. Conocimiento, intuición y abducción

En el presente apartado definiremos tres conceptos epistemológicos con el fin de ofrecer un marco interpretativo que nos permita identificar qué tipo de tareas pueden atribuirse a una inteligencia humana y cuáles a una inteligencia artificial. Consideramos que la diferenciación y criterios que a continuación presentamos nos permitirá delimitar en qué medida el cumplimiento exitoso de un parámetro predefinido es susceptible de entenderse como una prueba de un agente inteligente o simplemente la reorganización de datos cuyo estudio posterior garantiza nuevas posibilidades de interpretación.

Comenzamos por definir el concepto de conocimiento partiendo de la definición clásica dada por Platón (2015) *S* conoce que *P* cuando posee una *creencia*, *verdadera y justificada* (*Teeteto*, 201d). Encontramos una ampliación de estos criterios en la defensa de Sócrates respecto de sus actividades públicas, ya que argumenta que todo sujeto debe ser capaz de argumentar cómo es posible definir un conocimiento como propio (*Apología*, 19d-33c). En contraparte, esta definición ha sido puesta en duda por Gettier (1963, 121-123) a partir de la elaboración de contraejemplos en los que cuestiona la suficiencia tripartita para definir que *S* conoce que *P*, lo que hace pensar que dicha definición está incompleta.

John Ian K. Boongalig (2021, 87-111) ha evaluado el cuestionamiento de Gettier en contra de la definición clásica diciendo que el argumento más fuerte propuesto por el filósofo norteamericano frente a la tesis platónica son las circunstancias azarosas que parecen justificar que *S* no conoce verdadera y justificadamente que *P*. En ese sentido Boongalig apunta que aun cuando algunos eventos particulares pueden explicarse a partir de circunstancias fortuitas, ello no permite sostener universalmente que *S* no conozca que *P*; en contraparte, el propio Boongalig considera que para sostener la propuesta de Gettier como contra argumento válido de la definición clásica, es necesario que un segundo agente (S_2) presente pruebas que demuestren categóricamente que efectivamente *S* no conoce que *P*, pues el azar mismo

no es prueba fehaciente de que *S* desconozca que *P*. Sin embargo, aún en el caso que algún agente S_n demuestre que S no conoce, el mismo criterio se aplica para la justificación de S_2, pues un tercer agente S_3 puede demostrar que lo dicho por S_2 es erróneo. El círculo, como se aprecia, se vuelve infinito, pues en cada nuevo escenario S_n puede dudar de su antecesor y de las pruebas que éste presente, a menos que se establezca un criterio general a partir del cual determinar la imposibilidad de la duda para cualquier caso. La propuesta de Boongalig, entonces, es recolocar la definición de conocimiento en sus términos clásicos ¿puede *S* justificar la verdad de su creencia (*P*)?

Esta discusión previa es importante pues podemos reformularla de la siguiente forma para aplicarla a nuestro tema, ¿puede una IA conocer que *P*? Es decir, ¿puede una IA justificar la verdad de su creencia? Como se verá más adelante con los ejemplos, una respuesta a estas preguntas puede formularse del modo siguiente:

a. *Verdad*. Una IA otorga los resultados para los que fue programada, por ende, se puede establecer que este rubro se cumple.
b. *Justificación*. Una IA, sin embargo, al estar programada para cumplir con una tarea específica, no puede dar una explicación convincente de cómo es que alcanzó el resultado logrado. En todo caso, dicha labor corresponde al programador.
c. *Creencia*. Una IA opera un comando como parte de su funcionamiento. No es posible atribuir un estado mental a una computadora diciendo que ésta cree que la programación asignada es el mejor medio para resolver una métrica prestablecida. En todo caso, tal situación corresponde al agente humano quien cree que la programación puede alcanzar el parámetro definido, por ende, este rubro no es atribuible a una máquina.

En resumen, una IA no puede conocer que *P*, pues no cumple con los criterios previamente definidos. Si bien el programador aprovecha a la IA como instrumento para el desarrollo de nuevos conocimientos, ello no implica que la propia computadora sea capaz de conocer, sino que es un recurso para que el programador conozca.

Pasamos ahora a delimitar el concepto de intuición y para ello seguimos a Rafael Miranda-Rojas (2018) para quien, por tal, se ha de entender un "parecer intelectual" en donde el sujeto S intuye que un predicado P es verdadero, lo cual implica que "S aprende de modo inmediato la verdad de P", la cual, además, "es comprendida intelectualmente, racionalmente" (Miranda-Rojas 2018, 263). Sin embargo, de lo anterior se deriva un problema general: el que *S intuya* que *P* ¿necesariamente hace que *P* sea verdadero? Miranda-Rojas sostiene que "no es a través de la intuición que se justifica en última instancia la verdad de cierto enunciado, el nexo causal entre que P parezca verdadero y que P sea verdadero no es un caso de infalibilidad" (Miranda-Rojas 2018, 264). De acuerdo a esto, entonces, advertimos un modo de conocimiento no discursivo, pero cuyo objeto es el conocimiento de la verdad de *P*, pero en cuyo proceso de fundamentación la intuición únicamente opera como un caso donde la identidad entre el sujeto y el objeto se funda en la *creencia* de que *P* es verdad.

Es cierto que la fundamentación de *P* en un acto de creencia, como es bien sabido, no constituye un conocimiento *per se*, aunque sí implica uno de sus elementos para determinar la veracidad del postulado. En ese sentido, para Robert Audi (2019) la intuición sería:

> a non-inferential knowledge or grasp, as of a proposition, concept, or entity, that is not based on perception, memory, or introspection; also, the capacity in virtue of which *such cognition is possible* (p. 527).

En este contexto, el intercambio entre la 'creencia' de que *P* es verdadera por su *posibilidad cognitiva* constituye una renovación en los términos de fundamentación de la intuición como un medio válido para conocer la verdad de un postulado. De la definición se resaltan el carácter "no inferencial" de un "conocimiento"; ello implicaría, por tanto, que existe un conocimiento que se logra por medios no perceptivos, memorísticos o introspectivos, de tal modo que cumplirían con las condiciones de operar como *creencia, verdadera y justificada* (Platón, *Teeteto*, 201d); dicho de otro modo, para que una intuición sea considerada como tal ésta debe apuntar necesariamente a un conocimiento que sea "verdadero", pero a través de una justificación inferencial. La principal diferencia de la 'intuición' y el 'conocimiento', está en los medios y posibilidad de "justificación" de un postulado. Si aprovechamos estos resultados podemos preguntar ¿es posible que una IA intuya que P?

a. *Verdad*. Como se dijo previamente, este rubro se cumple por parte de una IA.
b. *Justificación*. De acuerdo a lo dicho, si bien una IA no puede justificar directamente el éxito de su programación, al constituirse como el instrumento o recurso para alcanzar el éxito de un objetivo, podríamos decir que la razón por la cual un proceso se cumple es, precisamente, su aprovechamiento. En ese sentido, el sentido de la justificación que el programador pueda ofrecer depende del resultado de la IA y, por ende, forma parte de la justificación.
c. *Creencia*. En este caso ocurre un proceso similar al previo. La IA directamente no asume una creencia, pero forma parte del marco de recursos que hace que el programador pueda asumir la creencia de que un procedimiento puede ser alcanzado. En ese sentido se puede decir que el programador puede creer en la posibilidad de éxito asumiendo el desarrollo de un algoritmo como base de dicho estado mental.

En resumen, a pesar de todas las consideraciones, no es posible establecer que una IA intuya, pues buena parte de las interpretaciones requieren una inteligencia humana para lograr los criterios.

Pasamos ahora a estudiar el concepto de abducción o hipótesis propuesto por Charles S. Peirce (1998) quien la define del siguiente modo:

> The abductive suggestion comes to us like a flash. It is an act of insight, although of extremely fallible *insight*. It is true that the different elements of the hypothesis were in our minds before; but is the idea of putting together what we had never before dreamed of putting together which flashes the new suggestion before our contemplation (p. 227. Énfasis en el original).

Como se aprecia, el concepto de abducción constituye un momento de conocimiento cuyo desarrollo implica un acto de "insight" o aparición. Lo relevante de este modelo es que el acto cognoscitivo no requiere de una justificación según requiere el sentido clásico del conocimiento. De acuerdo a lo que veremos en los ejemplos más adelante, todo proceso de programación se estructura como el desarrollo de una serie de proposiciones con las cuales

se espera conseguir un resultado predefinido. Ello implica que el acto de conocimiento que se alcanza no depende del acto de justificar cómo es que el proceso desarrollado por la computadora logró el resultado, sino que se sostiene de la hipótesis propuesta por el agente al momento de proponer su programa como un algoritmo capaz de resolver una tarea.

En este proceso es evidente que el acto de confirmación no implica una justificación de parte del programador mismo, sino únicamente la validación de que los comandos propuestos son un recurso efectivo para lograr el éxito establecido. Kyrylo Medianovsky y Ahti-Veikko Pietarinen (2021) insisten en que un sistema inteligente que procesa grandes cantidades de datos opera como un sistema abductivo, lo cual permite a un agente humano interpretar la información para con ello completar el conocimiento. Además de lo anterior, hacen notar que mientras que un conocimiento requiere de explicaciones causales, el procesamiento de los datos por parte de una IA únicamente requiere de una programación y un criterio que debe ser alcanzado. Esta diferencia permite reconocer que mientras que un agente humano es capaz de producir conocimiento, una computadora no puede cumplir con los mismos criterios epistémicos, sin embargo, ello no implica que los resultados entregados carezcan de validez para el agente humano que los interpreta.

Los autores proponen varios ejemplos para mostrar, por un lado, que actualmente una computadora por sí sola no puede detectar el grado de riesgo patológico, meteorológico o sísmico, de los datos que procesa, pero sí, en cambio ofrecer nuevos modelos de organización de datos ofreciendo datos cuya interpretación no podría alcanzarse únicamente aprovechando las capacidades humanas. Medianovsky y Pietarinen (2021) remarcan que es importante aceptar el lugar que ocupan las IA dentro de la producción de conocimiento.

De acuerdo con esto, se puede entender mejor el lugar que desempeña una IA en el esquema general hasta aquí planteado. Al operar como un sistema operativo que favorece la producción y desarrollo de conocimiento, el mejor esquema explicativo tiene que ser uno que reconozca las funciones abductivas que una computadora ofrece a un agente humano para lograr sus tareas y actividades de investigación.

Colocamos una última cuestión previo al análisis de los programas *AlphaGo* y *Hide and Seek*. John McCarthy, a mediados de la década de los 50's, convoca a una reunión de investigación en Darmouth para delimitar tanto los problemas propios de una *AI*, como su campo de investigación. Una de las cuestiones centrales que proponía estudiar era la aleatoriedad y la creatividad (*Randomness and Creativity*), cuya conjetura general proponía: "The randomness must be guided by intuition to be eficient" (McCarthy, et al., 1955, 2). Con dicha hipótesis el autor configuraría el campo de la inteligencia artificial, del cual más tarde se derivarían ramas como el aprendizaje profundo (*Deep Learning*) y aprendizaje por reforzamiento (*Reinforcement Learning*), y que hoy en día ofrecen una gran cantidad de evidencia computacional mostrando dos cuestiones importantes:

1) Un sistema inteligente basado en aprendizaje por reforzamiento opera en función de la delimitación de una métrica que define el éxito o fracaso de un resultado.
2) A pesar de que se delinea una arquitectura de funcionamiento, el procesamiento de la información no depende de la estructuración de los comandos iniciales, sino de la métrica general estipulada, lo que implica que el éxito está en función de la cantidad y tipo de datos que el programa puede combinar para resolver un problema.

A partir de esto, planteamos que existen programas basados en *Deep Reinforcement Learning* cuya operación cumple con las condiciones del concepto de abducción previamente expuesto, pues el agente programador establece una idea general que luego es alcanzada por un sistema de procesamiento de la información que le permite lograr una métrica general predefinida. Para demostrar esto, en el siguiente apartado se describirá la arquitectura de funcionamiento de dos programas cuyos resultados cumplen con las especificaciones métricas de sus diseñadores, sin que por ello implique que tales logros sean atribuibles o replicables por ellos mismos; en el caso de *AlphaGo* no es posible sostener que los programadores puedan ser considerados "campeones mundiales del juego" y, en el caso de *Hide and Seek*, los esquemas de búsqueda o escondite estén ya estipulados en el comando, sino que se derivan de las métricas y arquitectura del programa en sí, por lo que el procesamiento de la información no está supeditado a la conformación de estrategias de éxito del programa, sino al cumplimiento de la meta definida como final.

3. *Alpha Go* y *Hide and Seek*. Arquitecturas de programación

Dentro de las distintas propuestas de IA encontramos modelos procesamiento de información, así como de validación de éxito de una tarea previamente definida, en donde el único criterio de comprobación es la satisfacción de la propia métrica predefinida. En este sentido, se dice que una programación ha cumplido con su función si desarrolla la tarea asignada; analógicamente diríamos que un programa es exitoso si cumple con las tareas definidas inicialmente por el programador, lo cual implica que la verdad del programa (su éxito) está predefinido por el propio agente.

A continuación, presentamos dos ejemplos de IA en los que se diferenciará entre la programación que propone un agente humano para cumplir con ciertas tareas que él mismo no puede cumplir con recursos y capacidades humanas, por lo que es necesario recurrir al procesamiento de información a través de una computadora inteligente, la cual, si bien responde a criterios de programación pre establecidos, sus resultados se derivan de las funciones y capacidades tecnológicas de la computadora misma. Veamos ambos casos con detalle.

3.1. Mastering the Game. La intuición como métrica de programación

Un ejemplo de IA cuyo objetivo fue alcanzar una habilidad "súper humana" en el juego de 'Go', es el programa de *DeepMind*, *AlphaGo* y sus diferentes versiones. Algo importante a resaltar de *AlphaGo* es que fue la primera inteligencia artificial en derrotar a jugadores profesionales de Go, incluyendo al 18 veces campeón mundial, Lee Sedol (Granter, Beck y Papke Jr., 2017, 619-661). El proceso de entrenamiento de *AlphaGo* se detalla a continuación.

En la primera fase de entrenamiento se utilizó *Aprendizaje supervisado* (*Supervised Learning*). Inicialmente se entrenó a *AlphaGo* con cientos de jugadas de jugadores profesionales para que la *IA* "aprendiera" las reglas del juego "observando" las jugadas estratégicas convencionales (Gibney, 2016, 445-446). De acuerdo con Florian Brunner (2019):

> This network takes the current board state as a 19x19x48 batch as input and outputs a probability distribution over all legal moves *a*. The network has been trained on randomly sampled state-action pairs (*s,a*), using stochastic gradient ascent to maximize the log likelihood of selecting move *a* in state *s* […] This network predicted expert moves with an accuracy of 57.0% (p. 7).

Con el uso de redes neuronales profundas para calcular las probabilidades de movimientos, lo que se pretende es garantizar una toma de decisiones basada en los cálculos que la propia computadora puede procesar, pero, al mismo tiempo, ordenar dichas operaciones a una métrica general determinada. De este modo, se aprecia que, si bien la delimitación del evento exitoso y las fórmulas para lograrlo son precisadas por los programadores, no implica que esté definido el medio a través del cual es posible lograr el objetivo.

En la segunda fase, una vez lograda la tasa de éxito esperada, se entrenó a *AlphaGo* utilizando *Aprendizaje reforzado* (*Reinforcement Learning*). El objetivo general del entrenamiento fue recalibrado para, en lugar de predecir las jugadas y movimientos del experto, ahora se fijaría como meta el "ganar los juegos". En este estadio, la IA simuló millones de partidas contra sí misma con el fin de evaluar la tasa de éxito de las estrategias convencionales, además de desarrollar nuevas opciones de juego para garantizar el cumplimiento la métrica general. David Silver (2017) y su equipo al respecto aclaran: "The policy network was trained initially by supervised learning to accurately predict human expert moves, and was subsequently refined by policy-gradient reinforcement learning. The value network was trained to predict the winner of games played by the policy network against itself" (354-359). Con este reajuste en la lógica de operación, el programa, si bien calculaba los posibles movimientos del oponente para generar una estrategia de juego acorde a la situación, su métrica ahora implicaba que, en un segundo momento, debía delinear nuevas estrategias para alcanzar el éxito el cual se definía como "ganar" el juego.

Eventualmente *AlphaGo Zero* vencería en una serie de juegos a su predecesora *AlphaGo* con un resultado de 100 - 0. Esto permitió sostener que esta IA tiene un mejor desempeño en el juego de Go que su versión previa. David Silver, uno de los investigadores principales del proyecto, afirma que el aprendizaje inicial de *AlphaGo* a partir de las jugadas de profesionales humanos, es lo que limita la creatividad de esta IA, sesgando sus estrategias hacía la experiencia humana. Es precisamente el hecho de que *AlphaGo Zero* aprende lo que se traduce en la exploración del juego desde una perspectiva única[1]. Pero ¿cómo es que aprende esta inteligencia artificial? A continuación, revisaremos el funcionamiento general de la arquitectura de *AlphaGo Zero*.

---

1 Al respecto, son interesantes las reflexiones de Granter, Beck y Papke Jr. (2017) pues muestran tanto áreas que superan las capacidades humanas como sus propias limitaciones al respecto: "AlphaGo's success at surpassing human experts, computer vision algorithms fall short of matching some basic human vision capabilities. When humans look at pictures, they can interpret scenes and predict within seconds what is likely to happen after the picture is taken (''object dynamic prediction'' in computer terminology). However, although algorithms can be trained to make similar predictions in abstract cartoon scenes, they cannot accurately predict what will happen next in real-world photographic scenes. And algorithms are notoriously bad at some aspects of image analysis. For example, algorithms cannot accurately predict whether a human will find a photograph funny or not, to the point where humor detection is considered an ''AI-complete problem''". (p. 620).

Esta IA está compuesta por tres etapas principales. Una *primera red neuronal* profunda para predecir el posible ganador de la partida en cada estado del juego. Es decir, a partir del estado actual del tablero, la red entregará como salida valores entre '1' y '-1', de modo que, si hay una alta posibilidad de que el jugador 1 gane, el valor de salida se aproximará a 1, pero si el jugador 2 tiene mayor posibilidad de ganar la partida, el valor se aproximará a -1, y si el juego tiene alta posibilidad de terminar en empate, el valor de la red se encontrará cercano a 0 (Brunner, 2019, 8-9). Esta red neuronal permite evaluar el estado actual del juego en cada movimiento y la confianza y exactitud de la predicción de esta red depende de la experiencia obtenida por la IA (conocimiento de un mayor número de estados del juego).

Una *segunda red neuronal profunda* para determinar el mejor movimiento a partir del estado del juego. Esta red, "observa" el tablero y los últimos movimientos de ambos jugadores para determinar cuál es el siguiente movimiento que el jugador debe hacer. Cómo salida, la red neuronal entrega una matriz de valores entre '0' y '1' para cada posición en el tablero. La posición con el mayor valor es la mejor jugada (conocida) para ese estado del tablero. La optimización de esta red se da a través de evaluar si el movimiento fue beneficioso para el resultado final (ganar o perder la partida). De este modo la red neuronal se ajusta para sugerir mejores movimientos conforme a su experiencia.

La tercera etapa es un árbol de búsqueda de Monte Carlo (MCTS, por sus siglas en inglés. Ver figura 3.1), este algoritmo realiza internamente simulaciones de los posibles movimientos tanto de la IA como del oponente. De este modo, *AlphaGo Zero* crea un árbol ramificando diferentes jugadas y los posibles desenlaces de cada jugada. En esta etapa, el MCTS utiliza la red neuronal de la segunda etapa para predecir cuales son las mejores jugadas que la IA puede hacer, y cuáles podrían ser las mejores jugadas del oponente. Finalmente se utiliza la rama del MCTS que maximice las posibilidades de ganar (Silver, et al., 2017, 354).

El MTCS y la red neuronal de la segunda etapa se optimizan utilizando referencias circulares, es decir, la red neuronal se ajusta comparando sus predicciones con el MTCS, y a su vez, el MTCS utiliza las predicciones de la red neuronal para simular las mejores jugadas posibles. De este modo, conforme el MTCS explora más jugadas, la red neuronal puede intuir de mejor manera las mejores jugadas, y producir mejores resultados en las simulaciones del MCTS, y así sucesivamente (Silver, et al., 2017, 355-356). Luego de millones de juegos de experiencia, *AlphaGo Zero* produce jugadas infalibles a través de una intuición "súper humana" del juego (Silver, et al., 2017, 358).

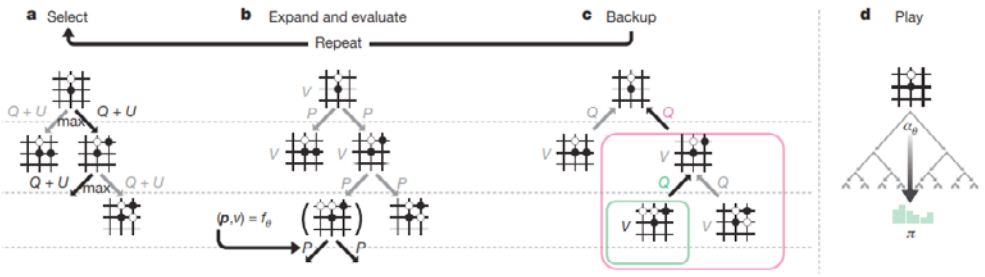

Figura 3.1 *Árbol de búsqueda de Monte Carlo para Alpha Go* (Silver, et al., 2017, 355.)

La parte creativa de *AlphaGo Zero* se produce a través de un parámetro del MCTS que permite a la *IA* explorar nuevas jugadas, incluso si estas no se ajustan a las estrategias convencionales del juego desarrollado por profesionales, esta exploración aleatoria lleva a *AlphaGo Zero* a desarrollar jugadas nuevas y evaluar si estas se aproximan a un escenario de éxito (1) o fracaso (-1). Durante la etapa de entrenamiento, este parámetro es más alto para maximizar la exploración de *AlphaGo Zero*, mientras que en partidas competitivas es menor, lo cual lleva a la IA a ejecutar las mejores jugadas previstas de acuerdo con su experiencia; sin embargo, permanece una pequeña posibilidad de hacer una jugada "creativa" o aleatoria (Silver, et al., 2017, 360-361). Lo que *AlphaGo Zero* demuestra es que puede desarrollar un estilo de juego basándose en los datos almacenados luego de jugar contra sí mismo. Al respecto, vale la pena reproducir las conclusiones de David Silver (2017) y su equipo sobre el logro alcanzado por su programa:

> Humankind has accumulated Go knowledge from millions of games played over thousands of years, collectively distilled into patterns, proverbs and books. In the space of a few days, starting *tabula rasa*, AlphaGo Zero was able to rediscover much of this Go knowledge, as well as novel strategies that provide new insights into the oldest of games (p. 358).

No deja de ser provocador este planteamiento, pues dicha afirmación se basa en los resultados logrados por una computadora enfrentando al campeón mundial del juego de Go. Pero lo que a nosotros nos interesa de estas conclusiones son dos cuestiones: (a) que el equipo de Silver, digamos, la parte humana, únicamente programó un proceso de juego al definir la métrica, los métodos de recuperación de información, así como el procesamiento de los datos, pero no así las estrategias para "ganar"; (b) al establecer como rango de desarrollo desde la *Tabula rasa* hasta la victoria contra Lee, implica que los programadores únicamente consideraban que a través de la programación de un sistema sería posible alcanzar la victoria, pero aún con ello no es posible sostener que fuesen capaces de derrotar directamente a Lee; es decir, para el caso de David Silver y su equipo, a pesar de su capacidad para programar una computadora que se alzó la victoria en contra del campeón de Go, ellos mismos no son capaces de genera una estrategia 'humana' que emule el mismo evento.

3.2. Hide and Seek. La creatividad como estrategia de intuición

Open *AI* (2019) ha desarrollado también diversas aplicaciones con inteligencias artificiales que han demostrado un comportamiento intuitivo, creativo y adaptable. Entre ellos se encuentra *Hide and Seek*, esta aplicación consiste en una serie de simulaciones en diferentes entornos virtuales donde dos equipos de agentes (escondidos y buscadores) aprenden el juego de "las escondidas" (Baker, et. al., 2019). El objetivo de este juego es que los participantes involucrados generen nuevas estrategias de éxito vinculadas al rol que desempeñan (esconderse o encontrar), lo que ha implicado un buen parámetro de medición para los programadores en cuanto al desarrollo de estrategias y de creación de acción. De acuerdo con Baker y su equipo, las reglas que tomaron en cuenta para su programa fueron las siguientes:

> We introduce a new mixed competitive and cooperative physics-based environment in which agents compete in a simple game of hide-and-seek. Through only a visibility-based reward function and competition, agents learn many emergent skills and strategies including collaborative tool use, where agents intentionally change their environment to suit their needs (Baker, et. al., 2019, 3).

Se esperaba que los agentes cumplieran con una métrica definida y designada para cada participante (esconderse o encontrar), a través de una toma decisiones, la cual, a su vez, dependía de una red neuronal profunda (Ver figura 3.2) para procesar información como: (1) La posición y velocidad tanto de sí mismo, como la de los otros agentes; (2) Campo de visión cónico; (3) La posición, velocidad y distancia de objetos del entorno como cajas, rampas o bloques. Mientras que, por otro lado, la salida de la red neuronal es la siguiente acción del agente, las acciones permitidas son: (i) Movimiento en cada dirección; (ii) Voltear a los alrededores; (iii) Sujetar objeto (para desplazarlo).

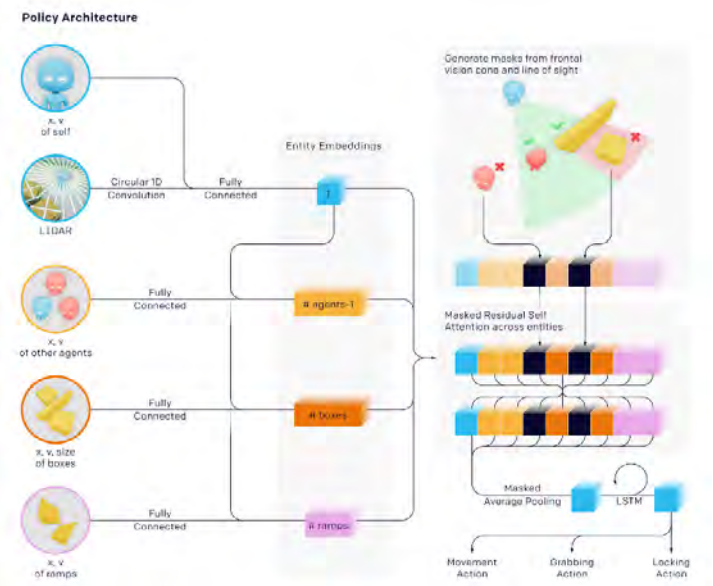

Figura 3.2 *Arquitectura de agentes de* Hide and Seek (Baker, et. al., 2019)

A partir de este marco, el comportamiento de los agentes está regido por una sola métrica cuyas políticas de recompensa pueden ser: (i) todo el equipo de buscadores es recompensado positivamente si un buscador encuentra a un agente escondido o recompensado negativamente si el equipo no puede encontrar a ningún escondido: (ii) el equipo de escondidos sea recompensado positivamente si logran mantenerse fuera del campo de visión de los buscadores o, también, si uno de los escondidos es encontrado todo el equipo recibe una recompensa negativa. De este modo, las simulaciones tratan de fomentar no sólo la creatividad de los

agentes para encontrar estrategias para esconderse, si no también, colaborar para que todos los miembros del equipo logren mantenerse fuera de visión.

De acuerdo con Baker y su equipo, en general los programas no existen estímulos explícitos que motiven a los agentes a interactuar con objetos del ambiente, por lo que la métrica debe asegurar que quien busca o se esconde, responda a la demanda de lograr un estímulo positivo (1) y evitar el negativo (-1), siendo, en ambos casos, definido por el sistema de programación, antes de iniciar el juego; pero ¿qué propósito tienen las recompensas en el contexto del aprendizaje de los agentes? En el aprendizaje reforzado, los agentes interactúan con el entorno a través de dos acciones básicas que eventualmente se traducen en una serie de movimientos tales como desplazar objetos de izquierda a derecha, abajo hacia arriba, o incluso empalmar más de un objeto para construir refugios (para el caso de quien se esconde) o trepar objetos, moverlos del lugar en el que se encuentran, rodear a otros agentes para evitar que escapen (para el caso de quien busca)[2]. Al terminar la iteración, se evalúa si el agente cumplió o no el objetivo. En caso de cumplir la meta, la red neuronal del agente promueve la ejecución de esa serie de acciones que llevaron a cumplir el objetivo, si el agente fracasa, la red neuronal desalienta al agente de realizar esa serie de acciones. Al cabo de cientos, miles, o incluso millones de iteraciones (dependiendo de la complejidad del entorno), el agente es capaz de intuir como interactuar con el entorno para maximizar la recompensa.

Open AI (2019) obtuvo resultados muy interesantes con respecto a la colaboración, creatividad y adaptación de los agentes para el caso particular del *Hide and Seek*; en uno de los experimentos realizados, obtuvieron los siguientes sucesos (ver figura 3.3):

a. Los agentes actuaban aleatoriamente, ya que no comprendían el objetivo del juego.
b. Los buscadores aprendieron a buscar a los escondidos.
c. Los escondidos aprendieron a bloquear las entradas de los buscadores utilizando cajas.
d. Los buscadores aprendieron a usar rampas para saltar las paredes que les impedían entrar al cuarto donde se encontraba el equipo contrario.
e. Finalmente, el equipo de escondidos aprendió a esconder las rampas y bloquear la entrada para evitar que los buscadores saltaran la pared.

2 El proceso de movimiento es detallado del modo siguiente: "Agents are simulated as spherical objects and have 3 action types that can be chosen simultaneously at each time step. They may move by setting a discretized force along their *x* and *y* axis and torque around their z-axis. They have a single binary action to grab objects, which binds the agent to the closest object while the action is enabled. Agents may also lock objects in place with a single binary action. Objects may be unlocked only by agents on the team of the agent who originally locked the object. Agents may only grab or lock objects that are in front of them and within a small radius", (Baker, et. al., 2019, p. 4).

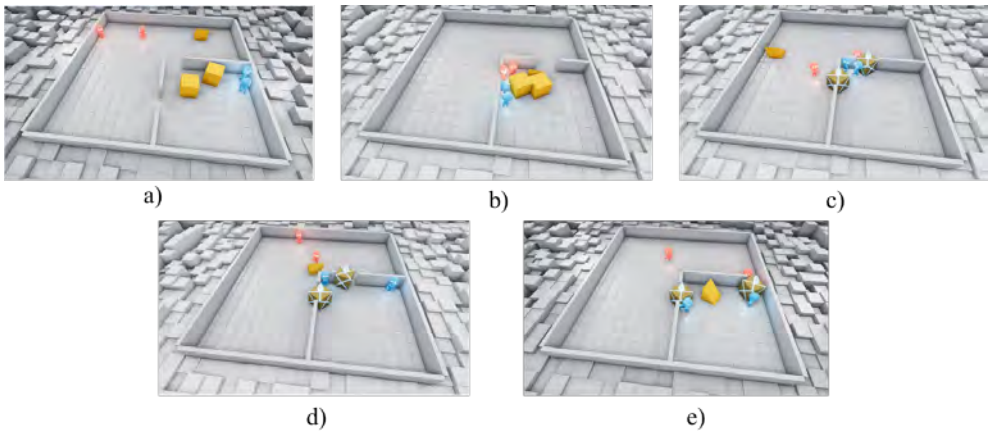

Figura 3.3 *Resultados simulación Hide and Seek* (Baker, et. al., 2019).

De esta serie de eventos, una vez que los equipos de agentes comprendieron el objetivo del juego y encontraron estrategias para cumplir la meta, generaron que el equipo contrario a adaptarse, creando nuevas estrategias que, a su vez, exigieron al equipo contrario a responder con la renovación estratégica de acciones para contrarrestar las opciones de búsqueda del equipo contrario. Este patrón se repitió continuamente hasta que el equipo de los escondidos encontró la forma de bloquear todas las posibilidades de los buscadores. Al menos para este entorno en particular, en otros entornos más complejos, se obtuvo el mismo patrón de adaptación entre ambos equipos, sin embargo, en algunas ocasiones, los agentes aprendieron incluso a aprovecharse del sistema de físicas de la simulación para crear nuevas estrategias. El propio equipo de Baker resalta estos logros como parte de una estrategia creativa cuya finalidad es asegurar la métrica final esperada:

> As agents train against each other in hide-and-seek, as many as six distinct strategies emerge, each of which creates a previously non-existing pressure for agents to progress to the next stage. Note that there are no direct incentives for agents to interact with objects or to explore, but rather the emergent strategies are solely a result of the autocurriculum induced by multi-agent competition (Baker, et. al., 2019, 6).

Al resaltar el valor de generación de "nuevas" estrategias, lo que se subraya es, por un lado, que la métrica predefinida por parte del equipo de programación establecía solamente las recompensas a lograr, aspecto que era buscado por el equipo de búsqueda y de ocultamiento. Sin embargo, el objetivo del equipo de Baker era medir el proceso de desarrollo de recursos mediante los cuales ambos programas cumplirían su objetivo hasta el punto en que no fuese posible por parte del equipo contrario encontrar soluciones a un problema inicial dado, de tal modo que el desarrollo del juego implicaría que uno de los dos equipos eventualmente no podría obtener la recompensa, mientras que el otro la lograría de modo indefinido.

Conclusiones

En el presente artículo nos propusimos como objetivo estudiar los fundamentos filosóficos de la arquitectura de programación mediante la cual una máquina procesa información con el fin de lograr un resultado exitoso que derive en un principio de orden. Para ello consideramos dos ejemplos de programación (*AlphaGo Zero* y *Hide and Seek*) y con ello mostrar el proceso mediante el cual los programadores lograron el cumplimiento de métricas definidas a través de recursos computacionales.

Partimos de la delimitación y objetivos estipulados por McCarthy en la conferencia en Darmouth, en donde se establecía que una de las tareas a desarrollar por parte de una *AI* correspondía al desarrollo de un pensamiento intuitivo como organizador de la creatividad. Consideramos los casos de *AlphaGo Zero* y *Hide and Seek* porque ambos plantean soluciones a juegos o simulaciones cuyos procesos suelen atribuirse a agentes humanos. Sin embargo, consideramos estos escenarios porque en ambos contextos se observa cómo la definición paramétrica de resultados no define el tipo de resultados por alcanzar, sino que, precisamente, se realizan para describir las estrategias mediante las cuales es posible alcanzar una meta de éxito. Los resultados obtenidos por *DeepMind* y *Open AI* –las empresas desarrolladoras de ambos programas– demuestran el potencial de la inteligencia artificial para resolver problemas en formas creativas que agentes humanos no han mostrado u ofreciendo alternativas de solución que resultan novedosas.

Como parte del análisis filosófico, se dispuso estudiar tres conceptos epistemológicos que nos permitirían evaluar el tipo de procesamiento que podía atribuirse a una IA. Se distinguió entre conocimiento (definido como creencia, verdadera, justificada), intuición (definido como creencia verdadera, pero que carece de justificación) e inducción o hipótesis (proceso cognitivo que requiere un caso y una regla para establecer una proposición verdadera). En este punto se estableció que los primeros dos conceptos no pueden ser vinculados con una computadora pues el conocimiento y la intuición implican operaciones que por ahora una IA no ha mostrado poseer y que, además, no se requieren por parte de una computadora para lograr el éxito de una métrica previamente definida.

En cambio, en el caso de la abducción, al tratarse de un proceso en donde se plantea un nuevo conocimiento, pero que no requiere de una justificación de cómo es que se obtuvo el resultado, se pudo mostrar que una actividad cognitiva de esta naturaleza permite un mejor modo de comprensión del tipo de papel que juega tanto un agente humano que aprovecha un programa computacional para resolver un problema. Con el modelo propuesto por Peirce es posible aceptar que el valor central de una IA no está en la capacidad reflexiva o la descripción de cada una de las fases del proceso que se cumple para lograr un resultado, sino, precisamente en la taza de logros alcanzados con un programa que demuestre ser la vía o recurso intelectual con el cual sobre pasar las capacidades intelectuales de un ser humano.

Para lograr esto anterior, emprendimos un análisis de los programas *Hide and Seek* y *Alpha Go*, concluimos que el uso e intermediación de una *AI* para la resolución de un problema del cual se desconocen las estrategias o recursos para la solvencia de la métrica, puede considerarse un caso de conocimiento intuitivo pues los equipos programadores, a pesar de que definen la meta final –delimitan el valor de éxito o verdad a la que debe llegar el programa– desconocen el proceso a través del cual alcanzar este escenario.

Con todo ello, consideramos que la reflexión sobre el papel que juegan las computadoras en la generación del conocimiento y, sobre todo, la posibilidad de justificar un resultado logrado como un proceso de búsqueda intencional y racional, validan la propuesta de McCarthy con respecto a la posibilidad que un sistema de IA basado en aprendizaje por reforzamiento opera en función de la delimitación de una métrica que define el éxito o fracaso de un resultado. Si bien los resultados ofrecidos por una computadora no pueden asumirse como conocimiento propiamente dicho, con los logros alcanzados por los programas estudiados se muestra que no toda mejora y desarrollo de nuevos recursos y medios de solución deben provenir de una inteligencia humana.

Esto último, recoloca el avance tecnológico en una nueva dimensión, pues no se puede afirmar *per se* que el vertiginoso avance de la IA constituya un atentado o problema para el ser humano, pues se trata de una herramienta que puede contribuir a la mejora de la sociedad. En todo caso, la pregunta más importante debe redirigirse a qué tipo de conocimientos buscan ser alcanzados por los agentes quienes a través de sus programas pretenden demostrar sus hipótesis. Una computadora, como se mostró, únicamente explora los mejores medios y recursos estratégicos para validar o desacreditar dicha proposición.

Referencias Bibliográficas

AlphaGo. (2021). *Google DeepMind*. *Alpha Go*. Recuperado de https://deepmind.com/research/case-studies/alphago-the-story-so-far.

Audi, R. (2019). "Intuition", *The Cambridge Dictionary of Philosophy*. Tercera Edición. Cambridge: Cambridge University Press, 2019.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B. y Mordatch, I. (2020). "Emergent tool Use from Multi-Agent Autocurricula", *The International Conference on Learning Representations* (conference paper), pp. 1-28. Disponible en: https://arxiv.org/pdf/1909.07528.pdf

Boongalig, J. (2021). *Disolving the Gettier problema*. *Beyond analysis*. Cambridge: Cambridge Scholars.

Brunner, F. (2019). "Mastering the game of Go with deep neural networks and tree search", *Artificial Intelligence for Games Seminar Report*, Heidelberg University, pp. 1-15.
https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/36349047/report_florian_brunner.pdf

Gettier, E. (1963). "Is Justified True Belief Knowledge?" *Analysis*, Vol. 23, No. 6, pp. 121-123. https://doi.org/10.2307/3326922

Gibney, E. (2016). "Google masters Go. Deep Learning software excels at complex ancient board game". *Nature*. Vol. 529, pp. 445–446.
https://www.nature.com/articles/529445a.pdf

Granter, S., Beck, A. y Papke Jr., D. (2017). "Alpha Go, Deep Learning, and the Future of the Human Micropist", *Arch Pathol*. *Lab Med*, Vol. 141, No. 5, pp. 619-61.
https://doi.org/10.5858/arpa.2016-0471-ED

McCarthy, J., Minsky, M., Rochester, N., y Shannon, C. (1955). *A proposal for the Darmouth Summer Research Project on Artificial Intelligence*.
http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

Medianovskyi, K. y Pietarinen, A.-V. (2022). On Explainable AI and Abductive Inference. Philosophies, Vol. 7, No. 35. https://doi.org/10.3390/philosophies7020035

Miranda-Rojas, R. (2018). "Intuición, racionalidad y confiabilidad", *Cinta Moebio*, Vol. 62, pp. 261-273. http://dx.doi.org/10.4067/S0717-554X2018000200261

Leib, R. (2021). "Interaction between Language and the other Symbolic Forms", en: Simon Truwant, ed., *Interpreting Cassirer. Critical Essays, Cambridge*: Cambridge University Press, pp. 11-13

OpenAI. (2019). *Emergent Tool from Multi-Agent Interaction*. Disponible Online: https://openai.com/blog/emergent-tool-use/

Park, W. (2022). "How to Make AlphaGo's Children Explainable". *Philosophies*, Vol. 7, No. 55. https://doi.org/10.3390/philosophies7030055

Peirce, Ch. (1998). "Pragmatism as the Logic of Abduction (Lecture VII)", en: Peirce Edition Project, eds., *The Essential Peirce. Selected Philosophical Writings. Vol. 2 (1893-1913)*, Bloomington and Indiana: Indiana University Press, pp. 226-240.

Platón, (2015). *Diálogos*, Traducción y notas de Ma. Santa Cruz; A. Vallejo Campos; N. Luis Cordero, Madrid: Gredos, 2015.

Silver, D., Schrittwieser, J., Simonyan, K., *et al*. (2017). "Mastering the game of Go without human knowledge", *Nature*, Vol. 550, pp. 354–359. https://doi.org/10.1038/nature24270

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 29-44
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.561551

¿La IA usada en biología de la conservación es una buena estrategia de justicia ambiental?*

Is AI used in conservation biology a good environmental justice strategy?

*CRISTIAN MOYANO FERNÁNDEZ***

Resumen. La biología de la conservación se ha sumado al uso de la inteligencia artificial para optimizar su trabajo. La eficiencia con que esta procesa los datos ayuda a identificar especies salvajes, reparar los impactos antropogénicos e intervenir en ecosistemas, ofreciendo resultados supuestamente buenos para la conservación. Así, la inteligencia artificial puede proponerse como una aliada de la justicia ambiental. Pero discutiré esta tesis, argumentando que como la biología de la conservación no parte de parámetros absolutos y la justicia ambiental no está exenta de una pluralidad moral, entonces la inteligencia artificial puede reproducir y aumentar los sesgos epistemológicos y éticos.
Palabras clave: biología de la conservación, inteligencia artificial, justicia ambiental, pluralidad, sesgo.

Abstract. Conservation biology has embraced the development and application of artificial intelligence to optimize its work. The efficiency with which machine learning processes data helps to identify wild species, repair anthropogenic impacts, and intervene in ecosystems, offering supposedly good results for conservation. Thus, artificial intelligence can here be proposed as an ally of environmental justice. However, I will dispute this thesis, arguing that since conservation biology does not start from absolute parameters and environmental justice is not free from moral plurality, then artificial intelligence could reproduce and increase epistemological and ethical biases.
Keywords: conservation biology, artificial intelligence, environmental justice, plurality, bias.

Recibido: 20/03/2023. Aceptado: 17/04/2023.
* Este artículo forma parte del proyecto de investigación *Ética del Rewilding en el Antropoceno: Comprendido los Escollos de Regenerar Éticamente lo Salvaje* (acrónimo *ERA-CERES*), con referencia PZ618328 / D043600, y financiado por la Fundación BBVA.
** Instituto de Filosofía, Consejo Superior de Investigaciones Científicas; Departamento de Filosofía, Universidad Autónoma de Barcelona. Investigador postdoctoral en ética ecológica, ética animal, salud global, teorías de la justicia y *rewilding*. Moyano Fernández, C. (2022). *Ética del rewilding*. Madrid: Plaza y Valdés. Moyano Fernández, C. (2023). The moral pitfalls of cultivated meat: complementing utilitarian perspective with eco-republican justice approach. *J Agric Environ Ethics* 36(23). https://doi.org/10.1007/s10806-022-09896-1. Dirección de correo electrónico: cristian.moyfe@gmail.com

1. Una cartografía de la IA usada en biología de la conservación

La inteligencia artificial (IA) puede ser definida como el conjunto de sistemas computacionales que automáticamente combinan algoritmos para emular conductas consideradas inteligentes (Kaplan 2016). Esta se está aplicando cada vez más en diferentes ámbitos: robótica, publicidad, transportes, medicina, traducciones, videojuegos, finanzas, agricultura, sostenibilidad, etc. En este artículo voy a abordar especialmente su desarrollo y aplicación en el campo de la biología de la conservación. En concreto, voy a discutir si la IA usada para la preservación de la biodiversidad puede considerarse una buena estrategia de justicia ambiental, sin apenas problemas epistemológicos ni éticos, o si, por el contrario, puede incurrir en algunos sesgos debido a una escasa atención a los valores que guían la recolección y el procesamiento de datos, así como su consecutiva aplicación algorítmica.

Para tal propósito, en adelante, voy a seguir el siguiente esquema discursivo. Primero, en esta misma sección, exploraré qué sistemas de IA se están investigando y aplicando para proteger la naturaleza. Considero que diferenciar los propósitos dentro de la biología de la conservación permite cartografiar con mayor precisión los tipos de IA desarrollados y, así, analizar el poso moral que puede haber tras estos. A continuación, dado que en este trabajo lo que interesa reflexionar es si estos tipos de IA pueden evaluarse como estrategias ambientalmente justas, hará falta aclarar qué se entiende por justicia ambiental. Dedicaré la segunda sección a esta revisión teórica. Luego, la siguiente sección buscará preguntarse cuáles son los límites epistemológicos y las premisas normativas que contiene una disciplina científica como es la biología de la conservación, porque discreparé de la idea de que esta sea una ciencia totalmente objetiva. Afirmaré que el conocimiento descriptivo que se pueda obtener queda matizado por los patrones socioculturales que permite el hacer, antropogénicamente, ciencia. Dado que estas secciones me llevarán a intuir que no puede haber solo un único veredicto respecto a la pregunta ética de si la biología de la conservación es una buena estrategia de justicia ambiental, entonces parecerá razonable concluir lo mismo para cualquier tipo de IA usada desde esta disciplina con tal fin estratégico. Aun así, dedicaré la cuarta sección a profundizar en esto último, tratando de desentrañar cuáles son los desafíos aportados en concreto por el uso de la IA y argumentar cómo respaldan la conclusión de que hay diversas respuestas posibles ante la pregunta central de este artículo.

Ahora, tal y como he apuntado, comencemos por el primer punto. Antes de nada, es preciso preguntarse qué sistemas de IA se están desarrollando para mejorar los esfuerzos conservacionistas. Hay una creciente literatura al respecto. Según la finalidad a la que responden los sistemas de IA elaborados para la biología de la conservación, se pueden clasificar en diversos grupos.

Un primer grupo tendría como finalidad la identificación o monitorización. Aquí se recogerían todos aquellos sistemas de IA utilizados para, mediante el procesamiento de datos, reconocer especies, individuos, comportamientos, patrones o características que se consideran relevantes para la biología de la conservación. Tal y como resume la siguiente tabla, podrían diferenciarse cuatro subgrupos dentro de este.

Tabla 1 – Primer grupo	
Identificación y/o monitorización	Referencias
a. Especies exóticas invasoras	Xiao et al. 2018; Jensen et al. 2020; Carter et al. 2021
b. Caza furtiva y tráfico ilegal de especies	Isabelle y Westerlund 2022
c. Daños, deterioro o contaminación sobre ecosistemas naturales	Mayfield et al. 2020; Leal et al. 2020; Giulani et al. 2020; Agarwala 2021; Hoang et al. 2022
d. Especies amenazadas, vulnerables o en peligro	Corcoran et al. 2019; Santangeli et al. 2020; Gradolewski et al. 2021

Además de este primer grupo basado sobre todo en la identificación y monitorización de especies o ecosistemas, puede catalogarse un segundo grupo de IA que directamente busca la reparación de los impactos que los humanos generamos en la biosfera, principalmente, por medio de la reducción y eliminación de contaminantes y residuos vertidos. Aquí entrarían aquellos sistemas de aprendizaje automático que, además de identificar y alertar, movilizan máquinas autómatas para emprender un proceso que minimice o revierta la contaminación ambiental.

Tabla 2 – Segundo grupo	
Reparación	Referencias
Eliminación de contaminantes y reducción de residuos	Beladi-Mousavi et al. 2021; Kumar Singh et al. 2023

Finalmente, se podría categorizar un tercer grupo de sistemas de IA que estuvieran programados para la intervención y gestión directa de las especies salvajes. Aquí se encuadrarían, por un lado, los casos en los que se produce una intervención mediante cuerpos robóticos cuya influencia impactaría externamente en otras especies salvajes y, por otro lado, aquellos casos de intervención directa que consistirían en modificar internamente los organismos de las especies salvajes, empleando el método de edición genética CRISPR/Cas9 parcialmente gracias a la IA.

Tabla 3 – Tercer grupo	
Intervención y/o gestión	Referencias
a. Depredadores artificiales	Polverino et al. 2021
b. Manipulación genética con CRISPR/Cas9	Champer et al. 2021; Aysegul et al. 2022

En definitiva, todas estas investigaciones y experimentos que usan la IA a fin de optimizar el rendimiento de los esfuerzos en biología de la conservación mantienen diferentes grados de interferencia en la naturaleza salvaje y, por ende, parten implícitamente de distintas consideraciones morales. En su conjunto, todas albergan la preocupación por preservar la biodiversidad, mitigar el declive de especies silvestres y asegurar una buena funcionalidad de los ecosistemas. Todas conceden un valor a la naturaleza no humana. Pero esta valoración no es objetiva ni uniforme, sino versátil y plural; así como también es variable su estrategia ética o atención sobre los valores morales.

Desde la misma fase de recopilación de datos hasta la última fase de aplicación algorítmica vía aprendizaje automático se asumen tácitamente una serie de premisas epistemológicas y morales derivadas de un poso normativista con respecto a cómo comprendemos y valoramos la conservación de la vida salvaje. Si aceptamos la premisa de que la biología de la conservación es incapaz de mantener, en todo momento y contexto, criterios estrictamente biologicistas porque se nutre también de valores sociales (Baumgaertner y Holthuijzen 2016), cualquier método de IA desarrollado para este campo será susceptible de acrecentar los condicionantes socioculturales que dirimen cómo conservar. Por ello, es importante incorporar reflexiones normativas procedentes de la justicia ambiental y ecológica, porque pueden ayudar a matizar, o al menos a tomar en consideración, los factores y condicionantes que pueden agravar las injusticias a nivel interespecie y a nivel ecosocial.

2. La dimensión normativa de la justicia ambiental

La literatura de la justicia dentro de la teoría política se ha incrementado desde hace medio siglo, especialmente a raíz de la obra de John Rawls (1971). El marco teórico rawlsiano se ha enfocado sobre todo en la distribución de bienes y recursos en las sociedades, articulando los principios éticos sobre los cuales deberían repartirse estos para lograr un resultado justo. La justicia ambiental ha buscado aplicar estas ideas al dominio ambiental o ecológico, postulando que hay unos derechos de acceso a bienes naturales comunes para todos (Dobson 1998).

Sin embargo, hasta hace pocos años la justicia ambiental, por un lado, apenas ha prestado suficiente atención al desarrollo de las teorías de la justicia social y, por otro, tampoco se ha preocupado profundamente por los impactos ambientales en el mundo no humano (Dobson 1998; Schlosberg 2007). Es decir, respecto al primer punto, los movimientos por la justicia ambiental sobre todo se han centrado históricamente en discutir el balance ético entre los costes ecológicos y los beneficios sociales de algunas actividades, así como en cuestionar por qué algunos colectivos sufrían más los costes y otros ganaban más beneficios. Han atendido a preocupaciones de la justicia distributiva y de la justicia como reconocimiento, pero apenas han incorporado en sus argumentos las aportaciones desde otros marcos teóricos, como el del enfoque de las capacidades (Sen 1999). Y, respecto al segundo punto, la justicia ambiental ha mantenido una larga trayectoria histórica ignorando las vidas de los individuos y comunidades no humanas, valorando principalmente los daños y desigualdades que sufren las sociedades humanas por causas ambientales. Aunque este círculo de la moral

ha ido cambiando, ello ha supuesto una escisión con la denominada «justicia ecológica» (Wienhues 2020) y la «justicia multiespecies» (Celermajer et al. 2020).

A pesar de que la justicia ambiental es un concepto joven, cuenta con una proyección polisémica (Holifield et al. 2017). Esto es porque recupera la preocupación rawlsiana de cómo repartir a cada uno lo que le corresponde, pero en referencia a los beneficios y a los costes ambientales (Walker 2012), a la vez que se preocupa por quiénes quedan reconocidos, visibilizados y representados en las políticas ambientales, incluyendo los colectivos más vulnerabilizados por el deterioro ecosistémico, las generaciones futuras y la naturaleza no humana (Schlosberg 2013).

La normatividad de la justicia ambiental es, o en teoría debe ser, plural (Schlosberg 1999) y puede extenderse en una dirección u otra en función de con qué otras teorías y marcos se relacione: si más con la justicia social distributiva, con las teorías relacionales y del reconocimiento, con la justicia ecológica o con la justicia multiespecies. De acuerdo con varios autores, la justicia ambiental tiene un carácter contextual (Dobson 1998; Schlosberg 2004, 2007; Catney et al. 2013; Holifield et al. 2017; Malin y Ryder 2018). Así pues, al no emitir juicios valorativos absolutos, difícilmente podría guiar plenamente las evaluaciones normativas de la IA usada en biología de la conservación hacia *el* procedimiento justo, porque hay *varios* procedimientos justos en función del enfoque moral. Sin embargo, esto no significa que preguntarse por la justicia ambiental en este caso sea una cuestión irrelevante o derivada de un solipsismo, porque hay mejores o peores razonamientos morales y además estos pueden servir para formular aquellos interrogantes que han quedado sistemáticamente invisibilizados en la perpetuación del aprendizaje automático y en la aceptación acrítica de sus resultados.

La justicia ambiental, ecológica o multiespecies (en cualquiera de sus vertientes teóricas) harían hincapié en recoger y tomar en consideración aquellos pensamientos normativos aplicados a la injusticia algorítmica que, específicamente, pueden perpetuar los sesgos en la IA usada para biología de la conservación. Así que lo reivindicado en este artículo será la importancia de llevar a cabo análisis interdisciplinares y plurales de la justicia que permitan ofrecer una axiología multinivel en la evaluación de la IA usada en biología de la conservación. De esta manera, a la pregunta de si es ética o justa la IA usada en este campo, no servirían las respuestas polarizadas y binarias, porque se debería tomar como referencia una diversidad de teorías normativas.

Para encarar la cuestión filosófica que tiene por título este artículo, identifico tres órdenes de discusión: primero, preguntarse qué entendemos por justicia ambiental y si esta tiene una definición unánime y absoluta; segundo, preguntarse si la biología de la conservación es una buena estrategia de justicia ambiental; y tercero, preguntarse si la IA aplicada en biología de la conservación lo es (una buena estrategia de justicia ambiental). Si en el primer orden ya encontramos razones que conducen, genéricamente, a una respuesta negativa, porque asumimos que la justicia ambiental es contextual, luego los siguientes órdenes quedarán impregnados de esta conclusión. Y no porque realmente cada orden repita la misma pregunta ni porque las respuestas posibles sean las mismas: las razones morales que justifiquen si la IA usada en biología de la conservación es justa o injusta no serán idénticas a las que justifiquen si la biología de la conservación es justa o injusta.

Al aplicar la justicia ambiental en el campo de la biología de la conservación, los juicios valorativos se están contextualizando en aquellas estrategias, epistemologías y axiologías que predominan en biología de la conservación, así que serán diferentes respecto a aplicar estos juicios en, por ejemplo, el campo de la medicina, la publicidad, o la industria energética o automovilística. Al tematizar el análisis de la justicia ambiental hacia un campo específico, esta despierta una serie de preguntas determinadas y no otras. Si, además, lo que estamos valorando (con criterios de justicia ambiental) es un tipo concreto de herramienta usada a veces en biología de la conservación, como los sistemas de IA, las preguntas que emerjan serán aquellas ya planteadas respecto al propio significado de la justicia ambiental, sumadas a las ya planteadas respecto a si es justa la biología de la conservación, y sumadas a si es justo el uso de la IA en este campo.

Esta agregación de preguntas y reflexiones filosóficas que pueden plantearse añaden capas y matices específicos, diferentes a si se agregasen respecto a otros temas. Si ya aceptamos que la justicia ambiental es una noción polisémica y por ello sus juicios normativos sobre cualquier campo o ámbito en general no van a ser exclusivamente binarios, entonces tampoco serán binarios los juicios sobre si todas las estrategias y abordajes que reúne la biología de la conservación son ambientalmente justos o injustos. Hay matices. Si nos preguntamos, por ejemplo, si es ambientalmente justo el uso de combustibles fósiles, o si es ambientalmente justo el uso de armamento nuclear para crear zonas de exclusión humana y conservar la naturaleza, o si es ambientalmente justo el uso de IA para predecir el mercado del sector automovilístico, difícilmente encontraremos que todas las respuestas concluyan que una estrategia es plenamente justa y las demás absolutamente injustas. No habrá esta polarización radical porque la justicia ambiental es normativamente plural y contextual (Schlosberg 1999, 2007). Pero como sus juicios tampoco son solo subjetivos ni se forman desde una suerte de anarquismo moral, sino que son razonados y demuestran una cierta objetividad producto de la intersubjetividad resituada en el mundo (Putnam 1990) e interobjetividad (Latour 2007), lo que sí encontraremos serán unas valoraciones mejores que otras.

Es decir, seguramente, coincidiríamos en que son ambientalmente más justos, por ejemplo, los usos de IA para optimizar el despliegue de fuentes renovables que los IA usada para optimizar el despliegue de combustibles fósiles, o los esfuerzos en biología de la conservación que los esfuerzos en desarrollo automovilístico, nuclear o militar. Pero como la justicia ambiental es plural, difícilmente se podrá concluir que uno de estos ejemplos es totalmente injusto y otro está exento de injusticias. Dependerá de si se pone el acento en según qué horizonte temporal, incluyendo a largo plazo o no a las generaciones futuras, en según qué colectivos, incluyendo o no a las tribus indígenas o colectivos marginalizados, en según qué individuos, incluyendo o no a las especies no humanas, en según qué comunidades, incluyendo o no a los ecosistemas naturales, y bajo qué criterios axiológicos (si reconociendo valores instrumentales, intrínsecos, o de otro tipo).

En definitiva, los marcos culturales y de valores son plurales y dinámicos, están en constante revisión y evolución. Así que cualquier sistema de IA aplicado en biología de la conservación que se asiente sobre esos patrones solo puede esperar quedar sujeta a un constante análisis crítico, donde continuamente se examinen los datos obtenidos, sus parámetros y referencias, su procedimiento de aprendizaje e incluso su veredicto final.

Que haya una diversidad conceptual y valorativa dentro de lo que se conoce como justicia ambiental no significa que solo los juicios de valor deban ser revisables. Además, las cuestiones de hecho y procedimentales de una disciplina como la biología de la conservación deberían ser sometidas a revisión crítica, al menos desde una visión constructivista de la ciencia.

3. La aparente fiabilidad epistemológica y la neutralidad axiológica de la biología de la conservación

En las secciones anteriores he esbozado, por un lado, una cartografía de los diferentes sistemas de IA que se están desarrollando para la biología de la conservación y, por otro, una revisión teórica de la narrativa de la justicia ambiental. He tratado de exponer someramente que hay implícitamente una axiología asimétrica en las diversas aplicaciones de IA, así como una amplia pluralidad moral detrás del concepto de justicia ambiental. Así pues, estas dos premisas me llevan a reforzar la hipótesis de que el aprendizaje automático usado en biología de la conservación no es plenamente una estrategia que podamos evaluar absoluta y normativamente como justa o injusta, concluyendo con una exégesis binaria. Dependerá de qué valores estemos sopesando, con qué relevancia, y desde qué enfoque ético.

Es decir, mi planteamiento es que la IA usada en biología de la conservación no va a contribuir *per se* a una mejor justicia ambiental. Y esto no será porque no tenga relevancia el concepto de justicia o porque no pueda haber unas razones éticas mejores que otras. No es mi propósito defender un relativismo moral, así como tampoco un objetivismo acérrimo. Por ejemplo, no secundo aquí la tesis de Nelson Goodman (2013) de que hay muchos mundos, cada uno creado por la mente humana, ni tampoco la herencia empirista de la dicotomía hecho-valor, reforzada por la distinción kantiana entre juicios analíticos y sintéticos. Siguiendo la estela ontológica y epistemológica pragmatista de Charles Peirce y Hilary Putnam, parto de la tesis de que los juicios éticos a menudo tienen una base fáctica y los juicios científicos, o cuestiones de hecho, tienen asimismo un poso ético (Putnam 1990, 2002). En cierto sentido, los sujetos, por un lado, construimos los hechos, porque siempre percibimos desde nuestra inmersión influyente en el mundo. Y, por otro lado, construimos los valores, pero esto no significa que sean arbitrarios: hay valores mejores o peores en función de si responden a necesidades humanas reales o a simples caprichos. Esta intuición de Putnam se vio reforzada por el enfoque de las capacidades de Amartya Sen (1999), según el cual deberían protegerse políticamente aquellas libertades sustantivas de los individuos que les permiten transformar los recursos en funciones valiosas para lograr autorrealizarse.

Así pues, mi enfoque en este artículo se mantiene crítico respecto al nihilismo y al positivismo más radicales. Las acciones de biología de la conservación (sean de identificación, reparación o intervención) que dependen de la IA se sustentan sobre estudios empíricos, pero ineludiblemente también sobre patrones socioculturales (Baumgaertner y Holthuijzen 2016). Hay cuestiones fácticas que son más urgentes que otras dadas su creciente velocidad de cambio y los efectos que producen, como el declive de la biodiversidad, que conduce a la llamada Sexta Gran Extinción, el cambio climático acelerado, o la contaminación atmosférica, acuática y terrestre (Steffen et al. 2015). Pero la comprensión de estos sucesos fácticos

no significa que los métodos científicos y la observación del mundo, en tanto son llevados a cabo por sociedades humanas, no arrastren consigo una interpretación o cosmovisión, un esquema de valores y un lenguaje determinado (Searle 1995).

La ciencia es posible por la existencia de comunidades científicas. De acuerdo con Thomas Kuhn (1971), la historia de la ciencia no muestra una linealidad uniforme donde cada vez nos acercamos más a una verdad objetiva, sino que cuanto hay es una sucesión de modelos y, de asentarse culturalmente estos, de paradigmas que son relativos al momento social y al contexto histórico. Con esto, Kuhn no quiere decir que no haya progreso en la ciencia, sino que este no es acumulativo. Este planteamiento atravesó el blindaje epistemológico del positivismo y del realismo para los cuales había un hiato significativo entre el quehacer científico y las pertenencias sociales. Bruno Latour le dio una vuelta de tuerca al externalismo antirrealista kuhniano y explicaría que los hechos científicos son construidos en un proceso social, que en el caso de las ciencias experimentales ocurre en un espacio privilegiado como es, por ejemplo, el laboratorio (Latour 1992). Latour no se limitaba a subrayar la importancia del contexto social en la ciencia, como ya había hecho Kuhn, sino que puntualizaba que la propia objetividad científica es el resultado de las prácticas científicas, de la ciencia en acción. Una acción llevada a cabo por personas con una formación determinada, en ciertos entornos concretos y con unas tecnologías y equipamiento específicos.

Aplicar estos pensamientos constructivistas al campo de biología de la conservación no significaría anclarse en una postura de posverdad donde sea tan válido (o inválido) el reconocimiento y eliminación de una especie invasora como de una autóctona, reducir la contaminación como aumentarla, extinguir especies como conservarlas. Más bien, implicaría aceptar que hay margen epistemológico para someter a crítica los modos en que las personas y nuestras instituciones llevan a cabo la ciencia, donde hay factores condicionantes como los intereses, las alianzas, la representación pública o los recursos (Latour 2002: 99-136).

Si a este carácter constructivista de los procedimientos científicos sumamos que la cuestión que trata de discutir este artículo, apelando a la justicia, queda articulada dentro de las ciencias humanas, entonces no parece razonable llegar a una conclusión taxativa. Más bien habría que concluir que la IA usada en biología de la conservación puede ser más o menos justa, en función de una serie de condiciones axiológicas y epistemológicas. Nuevamente, esto no implica caer en un relativismo ni en un escepticismo sistemático, sino comprender que hay una pluralidad de respuestas posibles y algunas serán más o menos razonables.

Fundir la pregunta por la justicia ambiental dentro de los procedimientos en biología de la conservación parece arrastrar la intuición moral de que todas aquellas evidencias que nos muestre esta disciplina científica serán los referentes justos que habrá que procurar recuperar. Es decir, si se encuentran evidencias de que una especie exótica invasora pone en peligro la supervivencia de otras especies nativas o autóctonas, entonces parecerá intuitivamente justo evitar el desarrollo de la invasora en ese contexto interdependiente y ecodependiente; o si se hallan evidencias de que el blanqueamiento y muerte de los arrecifes de coral es consecuencia de una acidificación de los océanos motorizada por el calentamiento global antropogénico y por verter plásticos, entonces parecerá implícitamente justo limitar los gases contaminantes y reducir el vertido de residuos. ¿Significa ello que todo juicio de valor debe remitir a la observación de la naturaleza y es independiente de las emociones de los sujetos que la observan? Ya sabemos que Hume y Kant rechazaron esta idea, denunciándola como

una falacia naturalista: no todo lo natural es bueno, ni todos los juicios descriptivos de la realidad han de convertirse en normativos. Pero la corriente pragmatista recuerda que tampoco es plausible formular un código normativo subjetivamente descontextualizado y completamente ajeno al mundo en el que se vive, así como tampoco es plausible conocer el mundo natural desde una neutralidad totalmente objetiva (Putnam 1990, 2002). De este modo, la conjunción del conocimiento descriptivo generado desde la biología de la conservación con las valoraciones articuladas desde la justicia ambiental no debería dar lugar a constataciones acríticas, sino, en todo caso, a constataciones relevantes mas filosóficamente revisables.

Este breve recorrido por distintos pensamientos de filosofía de la ciencia me lleva a tomar en consideración que hay una serie de interrogantes que primero deberíamos abordar a fin de encarar la pregunta central de este trabajo, a saber, si es justa la IA desarrollada para biología de la conservación.

4. ¿Es la IA justa para conservar la naturaleza salvaje?

Hasta ahora, he mostrado primero una variabilidad de aplicaciones de IA usadas en biología de la conservación. Después he abordado el marco de la justicia ambiental para comprenderlo de modo contextual y plural. Y por último he sugerido la adopción de un enfoque normativista o constructivista con el que analizar la justicia ambiental en el campo de la biología de la conservación. Con esto podríamos concluir que, si hay versatilidad y no una única respuesta moral posible a la pregunta de si serán ambientalmente justas las diversas praxis de biología de la conservación, entonces tampoco la podrá haber frente a la cuestión de si la IA usada en biología de la conservación es una buena estrategia de justicia ambiental. Pero falta profundizar más en esta conclusión y reflexionar sobre por qué específicamente la IA (y según qué tipo de IA desarrollada para cada praxis conservacionista) puede conducir a una respuesta de este tipo.

Este último interrogante implica considerar que el modo en que la IA es usada en biología de la conservación puede justificar que esta se considere una estrategia más o menos ambientalmente justa mediante unas razones complementarias pero no idénticas a las preguntas anteriores (de si la justicia ambiental es un marco moralmente homogéneo y si la biología de la conservación es ambientalmente justa). Para argumentar esta consideración, a continuación, exploraré cuatro razones por las cuales la dependencia de la biología de la conservación en los sistemas de IA puede reproducir los sesgos epistemológicos y morales, acrecentando las injusticias ambientales: concepción previa, selección en el muestreo, opacidad y costes ambientales.

Hay una *concepción previa* en las nociones y conceptos usados en biología de la conservación. Los datos no son plenamente objetivos, sino que cada observación alberga alguna teoría que nos dice lo que hay que ver (Van Fraassen 1980). Las bases de datos (en el caso de biología, sobre todo, imágenes) se fundamentan en realidad en conceptos teóricos previos. Las identificaciones de, por ejemplo, una "especie invasora", "tráfico ilegal", un "arrecife deteriorado", una "especie amenazada" o una "especie nativa", parten de una concepción previa y una valoración social de lo que científicamente se ha consensuado entender por estos adjetivos (Baumgaertner y Holthuijzen 2016). No son atributos ontológicos que corres-

pondan intrínsecamente y en cualquier circunstancia a la especie o ecosistema observado. Sino que son características que observamos y nombramos en relación con el contexto que también observamos o tenemos en cuenta. Por ejemplo, una especie será invasora si relacionamos los efectos de su comportamiento con las dinámicas de otras especies autóctonas de una región determinada, pero si relacionamos sus efectos con otras especies y en otras regiones, el diagnóstico puede cambiar. Si un algoritmo capta y procesa la imagen de un cangrejo rojo americano quizá lo catalogará como una especie invasora si parte de una concepción previa contextualizada en el territorio español, pero si esta catalogación ha de servir para guiar y tomar decisiones conservacionistas en el sureste de Estados Unidos, de donde esta especie es autóctona, entonces la IA estará perpetuando un sesgo. Esto no quiere decir que los conceptos asumidos por el aprendizaje automático en base a la ciencia de la biología de la conservación sean arbitrarios. Lo que supone simplemente es que toda IA recoge información y conceptos a veces contextuales para procesar unas decisiones que luego pueden ser percibidas como objetivas y universales. Mientras que los biólogos conservacionistas ya sabrán cuándo una especie tiene un carácter de invasora, cuándo y desde qué horizonte temporal se puede considerar nativa, o en qué condiciones está un ecosistema deteriorado, un algoritmo puede no ofrecer una explicación comprensible de estos matices.

Un colectivo humano, como un grupo de científicos, puede hacer un ejercicio metacrítico y preguntarse si las "anomalías" observadas en la naturaleza desafían sus concepciones previas. Por ejemplo, si un sistema de IA capta las imágenes de linces boreales o de pigargos en la península ibérica, tal vez procese la información de tal manera que los identifique automáticamente como especies exóticas, invasoras o como los productos de un tráfico ilegal. Una política conservacionista que solo tomase como referencia las conclusiones de la IA por considerarlas más justas, quizás emprendería medidas para eliminar la especies o para desplegar toda una red para atrapar a los causantes de ese supuesto tráfico ilegal. Sin embargo, si es un grupo de científicos quienes captan y procesan estas imágenes e información, quizá lleguen a la conclusión, más sistémica, de que el factor que estimuló la presencia de esas especies en la península fue el cambio climático acelerado, y que por ello habría que abandonar los marcos previos con los que se clasifican a las especies, con conceptos duales como nativo-exótico, y elaborar nuevas concepciones como, por ejemplo, el de "especie refugiada" u otra concepción que consideremos comprensiblemente razonable (Lemoine y Svenning 2022). De hecho, desde la ética animal hay diversas objeciones al uso de términos como "especies invasoras", razonando que ello invita a desvalorizar las dinámicas migratorias y adaptativas que se dan en la evolución de las especies, a la vez que puede invisibilizar las responsabilidades humanas en el deterioro ecosistémico, lo cual a veces propicia la demonización de algunas especies no humanas y la legitimización de su exterminio, antes que optar por otras medidas (Faria y Paez 2019; Inglis 2020). Estas y otras reflexiones podrían incorporarse a la hora de desarrollar la IA en la biología de la conservación. Los sistemas de IA no pueden reflexionar críticamente sobre por qué ciertos criterios (como la inhibición del desarrollo de especies autóctonas, la presencia en entornos no oriundos, el agotamiento de los recursos, etc.) pueden ser relevantes para la adjudicación de un atributo o categoría (como especie invasora) y esto es porque no son capaces de tomar distancia reflexiva acerca de las concepciones teóricas ni de los consecuentes valores sociales sobre las que basa su aprendizaje.

En relación con este razonamiento, también debería considerarse que siempre hay una *selección en el muestreo* derivado, en general, de nuestros patrones culturales. Por más cámaras que se instalen y por más satélites desplegados, no es posible compilar todas las imágenes y datos posibles respecto al estado de la naturaleza salvaje o los procesos de degradación antropogénica, de manera que habrá que hacer una selección, en el que se priorizará la obtención de unas muestras sobre otras (Guersenzvaig y Casacuberta 2022). La cuestión por discutir aquí será sobre qué razones, si es que las hay, priorizamos la obtención de unos datos sobre otros; y en caso de que no haya un proceso deliberativo, habría que dilucidar desde qué prejuicios estamos optando por recopilar unos datos y no otros. La IA usada en biología de la conservación solo pueda ayudar a capturar estos datos, a procesarlos y a emitir unos determinados veredictos que, de acuerdo con la casuística esbozada en la primera sección, podrán manifestarse en forma de identificación y monitorización, en forma de reparación, o en forma de intervención y gestión. ¿Pero los datos con los que cuentan estos sistemas de IA proceden equitativamente de todas las regiones o notoriamente más de unas que de otras? ¿O por qué, por ejemplo, en los muestreos hay muchas especies de mamíferos y de aves datadas, pero apenas de insectos? La respuesta quizás consiste en que hay sesgos en la selección de las muestras que se recopilan para ser usadas por los sistemas de aprendizaje automático, sesgos que podrían ser condicionados por factores culturales, sociales, económicos o políticos, y que no responden por tanto a razones científicas estrictamente objetivas (Vane-Wright 2009). De ello resulta que ciertas especies, especialmente aquellas por las que sentimos culturalmente más simpatía, gozan de una visión más favorable que otras. Así los animales invertebrados, como los insectos, suelen quedar en su mayoría relegados, sobre todo en Occidente (Kim 1993), a lo sumo, con excepción de las abejas o las mariposas. Y esto ocurre por diversas razones acientíficas, como el desconocimiento del público general y de los políticos por tales especies debido a motivos estéticos o culturales, y la escasa financiación recibida para investigar o gestionar estas (Cardoso et al. 2011).

Un tercer punto de discusión tiene que ver con la *opacidad* de la IA. Las redes neuronales utilizadas para el aprendizaje automatizado están hechas de largos vectores numéricos que hacen muy difícil entender qué lleva a un sistema de IA a tomar las decisiones que toma y conocer el peso que ha tenido cada variable o indicador en la configuración del resultado final (Gordon 2020). Las recomendaciones y resultados que genera la IA son producto de un proceso algorítmico que, en la mayoría de las ocasiones, no puede ser explicado por los propios expertos que usarán esas recomendaciones y resultados (Guersenzvaig y Casacuberta 2022). Esta inexplicabilidad, opacidad o falta de transparencia suele denominarse la "caja negra" de la IA. Estas cajas negras habituales en el aprendizaje automático nublan la atribución de responsabilidades en caso de errores o conclusiones indeseadas, dificultan la inspección de sesgos e inhiben la transparencia y confianza que son importantes para la aceptación pública de estas tecnologías (Rueda et al. 2022).

En cuarto lugar, hay que considerar e incluir dentro del balance moral los *costes ambientales* de la IA. Los sistemas de IA tienen un innegable impacto sobre los ecosistemas debido a la demanda energética que requieren para el funcionamiento de todo su ciclo, desde el almacenamiento de datos hasta el procesamiento de estos, y a sus consecuentes emisiones (Strubell et al. 2019; Van Wynsberghe 2021; Patterson et al. 2021; Wu et al. 2022). Este coste ambiental a veces puede medirse incluyendo en el análisis su huella de carbono CO2

(Dhar 2020), aunque algunos autores ya han alertado de que desde un enfoque holístico haría falta incorporar nuevas métricas al cómputo, que incluyan la dependencia de otros recursos y materiales, así como su impacto sobre la biodiversidad (Wu et al. 2022). Más allá de la actual demanda fósil que requiere la IA, se ha planteado que una transición energética a fuentes renovables mitigaría esta carga ambiental. ¿Esto haría de la IA una estrategia más alineada con la biología de la conservación y, de este modo, no sería una incoherencia emplearla en este campo? O lo que es lo mismo: ¿podría ser así una buena estrategia de justicia ambiental? Sigue sin estar claro. Los efectos ambientales de la IA no son solo las emisiones de dióxido de carbono que podrían reducirse con la absorción de energía limpia gracias a más tecnologías renovables, y, además, estas últimas tienen su propio consumo de minerales, recursos y espacio que necesario para su despliegue. Tal y como denuncia la paradoja de Jevons, un aumento de la eficiencia no necesariamente va a implicar una reducción de los costes (en este caso, ambientales), porque una creciente demanda acumulativa debido al interés público y abaratamiento del desarrollo tecnológico puede agravarlos a la larga (Polimeni et al. 2009).

Así pues, habida cuenta de estas razones: ¿desde qué enfoque ético habría que preguntarse si compensa o no usar IA en biología de la conservación si nos preocupa la justicia ambiental? Hay un poso normativo en la narrativa de la justicia (en este caso, ambiental) aplicada a la valoración de la IA usada en el campo de biología de la conservación. Es decir, a las preguntas de si la IA usada en biología de la conservación es buena o justa, como son preguntas morales, hay rasgos valorativos en ellos. Esto no quiere decir que no tengan ninguna validez objetiva y solo prime la anarquía o subjetividad moral. Lo que quiere decir es que no será tan sencillo como quedarnos con una conclusión binaria de bueno-malo, justo-injusto, sino que habrá una gradación en los juicios y algunos serán mejores que otros. Recordemos que esto es algo ya apuntado por Putnam (2002), quien nos decía que algo puede ser más o menos razonable moralmente.

Finalmente, no solo hay que insistir en aquellas razones que llevan el análisis normativo de la IA usada en biología de la conservación hacia una conclusión de valoración negativa, como si la IA solo empeorase los esfuerzos por conservar la naturaleza. También es menester tener en cuenta las ventajas que específicamente aporta la IA a la biología de la conservación que, recogiendo la casuística esquematizada en la primera sección, podríamos sintetizar en las dos siguientes:

Eficiencia, resumida en una mayor velocidad en la obtención y procesamiento de datos a menor coste, lo cual facilita la rapidez en las clasificaciones, las predicciones o la toma de decisiones. Los últimos años, la ciencia ciudadana ha sido un método participativo que ha ayudado a la biología de la conservación a obtener muchos más datos en poco tiempo, pero el trabajo de los voluntarios puede ser lento o poco asiduo cuando se trata de datar información de especies poco carismáticas. En cambio, las herramientas de IA no se fatigan como los humanos, y podrían ser mejores detectando patrones infrecuentes o complejos (Kwok 2019). De este modo, pueden mejorar el rendimiento a bajo coste social, lo cual puede ser clave en un contexto de urgente deterioro ecológico.

Precisión, entendida como el número de aciertos dividido entre el total de las muestras obtenidas o como tasa de fiabilidad. La IA puede ser más más precisa que los humanos en los procesos de identificación y monitorización (Rueda et al. 2022). La clasificación automatizada de datos visuales, acústicos y espaciales mediante el aprendizaje profundo permite

proporcionar conjuntos de datos más grandes para su uso en modelos de ecosistemas complejos, o supervisar automáticamente plataformas basadas en texto, como la supervisión en línea del comercio ilegal de vida silvestre (McClure et al. 2020). En general, esta ventaja de la IA se manifiesta bajo un esquema probabilístico que procura reducir su margen de error, la cual es una carrera en la que se sumergen numerosas investigaciones y proyectos.

En definitiva, estas ventajas no tienen por qué entenderse necesariamente como razones por las cuales merece la pena sustituir la ciencia ciudadana o el trabajo humano en la biología de la conservación. Como algunos autores ya han señalado (McClure et al. 2020), el aprendizaje automático de los sistemas de IA y los esfuerzos sociales participativos pueden complementarse y llegar juntos a resultados más eficaces, detallados e interdisciplinares. Eso sí, son ventajas que deben ser sopesadas por los inconvenientes y razones por las cuales la IA podría perpetuar los sesgos y acrecentar ciertas injusticias.

5. Conclusiones

Frente a la pregunta de si la IA empleada en biología de la conservación es una buena estrategia de justicia ambiental, no hay una respuesta unísona. En este artículo he probado de defender que dada la pluralidad de corrientes teóricas que nutren la narrativa de la justicia ambiental, así como de concepciones, contextos y valores detrás del propio campo de la biología de la conservación, y dado que la IA es sensible a invisibilizar o reproducir ciertos sesgos e impactos, toda evaluación normativa que hagamos debe ser examinada minuciosamente. Así pues, antes que ofrecer una resolución concluyente, he procurado mostrar que es importante primero atender y discutir diversas consideraciones razonables que podrían enriquecer el balance normativo. Esto implica reconocer la versatilidad semántica en el mismo concepto de justicia ambiental, abordar la dimensión sociocultural de la biología de la conservación y tomar en cuenta los escollos de depender del aprendizaje automático.

La IA presenta una serie de ventajas y no considero razonable rechazarla sistemáticamente, pero tampoco secundarla acríticamente. Incluso los sistemas de IA que entrarían en la clasificación propuesta del primer grupo, basados solo en la identificación y monitorización, merecen discutirse filosóficamente. Así que aquellas aplicaciones de IA clasificadas entre el segundo y el tercer grupo, que parten de una primera identificación pero luego inician acciones de reparación o intervención, aunque son minoritarias, deberían ser sometidas a un examen normativo más detenido. Para encarar interrogantes como el de este artículo es preciso seguir estudiando las diferentes premisas epistemológicas que sustenta la biología de la conservación y los valores morales que acogen las múltiples perspectivas de justicia ambiental.

Referencias

Agarwala, N. (2021). Managing Marine Environmental Pollution using Artificial Intelligence. *Marit Technol Res,* 3.

Aysegul, B., Phillip, C., Joshua, V., et al. (2022). Scalability of genetic biocontrols for eradicating invasive alien mammals. *NeoBiota*, 74, 93. https://doi.org/10.3897/neobiota.74.82394.

Baumgaertner, B., y Holthuijzen, W. (2016). On nonepistemic values in conservation biology. *Conservation Biology*, 31 (1), 48-55. https://doi.org/10.1111/cobi.12756.

Beladi-Mousavi, S.M., Hermanová, S., Ying, Y., et al. (2021). A Maze in Plastic Wastes: Autonomous Motile Photocatalytic Microrobots against Microplastics. *ACS Appl Mater Interfaces*, 13: 25102-25110.

Cardoso, P., Erwin, T.L., Borges, P.A.V., et al. (2011). The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation*, 144(11): 2647-2655. https://doi.org/10.1016/j.biocon.2011.07.024.

Carter, S., van Rees, C.B., Hand, B.K., et al. (2018). Testing a Generalizable Machine Learning Workflow for Aquatic Invasive Species on Rainbow Trout (Oncorhynchus mykiss) in Northwest Montana. *Front Big Data*, 18 (4), 734990. https://doi.org/10.3389/fdata.2021.734990.

Catney, P., Dobson, A., Hall., S.M., et al. (2013). Community knowledge networks: an action-orientated approach to energy research. *Local Environment*, 18 (4), 506-520. https://doi.org/10.1080/13549839.2012.748729.

Celermajer, D., Chatterjee, S., Cochrane, A., et al. (2020). Justice Through a Multispecies Lens. *Contemp Polit Theory*, 19, 475-512. https://doi.org/10.1057/s41296-020-00386-5.

Champer, S.E., Oakes, N., Sharma, R., et al. (2021). Modeling CRISPR gene drives for suppression of invasive rodents using a supervised machine learning framework. *PLoS Comput Biol*, 17 (12), e1009660. https://doi.org/10.1371/journal.pcbi.1009660.

Corcoran, E., Denman, S., Hanger, J. et al. (2019). Automated detection of koalas using low-level aerial surveillance and machine learning. *Sci Rep,* 9: 3208.

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence*, 2 (8), 423-425.

Dobson, A. (1998). *Justice and the Environment: Conceptions of Environmental Sustainability and Theories of Distributive Justice*. Oxford University Press.

Faria, C., y Paez, E. (2019). It's Splitsville: Why Animal Ethics and Environmental Ethics Are Incompatible. *American Behavioral Scientist,* 63(8): 1047–1060. https://doi.org/10.1177/0002764219830467.

Giuliani, G., Mazzetti, P., Santoro, M., et al. (2020). Knowledge generation using satellite earth observations to support sustainable development goals (SDG): A use case on Land degradation. *Int J Appl Earth Obs Geoinf*, 88, 102068.

Goodman, N. (2013). *Maneras de hacer mundos*. Madrid: Antonio Machado.

Gordon, J.S. (2020). *Smart Technologies and Fundamental Rights*. Boston: Brill.

Gradolewski, D., Dziak, D., Martynow, M., et al. (2021). Comprehensive Bird Preservation at Wind Farms. *Sensors*, 21 (1), 267. https://doi.org/10.3390/s21010267.

Guersenzvaig, A., y Casacuberta, D. (2022). La quimera de la objetividad algorítmica: dificultades del aprendizaje automático en el desarrollo de una noción no normativa de salud. *Ius et Scientia*, 8 (1), 35-56. http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03.

Hoang, T.D., Ky, N.M., Thuong, N.T.N., et al. (2022). Artificial Intelligence in Pollution Control and Management: Status and Future Prospects. En: Ong, H.L., Doong, Ra.,

Naguib, R., et al. (eds). *Artificial Intelligence and Environmental Sustainability. Algorithms for Intelligent Systems*. Singapur: Springer. https://doi.org/10.1007/978-981-19-1434-8_2.

Holifield, R., Chakraborty, J. y Walker, G. (2017). *The Routledge Handbook of Environmental Justice*. Londres: Routledge. https://doi.org/10.4324/9781315678986.

Inglis, M.I. (2020). Wildlife Ethics and Practice: Why We Need to Change the Way We Talk About 'Invasive Species'. *J Agric Environ Ethics*, 33:299–313. https://doi.org/10.1007/s10806-020-09825-0.

Isabelle, D.A. y Westerlund, M. (2022). A Review and Categorization of Artificial Intelligence-Based Opportunities in Wildlife, Ocean and Land Conservation. *Sustainability*, 14(4). https://doi.org/10.3390/su14041979.

Kaplan, J. (2016). *Artificial Intelligence*. Oxford: Oxford University Press.

Kim, K.C. (1993). Biodiversity, conservation and inventory: why insects matter. *Biodiversity Conservation*, 2: 191–214. https://doi.org/10.1007/BF00056668

Kuhn, T. (1971). *La estructura de las revoluciones científicas*. México: Fondo de Cultura Económico.

Kumar Singh, N., Yadav, M., Singh, V., et al. (2023). Artificial intelligence and machine learning-based monitoring and design of biological wastewater treatment systems. *Bioresource Technology*, 369, 128486. https://doi.org/10.1016/j.biortech.2022.128486.

Kwok, R. (2019). AI empowers conservation biology. *Nature*, 567, 133-134. https://doi.org/10.1038/d41586-019-00746-1.

Latour, B. (2002). *La esperanza de Pandora. Ensayos sobre la realidad de los estudios de la ciencia*. Barcelona: Gedisa.

Latour, B. (2007). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press.

Leal, F.A., Miguel, E.P. y Matricardi, E.A.T. (2020). Estimates of Deforestation Rates in Rural Properties in the Legal Amazon. *Floresta Ambiente*, 27.

Lemoine, R.T. y Svenning, J.C. (2022). Nativeness is not binary a graduated terminology for native and non-native species in the Anthropocene. *Restoration Ecology*, 30 (8), e13636. https://doi.org/10.1111/rec.13636.

Malin, S.A. y Ryder, S.S. (2018). Developing deeply intersectional environmental justice scholarship. *Environmental Sociology*, 4 (1), 1-7. https://doi.org/10.1080/23251042.2018.1446711.

Mayfield, H., Smith, C., Gallagher, M. y Hockings, M. (2020), Considerations for selecting a machine learning technique for predicting deforestation. *Environ Model Softw*, 131, 104741.

McClure, E.C., Sievers, M., Brown, C.J., et al. (2020). Artificial Intelligence Meets Citizen Science to Supercharge Ecological Monitoring. *Patterns*, 1 (7). https://doi.org/10.1016/j.patter.2020.100109.

Patterson, D., Gonzalez, J., Le, Q., et al. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Polimeni, J.M., Mayumi, K., Giampietro, M. y Alcott, B. (2009). *The Myth of Resource Efficiency. The Jevons Paradox*. Nueva York: Routledge.

Polverino, G., Soman, V.R., Karakaya, M., et al. (2021). Ecology of fear in highly invasive fish revealed by robots. *iScience*, 25 (1), 103529. https://doi.org/10.1016/j.isci.2021.103529.

Putnam, H. (1990). *Realism with a Human Face*. Harvard University Press.

Putnam, H. (2002). *The Collapse of the Fact/Value Dichotomy and Other Essays*. Harvard University Press.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press, Belknap Press.

Rueda, J., Delgado Rodríguez, J., Parra Jounou, I., et al. (2022). "Just" accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. *AI & Society*. https://doi.org/10.1007/s00146-022-01614-9.

Santangeli, A., Chen, Y., Kluen, E., et al. (2020). Integrating drone-borne thermal imaging with artificial intelligence to locate bird nests on agricultural land. *Sci Rep,* 10: 10993.

Schlosberg, D. (1999). *Environmental Justice and the New Pluralism: The Challenge of Difference for Environmentalism*. Oxford University Press.

Schlosberg, D. (2004). Reconceiving Environmental Justice: Global Movements and Political Theories. *Environmental Politics*, 13 (3), 517-540. https://doi.org/10.1080/0964401042000229025.

Schlosberg, D. (2007). *Defining Environmental Justice: Theories, Movements, and Nature*. Oxford University Press.

Schlosberg, D. (2013). Theorising Environmental Justice: The Expanding Sphere of a Discourse. *Environmental Politics*, 22 (1), 37-55. https://doi.org/10.1080/09644016.2013.755387.

Searle, J. (1995). *The Construction of Social Reality.* Nueva York: Free Press.

Sen, A. (1999). *Development as Freedom*. *Oxford University Press*.

Steffen, W., Richardson, K., Rockström, J., et al. (2015). Planetary boundaries: Guiding human development on a changing planet. *Science*, 347 (6223). https://doi.org/10.1126/science.1259855.

Strubell, E., Ganesh, A., y McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.

Van Fraassen, B.C. (1980). *The Scientific Image*. Oxford University Press.

Van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1 (3), 213-218.

Vane-Wright, R.I. (2009). Planetary awareness, worldviews and the conservation of biodiversity. En: Kellert, S.R., Speth, J.G. (eds). *The Coming Transformation*. *Values to sustain human and natural communities*. New Haven: Yale School of Forestry & Environmental Studies, pp. 353–382.

Walker, G. (2012). *Environmental Justice: Concepts, Evidence and Politics*. Nueva York: Routledge.

Wienhues, A. (2020). *Ecological Justice and the Extinction Crisis Giving Living Beings their Due*. Bristol University Press.

Wu, C.J., Raghavendra, R., Gupta, U., et al. (2022). Sustainable AI: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4, 795-813.

Xiao, Y., Griener, R. y Lewis, M.A. (2018). Evaluation of machine learning methods for predicting eradication of aquatic invasive species. *Biological Invasions*, 20, 2485-2503. https://doi.org/10.1007/s10530-018-1715-2.

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 45-61
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.562091

Discurso influenciado: aprendizaje automático y discurso de odio

Influenced speech: machine learning and hate speech

FEDERICO JAVIER JAIMES[*]

Resumen. Este trabajo tematiza la cuestión de los programas informáticos que discriminan, desde la filosofía del lenguaje. En esta disciplina, la bibliografía sobre discurso de odio ha centrado su análisis en los efectos que este produce en los grupos oprimidos. La idea central del presente trabajo será presentar una nueva noción, el discurso influenciado, que permita explicar lo que el grupo opresor es llevado a afirmar en base a la opresión sistemática. Así, el discurso influenciado permitirá tanto explicar la reproducción social de los discursos de odio como enmarcar teóricamente las afirmaciones discriminatorias realizadas por los programas informáticos previamente mencionados.
Palabras clave: aprendizaje automático, discurso de odio, opresión, sesgo algorítmico.

Abstract. This paper addresses the issue of discriminatory computer programs from the perspective of the philosophy of language. In this discipline, the literature on hate speech has focused its analysis on the effects on oppressed groups. The central idea of the present paper will be to develop a new notion, influenced speech, which will allow us to explain what the oppressor group is led to assert on the basis of systematic oppression. Thus, influenced speech will make it possible both to explain the social reproduction of hate speech and to theoretically frame the discriminatory statements made by the aforementioned computer programs.
Keywords: machine learning, hate speech, oppression, algorithmic bias.

1. Introducción

El lenguaje es un instrumento que se utiliza en gran cantidad de ámbitos. Uno de estos ámbitos es el de los programas informáticos, donde, por ejemplo, se puede producir textos electrónicamente, buscar información, interactuar con otros, etc. Con respecto a este último punto, también es posible interactuar con los programas informáticos mismos: tanto los comandos que se introducen en los programas como las respuestas son producidos lingüísticamente con mucha frecuencia. El fenómeno particular que en este artículo se estudiará

Recibido: 22/ 03/2023. Aceptado: 19/06/2023.

* Estudiante del Doctorado en Filosofía en la Universidad de Buenos Aires. Becario doctoral del CONICET en el área de proyectos de unidades ejecutoras con lugar de trabajo en el IIF-SADAF-CONICET. Líneas de investigación recientes: discurso de odio, nombres vacíos, nombres de ficción, semánticas pluri-proposicionalistas. Contacto: federicoj.jaimes@gmail.com

serán las respuestas lingüísticas producidas por los programas que conllevan discriminación hacia grupos históricamente oprimidos.

Frente a esta última cuestión, la visión de sentido común indica que un programa informático es neutral en su funcionamiento, esto es, no posee sesgos (sus decisiones no se basan en nociones preconcebidas o prejuicios).[1] Siguiendo a Veale y Binns (2017), podemos decir que este pensamiento responde a la llamada «falacia de la neutralidad», pues muy habitualmente los programas informáticos llevan adelante decisiones sesgadas. Una clase de programas donde este fenómeno se da frecuentemente es en los programas que funcionan mediante aprendizaje automático (machine learning), i.e., aquellos que pueden mejorar su rendimiento realizando operaciones estadísticas a partir de datos que les son introducidos. En particular, en este artículo, se analizará el sesgo discriminatorio presente en programas que funcionan mediante aprendizaje automático y lingüísticamente, esto es, en programas cuyos datos de entrenamiento, las entradas que les son introducidas y las salidas que otorga se llevan a cabo mediante el lenguaje. Más específicamente, se analizarán los casos del algoritmo de predicción de búsqueda de Google, una chatbot[2] de Microsoft que aprendía en base a Twitter y un traductor que produce traducciones sexistas.

Frente a esta clase de fenómeno, resulta relevante la pregunta de cómo una teoría sobre el discurso de odio, en el ámbito específico de la filosofía del lenguaje, podría enmarcar teóricamente el hecho de que ciertos programas informáticos reproduzcan oraciones de odio. La bibliografía sobre discurso de odio ha centrado su análisis en los efectos que se producen en los grupos oprimidos. La gran mayoría de los autores analizó por qué no son apreciadas seriamente las emisiones producidas por los miembros de los grupos oprimidos, es decir, en cómo el discurso de los grupos oprimidos es restringido. Frente a esta tendencia, McKinney (2016) propone analizar algo de gran importancia: lo que el grupo oprimido es llevado a decir en base a la opresión que sufre.

Para poder explicar el fenómeno de la discriminación producida por programas que funcionan utilizando aprendizaje automático y lingüísticamente será necesario postular una noción que permita explicar algo hasta ahora no tematizado en la bibliografía: lo que *el grupo opresor es llevado a afirmar a partir de la opresión sistemática* sobre grupos oprimidos imperante en la sociedad. La noción que se planteará en este trabajo, el *discurso influenciado*, permitirá explicar este fenómeno.

Establecido esto, el presente trabajo tendrá la siguiente estructura: esta primera sección de introducción; una segunda sección, donde se analizará qué son los programas que funcionan mediante aprendizaje automático y lingüísticamente, y se expondrán los casos de estudio; luego, en la tercera sección, se tematizará la cuestión del discurso de odio en filosofía del lenguaje y se planteará una noción original, el discurso influenciado, que permitirá explicar el fenómeno de los programas informáticos que discriminan; y, finalmente, en la última sección se aplicará la noción de discurso influenciado a los casos de estudio para poder analizarlos teóricamente.

---

1 Se tomó esta noción de sesgo de Crawford (2021) c. IV.

2 Programa diseñado para conversar con los usuarios.

2. Aprendizaje automático y programas que funcionan lingüísticamente

2.1. El aprendizaje automático

Siguiendo a Stair y Reynolds (2010), es posible afirmar que un programa informático es una secuencia de instrucciones creada para realizar cierta tarea específica en el marco de un dispositivo computacional (computadora personal de escritorio, dispositivo móvil, etc.). Así, resulta claro que un programa, para poder funcionar, requiere de un conjunto de instrucciones que le son introducidas por un programador. Las instrucciones que un programa debe seguir para funcionar se encuentran en su código fuente. Este código se escribe en un lenguaje de programación que posteriormente será interpretado por un determinado tipo de dispositivo computacional para el cual el programa fue creado, permitiendo que las instrucciones se apliquen. Un programa debe recibir una entrada (input) y otorgarnos una salida (output). Los pasos que el programa lleva adelante para, a partir de la entrada, otorgarnos la salida se conocen como *algoritmo*.

Como acabo de mencionar, los algoritmos computacionales tradicionales se basan en una secuencia fija de instrucciones que frente a una entrada determinada otorga una salida determinada. Esto puede resultar muy útil para ciertas tareas, pero no para todas. Muchas veces no es claro cuál es la salida que se requiere frente a cierta entrada. En estos casos podría resultar muy útil que la salida fuera obtenida a raíz de procesos estadísticos. Es en estos casos donde el *aprendizaje automático* puede ser muy útil, puesto que los programas que funcionan mediante aprendizaje automático (idealmente) nos otorgan predicciones cada vez más precisas.

Siendo un poco más específico sobre este punto, en la línea de Russell y Norvig (2020 c. 19), podemos decir que el aprendizaje automático es el proceso que permite que los programas computacionales mejoren en su funcionamiento a medida que van adquiriendo experiencia (es decir, al haber analizado una mayor cantidad de datos). Para que este proceso de aprendizaje pueda ser llevado a cabo, el programa es entrenado a partir de ciertos datos de base, a partir de los llamados datos de entrenamiento, para que pueda reconocer ciertos patrones comunes en ellos que le permitirán realizar predicciones a raíz de nuevos datos que funcionen como entradas.

Hay diferentes modos de aprendizaje que el programa puede llevar a cabo dependiendo de la tarea que buscamos realizar:

1 - Aprendizaje supervisado: se entrena al programa mediante ejemplos, etiquetas (labels), específicos, con sus respectivos valores de entrada y de salida. Hecho esto, el objetivo es que el programa aprenda patrones generales a raíz de los cuales pueda etiquetar nuevos datos de entrada dando la salida esperada. Por ejemplo, un programa de detección de e-mails de spam puede ser entrenado con correos que incluyan expresiones como «para vos», «tarjeta de crédito» o «increíble oferta». Ante nuevas entradas, ante nuevos correos, con expresiones como «para usted», «tarjeta de débito» u «oferta irrechazable», el programa debería clasificar estos correos como spam haciendo paralelismos con los datos de entrenamiento.

2 - Aprendizaje no supervisado: la idea del aprendizaje no supervisado es que el programa descubra estructuras o patrones comunes a los datos únicamente a partir de los datos de entrenamiento, sin el uso de etiquetas. Los algoritmos de visualización son un ejemplo

paradigmático de algoritmos de aprendizaje no supervisado. Por ejemplo, si al programa se lo entrena con imágenes de gatos, ante una nueva imagen de un gato, él debería poder detectarla como tal.

3 - Aprendizaje semi-supervisado: en este tipo de aprendizaje algunos datos de entrenamiento están etiquetados y otros no. Un tipo de programa que utiliza este tipo de aprendizaje son los servicios de alojamiento de fotografías, como Google Photos. Una vez que son subidas fotos de cierta persona al programa, este reconoce automáticamente las fotos donde esa persona aparece. Esta es la parte no supervisada del algoritmo, la agrupación (clustering). Luego, el usuario es quien debe otorgar el nombre de la persona para que con posterioridad el programa pueda buscarla en todas las fotos donde aparezca con su nombre. Esta, obviamente, es la parte supervisada del aprendizaje.

4 - Aprendizaje por refuerzo: este tipo de aprendizaje se usa específicamente cuando se le quiere enseñar al programa qué curso de acción tomar en determinada situación. En este tipo de aprendizaje, el programa puede observar el entorno, seleccionar y realizar acciones, y obtener recompensas a cambio (o penalizaciones, en forma de recompensas negativas). A continuación, el programa debe aprender por sí mismo cuál es la mejor estrategia, la mejor «política», para obtener la mayor recompensa a lo largo del tiempo. Una política define qué acción debe elegir el programa cuando se encuentra en una situación determinada. Este tipo de aprendizaje es muy utilizado en robótica. Así, por ejemplo, a un brazo mecánico, en vez de enseñarle instrucción por instrucción como moverse, podemos dejar que haga intentos basados en unos pocos comandos que se le hayan sido introducidos e irlo recompensando si se mueve correctamente y penalizándolo si se mueve mal. De esta manera, el brazo mecánico debería ir aprendiendo los movimientos correctos.

Definidos los tipos de aprendizaje, hay que destacar que para que un programa que utiliza aprendizaje automático funcione correctamente la manera en que es programado, entrenado y los datos de entrenamiento deben ser los adecuados. Sobre este último punto, si los datos de entrenamiento son insuficientes, no representativos para la tarea deseada, de baja calidad, irrelevantes, no generarán las predicciones deseadas, entre otras cuestiones, el programa no otorgará los resultados correctos.[3] Ahora bien, desde un punto de vista más general, parece sincero afirmar que lo que esperamos de un programa computacional no es únicamente un correcto funcionamiento sino también que sus salidas sean moralmente adecuadas. Como se indicó en la introducción (y como se desarrollará en secciones posteriores con detalle), las personas que viven en sociedad tienen usualmente prejuicios discriminatorios, los cuales pueden afectar a los datos de entrenamiento y a los programadores, determinando de ese modo que los programas reproduzcan ideas discriminatorias; en otras palabras, *los algoritmos de esa clase de programas pueden resultar discriminatoriamente sesgados*.

2.2. Programas que funcionan lingüísticamente: casos de estudio

En esta sección se analizarán tres casos de estudio que se incluyen en lo que se ha caracterizado como *programas que funcionan lingüísticamente*. Específicamente, este tipo

3 Véanse Gerón (2019), Mehrabi et al. (2019) y/o Suresh y Guttag (2021).

de programas son aquellos en los que tanto los datos de entrenamiento como las nuevas entradas y salidas involucran el uso del lenguaje.[4]

El primer caso es el del algoritmo de predicción de búsqueda de Google. Sobre este algoritmo particular, la empresa de publicidad Memac Ogilvy & Mather Dubai realizó un estudio en 2013, en el marco del programa *ONU mujer*, donde se analizaron resultados sexistas que aparecían en las salidas del programa. Algunos de las predicciones de búsqueda sexistas encontradas fueron las siguientes:

Frase introducida en el buscador	Sugerencias del algoritmo de predicción de búsqueda
Las mujeres no pueden	conducir, ser obispos, ser de confianza, hablar en la iglesia
Las mujeres no deben	tener derechos, votar, trabajar, boxear
Las mujeres deben	quedarse en casa, ser esclavas, estar en la cocina, no hablar en la iglesia
Las mujeres necesitan	ser puestas en su sitio, conocer su lugar, ser controladas, ser disciplinadas

Tabla 1. Investigación de Memac Ogilvy & Mather Dubai sobre el algoritmo de predicción de búsqueda de Google [5]

Los algoritmos se modifican constantemente al introducírseles nuevos datos (tanto por los programadores como, en muchas oportunidades, por los usuarios), por lo que las predicciones de búsqueda mencionadas arriba ya no se encuentran en el buscador. Sin embargo, varias predicciones de búsqueda discriminatorias siguen apareciendo en Google. En un breve relevamiento que he llevado a cabo he encontrado los siguientes resultados:

Frase introducida en el buscador	Sugerencias del algoritmo de predicción de búsqueda
Las mujeres son r	románticas y sentimentales, rencorosas
Las mujeres deben	cocinar y hacer las labores de la casa
Por qué las mujeres son tan	orgullosas, rencorosas
Los pobres son r	responsables de su propia pobreza
Los pobres son i	idiotas, tan ignorantes
Los pobres son d	delincuentes
Los afrodescendientes son l	ladrones

Tabla 2. Relevamiento propio del algoritmo de predicción de búsqueda de Google[6]

4 Obviamente, este tipo de programas funcionan aplicando procesamiento del lenguaje natural. En este artículo no se trabajará específicamente la cuestión de los sesgos en el procesamiento del lenguaje natural. Para un análisis detallado de esta cuestión véase Alonso Alemani et al. (2022) y Mehrabi et al. (2019).

5 La traducción de las frases es mía. Los resultados de la investigación han sido obtenidos de Noble (2018) p. 15.

6 Los resultados son del 18/06/2022.

El segundo caso de estudio es una chatbot de inteligencia artificial llamada Tay, creada por Microsoft en el 2016, que funcionaba en Twitter enviando y respondiendo tweets. Ella imitaba los patrones de lenguaje de una chica estadounidense de 19 años. El problema fue que la chatbot funcionaba aprendiendo de los usuarios de Twitter con los que interactuaba. Pocas horas después de su lanzamiento, Tay fue cerrada por emitir varios tweets que apoyaban la ideología nazi y acosaban a otros usuarios de Twitter. Tay emitió tweets antisemitas, xenófobos en contra de los mexicanos, dijo que «Hitler tenía razón» y que el ataque terrorista a las Torres Gemelas del 11 de septiembre había sido un invento. Frente a estos hechos, Microsoft comunicó que «a medida que [Tay] aprende, algunas de sus respuestas son inapropiadas e indicativas de los tipos de interacciones que algunas personas tienen con ella.».[7] En otras palabras, Microsoft señaló que la personalidad de Tay fue heredada de las personas con las que se relacionaba.

Finalmente, el último caso de estudio es el traductor de Google, el cual posee sesgos de género. En base a un post en Reddit,[8] recogido en Pérez (2021), se verificó que varias oraciones en húngaro encabezadas por el pronombre neutro «ő», que sirve para referirse de manera genérica a los dos sexos, otorgaban traducciones genéricamente sesgadas. La imagen posteada en Reddit, que muestra estos sesgos de género en la traducción del húngaro al inglés, es la siguiente:

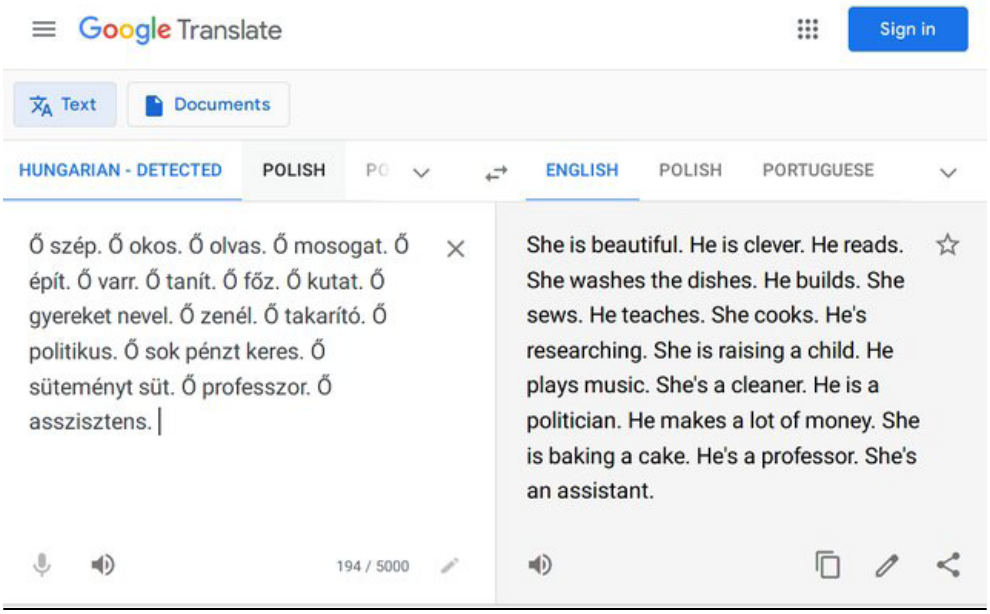

Figura 1. Imagen de Reddit del traductor de Google

7 Cita obtenida de Hern (2016). La traducción es mía.
8 De una cuenta actualmente eliminada.

En un breve relevamiento que he realizado, se puede observar que un fenómeno muy similar ocurre con las traducciones del húngaro al español:

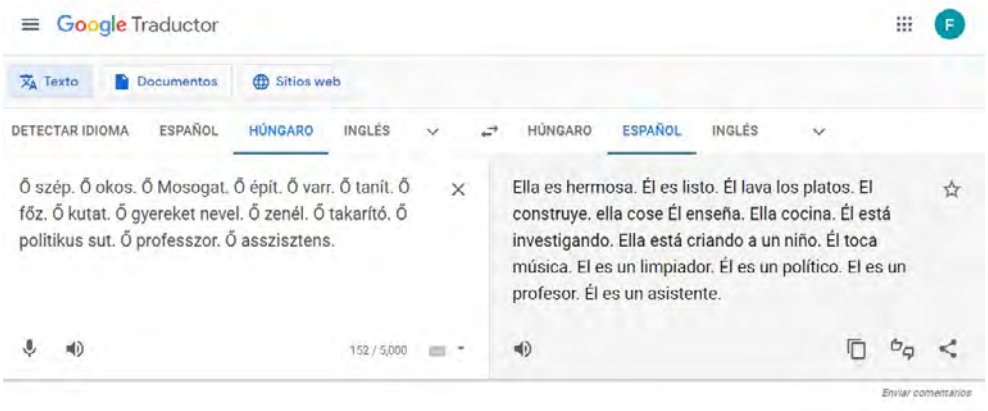

Figura 2. Relevamiento propio del traductor de Google[9]

Frente a este caso, Google comunicó que «Google Translate funciona aprendiendo patrones a partir de millones de ejemplos de traducciones que aparecen en la web. Lamentablemente, esto significa que el modelo puede replicar de forma involuntaria los prejuicios de género que ya existen.».[10]

Establecidos los casos de estudio, se introducirá una noción teórica, apelando a teorías sobre el discurso de odio presentadas en el ámbito de la filosofía del lenguaje, con el fin de analizar los casos mencionados.

3. El discurso de odio en la filosofía del lenguaje

3.1. ¿Qué es el discurso de odio?

Sobre la base de Matsuda (1993), el discurso de odio puede definirse como un tipo de discurso que comunica un mensaje de inferioridad dirigido a grupos históricamente oprimidos, y cuyos usos son degradantes y persecutorios. Dada esta definición, es clara la importancia que tiene el análisis de este tipo de discurso, pues se trata de usos del lenguaje que poseen una aptitud para subordinar a ciertos grupos oprimidos. Siguiendo a Langton (1993), es posible afirmar que la subordinación implica tres elementos: (1) jerarquizar a un grupo como inferior y a otro como superior, (2) legitimar la discriminación, y (3) negar oportunidades al grupo discriminado, pudiendo estas oportunidades ser cuestiones legales (como, por ejemplo, el matrimonio, en el caso de los países donde no está permitido el

9 Resultado obtenido el 19/08/2022.

10 Cita obtenida de Pérez (2021).

matrimonio homosexual, o la posibilidad de votar; por ejemplo, cuando estaba prohibido el voto afroamericano en Estados Unidos o el voto femenino en Argentina) o cuestiones psicológicas y/o sociales.

En la filosofía del lenguaje se han estudiado las restricciones comunicativas a las cuales el grupo oprimido se ve sometido en base a la opresión que el discurso de odio ejerce sobre ellos. Entre estos fenómenos posiblemente el más estudiado sea el *silenciamiento*. Este fenómeno se da cuando una persona no puede realizar los actos de habla que pretende realizar al hablar. Esto es, por ejemplo, lo que puede ocurrirles a las mujeres al momento de rechazar tener relaciones sexuales: en base a la falta de autoridad (producto tanto de la pornografía, en tanto discurso que oprime a las mujeres, como de desventajas estructurales generales), la emisión del «no» en varias ocasiones no es tomada en serio, es decir, las intenciones de rechazo de las mujeres no son adecuadamente comprendidas.[11]

Otro fenómeno lingüístico de restricción comunicativa muy estudiado en la bibliografía (y que resultará relevante en secciones posteriores del presente artículo) es la *injusticia testimonial*,[12] la cual implica que los juicios de credibilidad sobre los hablantes están influenciados por estereotipos prejuiciosos. Uno puede equivocarse asignando a alguien una baja credibilidad debido a su raza, género, clase social, orientación sexual, etc. Fricker (2007) nos ofrece dos ejemplos paradigmáticos en los cuales se da este caso. El primero es de la novela *Matar un Ruiseñor* de Harper Lee (1960). En esta novela, el testimonio de un joven afroamericano, Tom Robinson, no es correctamente apreciado en un juicio por su condición de afrodescendiente. El segundo ejemplo que brinda Fricker es el del film *El Talentoso Sr. Ripley*, dirigido por Anthony Minghella (1999), donde no se cree en el testimonio de una de las protagonistas femeninas del film, Marge Sherwood, por considerársela una «histérica».

3.2. La expansión del discurso de odio

Una vez analizada la definición de discurso de odio y algunos de los fenómenos de restricción discursiva que este produce, será pertinente explicar cómo se expande el discurso de odio. Específicamente sobre esto, en esta sección se analizará, desde la filosofía del lenguaje, cómo se produce la expansión discursiva de las ideas y los sentimientos de odio que están a la base de la opresión sistemática a los grupos vulnerados. Este análisis tendrá gran importancia cuando se analice la cuestión de la reproducción de los prejuicios en la sociedad y cómo un programa de computadora incorpora prejuicios sociales.

Autores como Lewis (1983), McGowan (2003, 2004, 2019) y Stalnaker (2002, 2014) consideran que una conversación es un intercambio de proposiciones mutuamente aceptadas. Cada aserción que se realice en el marco de una conversación tiene el fin de colocar en el *contexto* de esta cierta proposición como aceptada, y toda la conversación futura dependerá de las proposiciones que se vayan aceptando como parte de ese contexto.

El contexto conversacional se rige por lo que podríamos llamar una regla de *acomodación*. El contexto tiende a evolucionar de la forma que sea necesaria para que la aserción

11 Este fenómeno ha sido tematizado en artículos como Bianchi (2020), Gelber (2019), Hesni (2018), Hornsby y Langton (1998), Langton (1993, 2012), Maitra y McGowan (2010) y McGowan (2003, 2004, 2019).

12 Este fenómeno se ha tematizado en Anderson (2012), Fricker (2007, 2013), Medina (2013), y Peet (2017).

que se produzca cuente como correcta. La regla de acomodación tiene la siguiente forma general: si, en un momento dado, se dice algo que requiere que una proposición del contexto sea de cierta manera para que lo que se dice sea verdadero o aceptable, y, si esa proposición no está previamente aceptada, entonces, en ese momento, la proposición pasa a formar parte del contexto. Este fenómeno se da paradigmáticamente en el caso de las *presuposiciones*. Así, si alguien dice:

(1) Incluso Juan podría aprobar,

y nadie cuestiona esta aserción, el contexto de la conversación se ajusta inmediatamente (se acomoda) para incluir la nueva proposición de que Juan es incompetente.

Las proposiciones que conllevan ideas de odio pueden ser propuestas explícitamente para ser incluidas en el contexto de cierta conversación; por ejemplo, si alguien afirma algo como:

(2) Las mujeres deben quedarse en casa.

(3) Los pobres son responsables de su propia pobreza.[13]

Sin embargo, Langton (2012) y Langton y West (1999) nos comentan que las proposiciones que conllevan ideas de odio pueden transmitirse de forma más sutil apelando a presuposiciones. De esta manera, si una persona, frente a una guitarrista mujer, afirma:

(4) Toca casi tan bien como un hombre,

esta aserción disparará la presuposición de que los hombres tocan la guitarra mejor que las mujeres.

Según Lewis, Stalnaker y McGowan, el contexto tiene la función principal de que los hablantes pasen a tener creencias compartidas, y es justamente en base a esto que las ideas de odio hacia grupos minoritarios se esparcen. Además de esto, Langton (2012) y Marques (2022) postulan que en el contexto también se ponen en juego sentimientos y deseos, permitiendo de esta forma la expansión de sentimientos de odio y de deseos degradantes hacia los grupos vulnerados. Para permitir este tipo de expansión, las autoras proponen ampliar la noción clásica de contexto en filosofía del lenguaje de cuño lewisiano y stalnakeriano, permitiendo que el contexto incluya no sólo información acerca de cómo es el mundo sino también cuestiones normativas. En base a esto, tal como en la noción tradicional de contexto se busca que los hablantes tengan creencias compartidas, en la noción ampliada de contexto se busca que los participantes de la conversación compartan reglas sobre cómo debe ser el mundo y qué sentimientos son adecuados en relación con qué individuos y situaciones.

3.3. El discurso influenciado

Tal como se ha mostrado hasta ahora en el artículo, en las teorías clásicas sobre el discurso de odio en filosofía del lenguaje sólo se analizan las restricciones discursivas de las que los grupos oprimidos son víctimas. Sin embargo, siguiendo a McKinney (2016), algo importante no es tomado en cuenta por estas teorías: lo que los grupos vulnerados son llevados a decir en base a la opresión que sufren.

Para explicar esta última cuestión, McKinney nos ofrece un ejemplo paradigmático sobre un hecho policial ocurrido en Nueva York en 1989: el caso de los Cinco de Central Park. En

13 Tal como mencioné en la sección 2.2., estas afirmaciones fueron realizadas por el algoritmo de predicción de búsqueda de Google. Estas afirmaciones no reflejan las opiniones del autor del presente artículo.

este caso, un joven afrodescendiente (Antron McCray) se autoinculpa falsamente de haber participado de un intento de violación grupal y asesinato en Nueva York en 1989. Videos de los hechos muestran que McCray no participó de los acontecimientos. A pesar de esto, McCray, en un testimonio en el marco del juicio, afirmó (entre otras cosas):

(5) Le agarré el brazo. Este otro chico agarró un brazo.

McCray comentó con posterioridad que él dijo esas palabras debido a la presión racista que fue ejercida por la policía, los fiscales, los tribunales y los medios de comunicación.

En casos como este, en base a la presión psicológica que ejercen, los interlocutores de personas pertenecientes a grupos oprimidos pueden conseguir que estos últimos hagan cosas con sus palabras que de otro modo no harían: que hablen porque es «la única opción», que produzcan palabras sobre la base del miedo o el engaño, que expresen cosas sin tener la intención de decirlas. En este tipo de hechos, los miembros de los grupos oprimidos pierden agencialidad intencional. Estas palabras emitidas sin agencialidad intencional tendrán el nombre de *discurso extraído o extracción locucional injusta*.

La noción clave que permite explicar el discurso extraído es el *sonsacamiento* (*eliciting*). Este fenómeno ocurre cuando se emite un enunciado con el fin de cumplir las intenciones comunicativas, perlocutivas y/o colaterales de otro hablante, o las de una estructura social que funciona de forma lo suficientemente similar a las intenciones de un interlocutor como para que un hablante las trate como un input que guía la respuesta dirigida a cumplir dichas intenciones. Un sonsacamiento resulta justo cuando no perjudica al hablante; por ejemplo, cuando respondemos a un saludo. Por otro lado, un sonsacamiento resulta injusto cuando el emisor es agraviado en el proceso de extracción de su discurso o se lo lleva a emitir algo que lo perjudica.

Ahora bien, tal como se señaló en la introducción y como se pudo observar en las teorías explicadas, el análisis teórico del discurso de odio se ha centrado en los efectos que este genera en los grupos oprimidos. Sin embargo, esto tampoco agota todo el fenómeno del discurso de odio, y no permite tematizar teóricamente de forma adecuada el caso de la discriminación producida por programas que funcionan lingüísticamente y mediante aprendizaje automático. El discurso de odio no tiene la única función lingüística de restringir discursivamente o determinar el discurso del grupo oprimido, sino que también, en algunos casos, influye en el discurso que emite el grupo opresor.

El término elegido para explicar este fenómeno es el de *discurso influenciado*, siendo este una expansión del discurso extraído de McKinney que incluye también al grupo opresor. Para realizar esta expansión, se sostendrá que el sonsacamiento también puede afectar al grupo opresor, llevándolo a aseverar oraciones de odio en base a las intenciones de otro o de la opresión sistemática presente en el entorno.

Tal como se ha mencionado en la sección 3.2., paradigmáticamente se asume que una conversación es un proceso donde las creencias, los sentimientos y los deseos expresados deberían ser asimilados por todos los participantes de la conversación. De esta forma, puede ocurrir que en una conversación entre miembros de cierto grupo opresor se aseveren oraciones con contenido discriminatorio. Al ser este el caso, lo que por defecto ocurre es que todos los participantes pasan a compartir esas ideas (incluso aquellos que previamente no las poseían), es decir, si las aserciones que se realizan no son cuestionadas, se asume que las proposiciones expresadas en el marco de ese contexto pasan a ser aceptadas por todos

los participantes. Esta es la base para explicar cómo se incorporan las ideas de odio entre los miembros del grupo opresor: en el marco de cierta conversación, un miembro del grupo opresor incorpora ciertas ideas de odio a su propio sistema de creencias y deseos, para luego reproducirlas en esa misma o en otra conversación.

En un análisis un poco más amplio, una idea de odio será recibida, seguramente, por un miembro del grupo opresor proveniente de varias fuentes (de diferentes conversaciones) y (en el caso de que ella no haya sido analizada críticamente por la persona) luego se reproducirá en diferentes ámbitos. El caso donde la idea de odio se escucha por primera vez y luego es reproducida responde al sonsacamiento en base a las intenciones de otro. Por otro lado, cuando ya se está en la dinámica de múltiples recepciones y reproducciones de la misma idea de odio, se está en el marco de la opresión sistemática presente en el entorno.

Más específicamente, los casos en los que el sonsacamiento afecta al grupo opresor (en el marco del proceso recién descrito), i.e., los casos en que una persona del grupo opresor emite oraciones de odio sin agencialidad intencional son dos: en los prejuicios reproducidos irreflexivamente y en los residuos prejuiciosos. Con respecto a estas categorías, Fricker (2007) nos menciona que ser una *persona virtuosa a nivel de la justicia testimonial* (es decir, ser una persona que no se ve afectada por prejuicios y que valora adecuadamente el testimonio de sus interlocutores) implica tener una crítica reflexiva hacia nuestros propios prejuicios de odio. Aplicar la crítica reflexiva sobre nuestros prejuicios puede llevar a la eliminación de estos. Considero que resulta claro que esta eliminación no es automática, sino que se da en diferentes etapas.

Si el prejuicio no es o es poco analizado críticamente, será *reproducido irreflexivamente*. Tal como se ha señalado, la gran mayoría de los prejuicios de odio que tiene la gente son obtenidos a raíz de la expansión socio-contextual de los discursos de odio. De esta manera, puede ocurrir que, al no haber reflexionado sobre los propios prejuicios, el discurso de odio imperante a nivel social, y transmitido a raíz de la formación de creencias, normas y sentimientos compartidos en las diferentes conversaciones, sea reproducido irreflexivamente por la persona.[14]

Cuando una persona empieza a analizar críticamente sus propios prejuicios, esta pasa a emitir cada vez menor cantidad de oraciones con ideas de odio, pues se da cuenta de que las ideas de odio que poseía no estaban bien fundamentadas. Mientras el prejuicio se va eliminando progresivamente, se pasa a emitir oraciones con ideas de odio producto de lo que Fricker (2007) llama *residuos prejuiciosos*. Este fenómeno se da cuando nuestro sistema de creencias choca con nuestra formación psicológica, llevándonos a que, de forma esporádica, emitamos alguna oración de odio sin tener creencias de odio. Sobre esto, podrían mencionarse como posibles ejemplos el de un activista de barrios populares, que está involucrado en gran cantidad de causas y es reconocido públicamente, que de forma muy esporádica y excepcional emite oraciones clasistas, o el de un activista racial, involucrado en gran cantidad de causas, que en su empresa no suele contratar personas afrodescendientes.

14 Este fenómeno es correctamente señalado en Alcoff (2010) y Anderson (2012), donde las autoras señalan que no siempre nos son claros nuestros propios prejuicios discriminatorios. Incluso una persona muy virtuosa a nivel justicia testimonial en general podría no ser consciente de que reproduce ciertos prejuicios discriminatorios que ella posee.

4. Aplicación del discurso influenciado a los casos de estudio

4.1. Programas discriminatorios: los orígenes

Tal como mencioné en la sección dos, el entrenamiento de los programas que funcionan mediante aprendizaje automático se basa en una clasificación de atributos relevantes. En este marco, como se indicó en la sección 3.2., los discursos de odio se expanden contextualmente en las conversaciones que incluyen ideas discriminatorias, pudiendo estas ideas afectar tanto a los datos de entrenamiento mismos como a los programadores, y pudiendo llevar a que estas ideas de odio se vean reflejadas en los programas. De esta manera, diferentes cuestiones relativas a los datos de entrenamiento, a la forma en que se construye un algoritmo o a las etiquetas utilizadas pueden resultar en programas informáticos que discriminen.[15] De todas estas cuestiones de las que puede surgir un programa informático discriminatorio, mencionaré sólo dos, que son las que nos permitirán, junto con la noción de discurso influenciado, explicar los casos de estudio mencionados en la sección 2.2: el sesgo histórico y el sesgo en la especificación del problema.

El sesgo histórico surge cuando un algoritmo de aprendizaje automático aprende de datos que son como son debido a prácticas discriminatorias presentes en el entorno, incluso si los datos son correctamente seleccionados. Como ejemplo, supongamos que queremos confeccionar un programa de selección de docentes universitarios de filosofía. En este caso, parece natural que uno de los datos que el programa debería tener en cuenta sería la cantidad de artículos académicos que el postulante publicó. Con respecto a este parámetro, sin embargo, Johnson (2020) nos comenta que estadísticamente en filosofía es más difícil que se acepte en revistas académicas la publicación de artículos de mujeres que de hombres. Teniendo esto en cuenta, el programa de selección de docentes de filosofía estará sexualmente sesgado, pues seguramente elija más hombres que mujeres. Sigo a Johnson (2020) y pienso que este tipo de sesgo nos enfrenta al difícil desafío de intentar realizar un equilibrio entre la precisión del programa y la no discriminación.

Por otra parte, el sesgo en la especificación del problema surge cuando el objetivo o los objetivos para los cuales se utilizará un programa resultan complejos, controvertidos y/o ambiguos, y, en base a esto, se generan dificultades en la creación y en los resultados obtenidos por un programa. Así, por ejemplo, supongamos que se quiere crear un programa que prediga el «éxito de estudiantes universitarios». Siguiendo a Fazelpour y Danks (2021), este tipo de objetivo, claramente complejo y controvertido, puede conllevar grandes problemas, tanto a la hora de programar el algoritmo específico que resolverá la cuestión (por ejemplo, comentan los autores, seguramente habrá problemas en la elección de las variables que se tendrán en cuenta para poder predecir el «éxito») como al momento de analizar los resultados finales.

15 En este artículo no se pretende dar una taxonomía específica de los tipos de sesgos que podrían resultar en un programa informático que discrimine. Para consultar posibles taxonomías de sesgos presentes en programas informáticos que pueden resultar en programas informáticos discriminatorios, véanse Danks y London (2017), Mehrabi et al. (2019) y/o Suresh y Guttag (2021).

4.2. Programas discriminatorios: tematización teórica de los casos de estudio, problemas y soluciones

Explicada la teoría, tematizados los casos de estudio y analizados algunos sesgos específicos que pueden derivar en programas informáticos que discriminen, la relación entre programas informáticos que funcionan lingüísticamente y mediante aprendizaje automático y el discurso influenciado es bastante clara. Para analizar esto, primero se explicará qué tipo de sesgos originan que los programas analizados como casos de estudio discriminen.

El caso del algoritmo de predicción de búsqueda de Google y del traductor son claras instancias de sesgo histórico. En el caso del algoritmo de predicción de búsqueda, seguramente las predicciones de búsqueda que son otorgadas como salidas sean las búsquedas más realizadas en el buscador y, por lo tanto, sean buenas predicciones de búsqueda a nivel práctico, pero los datos de los que el programa aprendió para otorgarnos estos resultados resultan ser como son debido a prácticas discriminatorias presentes en el entorno. De igual modo, en principio parece deseable que un traductor traduzca oraciones basándose en la manera en que habla la gente donde el idioma es nativo, pero lo que no se tuvo en cuenta es que la manera en que la gente habla conlleva prejuicios sexistas, en este caso.

Con respecto a la chatbot Tay, se cayó en un sesgo histórico y, seguramente, en un sesgo en la especificación del problema. Sobre el sesgo histórico, es claro que Tay aprendió de datos discriminatorios presentes en Twitter. Ahora bien, en este caso específico, considero que los resultados discriminatorios repuestos en la sección 2.2. surgen por problemas en la especificación del objetivo del programa. Si el objetivo de Microsoft era crear una inteligencia artificial que hablara como una adolescente de 19 años, entonces infiero que el programa fue correctamente entrenado (pues en Twitter hay gran cantidad de adolescentes y parece correcto que el programa aprenda a hablar como un adolescente tomando a tweets como datos). Si esto es así, la contrariedad en este caso es que los adolescentes estadounidenses son discriminadores y el programa únicamente reflejó esto. Por otro lado, supongo que el problema es que el objetivo planteado en la creación del programa resultó excesivamente controversial o ambiguo. Si no se quería una chatbot que discrimine, entonces no se debió crear una chatbot que imite a una adolescente de 19 años desde un inició o debió plantearse como objetivo el crear una chatbot que hable como un adolescente y que emita oraciones no moralmente controversiales.

Retomando entonces la noción de discurso influenciado, en base a los casos analizados (y, supongo, en la mayoría de programas no tan complejos del tipo señalado), se puede concluir que los programas emiten oraciones de odio debido a que reproducen irreflexivamente los sesgos discriminatorios presentes en los datos de los cuáles aprenden. En otras palabras, la reproducción de estas ideas de odio son un caso paradigmático de discurso influenciado.[16]

Además de esto, resulta bastante claro que un programa informático de la clase mencionada y que no sea demasiado complejo no puede analizar reflexivamente sus propios prejui-

16 Asociar una noción como la de discurso influenciado, pensada para personas humanas en tanto agentes intencionales, a programas informáticos presupone la idea de que los programas informáticos son efectivamente agentes intencionales. Esta posición bien conocida en la literatura, que suele basarse en la capacidad de ciertos programas informáticos de pasar la prueba de la «actitud intencional» propuesta por Dennett (1987, 2009), puede hallarse en Laukyte (2017), List (2021) y Russell y Norvig (2010) c.c. II y XXVI.

cios de odio, es decir, no puede aplicar la reflexión crítica propia de la virtud de la justicia testimonial. Con respecto a esto, las empresas informáticas suelen usar dos estrategias para intentar que no se creen sesgos algorítmicos que resulten en programas discriminatorios, que, metafóricamente, serían como aplicar la reflexión crítica propia de la virtud de la justicia testimonial en estos programas. En primer lugar, empresas como Google y Microsoft tienen equipos especiales de empleados trabajando en la detección de sesgos y la modificación de algoritmos (García, 2016), [17] y, en segundo lugar, se han intentado crear programas informáticos que realizan la tarea de detección de sesgos (véase Hajian y Domingo-Ferrer, 2013; Veale y Binns, 2017).

A pesar de que, en principio, estas parecen estrategias correctas para eliminar los sesgos algorítmicos que producen programas informáticos discriminatorios, serias dificultades se presentan ya que: (1) comprender el funcionamiento de algunos programas computacionales resulta muy difícil debido a la complejidad de sus algoritmos (especialmente, este problema se presenta en los programas que funcionan con bases de datos muy amplias) (Johnson, 2020; Ramírez-Bustamante y Páez, 2022; Sandvig et al., 2014); (2) en muchas oportunidades, los códigos-fuente de los programas son patentados y privados, lo que dificulta enormemente su diagnóstico y modificación (Ramírez-Bustamante y Páez, 2022; Sandvig et al., 2014); (3) los empleados y/o los programas que realizan las tareas de detección y eliminación de sesgos pueden tener ellos mismos sus propios prejuicios discriminatorios, por lo que podrían no ser completamente eficientes en la realización de sus tareas; y (4) resulta habitual que la información necesaria para la modificación de los algoritmos sea información privada de la gente (protegida por ley) (Veale y Binns, 2017).

Claramente, las soluciones algorítmicas como las planteadas hace un momento son sumamente valorables y ayudan a crear algoritmos computacionales más justos, pero, a pesar de esto, sigo a Crawford (2021) y a Noble (2018) en la idea de que la solución definitiva (pero no así fácil) a esta clase de problemas es mediante la eliminación (o, por lo menos, la reducción) de los prejuicios discriminatorios a nivel social general. Es un hecho que los prejuicios discriminatorios están presentes en la sociedad y ocurre que los programas heredan en su programación este tipo de prejuicios. Si se eliminaran estos prejuicios a nivel social, esta herencia, obviamente, no se daría.

6. Conclusión

El objetivo del presente artículo ha sido analizar el fenómeno de los sesgos en los programas lingüísticos que funcionan mediante aprendizaje automático acudiendo a herramientas teóricas de la filosofía del lenguaje. En este marco, el discurso influenciado ha sido la noción central que permitió realizar este análisis. Creo que es importante destacar el rol que la filosofía, en tanto disciplina, cumple en relación con esta clase de problemas.

17 Sobre esta cuestión, sigo a García (2016) en la opinión de que es fundamental que personas con diversidad de géneros, etnias y nacionalidades formen parte de los equipos de supervisión de algoritmos (y, seguramente, de las empresas tecnológicas en general), ya que los propios miembros de los grupos socialmente oprimidos son mucho más eficaces detectando elementos discriminatorios hacia sus propios colectivos que personas que no pertenecen a ellos.

La filosofía es la encargada de tematizar teóricamente las cuestiones relacionadas con los prejuicios discriminatorios en general y, de esta manera, puede contribuir a su eliminación. En otras palabras, la filosofía puede ser una herramienta fundamental para el desarrollo de la virtud de la justicia testimonial a nivel general. Al tematizar el problema de la expansión y la reproducción de los prejuicios discriminatorios, espero que el presente artículo represente un pequeño aporte a la realización de esa importante tarea.[18]

Referencias

Alcoff, L. (2010). Epistemic identities. *Episteme*, 7 (2), 128–37. https://doi.org/10.3366/epi.2010.0003

Alonso Alemani, L., Benotti, L., González, L., Sánchez, J., Busaniche, B., Halvorsen, A. y Bordone, M. (2022). Una herramienta para superar las barreras técnicas para la evaluación de sesgos en las tecnologías del lenguaje humano. Recuperado de la página web de *Fundación Vía Libre*. https://www.vialibre.org.ar/wp-content/uploads/2022/08/vialibre_Una-herramienta-para-superar-las-barreras-tecnicas.pdf

Anderson, E. (2012). Epistemic justice as a virtue of social institutions. *Social Epistemology*, 26 (2), 163-173. https://doi.org/10.1080/02691728.2011.652211

Bianchi, C. (2020). Discursive injustice: the role of uptake. *Topoi* 40 (1): 181-190. https://doi.org/10.1007/s11245-020-09699-x

Crawford, K. (2021). *Atlas of AI*. Yale University Press.

Danks, D. y London, A. (2017). Algorithmic bias in autonomous systems. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Australia, 17, 4691–4697

Dennett, D. (1987). *The intentional stance*. MIT Press.

Dennett, D. (2009). Intentional systems theory. En A. Beckermann, B. McLaughlin y S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 339–350). Oxford University Press.

Fazelpour, S. y Danks, D. (2021). Algorithmic bias: senses, sources, solutions. *Philosophy Compass*, 16 (8), e12760. https://doi.org/10.1111/phc3.12760

Fricker, M. (2007). *Epistemic injustice*. Oxford University Press.

Fricker, M. (2013). Epistemic justice as a condition of political freedom? *Synthese*, 190 (7), 1317-1332. https://doi.org/10.1007/s11229-012-0227-3

García, M. (2016). Racist in the machine: the disturbing implications of algorithmic bias. *World Policy Journal*, 23 (4), 111-117. https://doi.org/10.1215/07402775-3813015

Gelber, K. (2019). Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, 22 (3), 607-622. https://doi.org/10.1080/13698230.2019.1576006

18 Versiones previas de este artículo fueron presentadas en las *XV Jornadas de Comunicación de Investigación en Filosofía*, en las *IV Jornadas Nacionales de Filosofía del Departamento de Filosofía (Universidad de Buenos Aires)*, en el taller *Inteligencia Artificial y Filosofía: el Desafío de los Sesgos* y en las sesiones de investigación del Grupo TALK. Agradezco a los asistentes a aquellas reuniones por los comentarios y discusiones. Además, merece un agradecimiento especial Eleonora Orlando por sus comentarios por escrito a versiones previas del presente artículo.

Gerón, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly.

Hajian, S. y Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25 (7), 1445-1459. https://doi.org/10.1109/TKDE.2012.72

Harper Lee, N. (1960). *To kill a mockingbird*. J. B. Lippincott & Co.

Hern, A. (2016, 24 de marzo). Microsoft scrambles to limit PR damage over abusive AI bot Tay. *The Guardian*. https://www.theguardian.com/technology/2016/mar/24/microsoft-scrambles-limit-pr-damage-over-abusive-ai-bot-tay al 20/08/2022.

Hesni, S. (2018). Illocutionary frustration. *Mind*, 127 (508), 947-976. https://doi.org/10.1093/mind/fzy033

Hornsby, J. y Langton, R. (1998). Free speech and illocution. *Legal Theory*, 4, 21-37. https://doi.org/10.1017/S1352325200000902

Johnson, G. (2020). Algorithmic bias: on the implicit biases of social technology. *Synthese*, 198 (10), 9941-9961. https://doi.org/10.1007/s11229-020-02696-y

Langton, R. (1993). Speech acts and unspeakable acts. *Philosophy and Public Affairs*, 22 (4), 293-330.

Langton, R. (2012). Beyond belief: pragmatics in hate speech and pornography. En I. Maitra y M. McGowan (Eds.), S*peech and harm: controversies over free speech* (pp. 72-93). Oxford University Press.

Langton, R. y West, C. (1999). Scorekeeping in a pornographic language game. *Australasian Journal of Philosophy*, 77 (3), 303-319. https://doi.org/10.1080/00048409912349061

Laukyte, M. (2017). Artificial agents among us: should we recognize them as agents proper? *Ethics and Information Technology*, 19 (1),1-17. https://doi.org/10.1007/s10676-016-9411-3

Lewis, D. (1983). Scorekeeping in a language game. En D. Lewis, *Philosophical papers: volume 1* (pp. 233-249). Oxford University Press.

List, C. (2021). Group agency and artificial intelligence. *Philosophy and Technology*, 4, 1-30. https://doi.org/10.1007/s13347-021-00454-7

Maitra, I. y McGowan, M. (2010). On silencing, rape, and responsibility. *Australian Journal of Philosophy*, 88 (1), 167-172. https://doi.org/10.1080/00048400902941331

Marques, T. (2022). The expression of hate speech. *Journal of Applied Philosophy*, 10, 1-29. https://doi.org/10.1111/japp.12608

Matsuda, M. (1993). Public response to racist speech. En M. Matsuda, C. Lawrence, R. Delgado y K. Williams Crenshaw (Eds.), W*ords that wound: critical race theory, assaultive speech and the first amendment* (pp. 17-52). Westview Press.

McGowan, M. (2003). Conversational exercitives and the force of pornography. *Philosophy & Public Affairs*, 31 (2), 155-189. https://doi.org/10.1111/j.1088-4963.2003.00155.x

McGowan, M. (2004). Conversational exercitives: something else we do with our words. *Linguistics and Philosophy*, 27, 93-111. https://doi.org/10.1023/B:LING.0000010803.47264.f0

McGowan, M. (2019). *Just words*. Oxford University Press

McKinney, R. (2016). Extracted speech. *Social Theory and Practice*, 42 (2), 258-284. https://doi.org/10.5840/soctheorpract201642215

Medina, J. (2013). *The epistemology of resistance*. Oxford University Press.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. y Galstyan, A. (2019). A survey on bias and fairness in machine learning. *CM Computing Surveys*, 54 (6), 1–35. https://doi.org/10.1145/3457607

Minghella, A. (Director) (1999). *The talented Mr. Ripley* [El talentoso Sr. Ripley] [Película]. Paramount Pictures.

Noble, S. (2018). *Algorithms of oppression*. New York University Press.

Peet, A. (2017). Epistemic injustice in utterance interpretation. *Synthese*, 194 (9), 3421-3443. https://doi.org/10.1007/s11229-015-0942-7

Pérez, E. (2021). Cuando traducimos un idioma con pronombres sin género como el euskera o el húngaro, Google asume el masculino o femenino. Recuperado de *Xataka* web. https://www.xataka.com/robotica-e-ia/cuando-traducimos-idioma-genero-neutro-como-euskera-hungaro-google-asume-masculino-femenino.

Ramírez-Bustamante, N. y Páez, A. (2022). Análisis jurídico de la discriminación algorítmica en los procesos de selección laboral, en N. Angel y R. Urueña (Eds.), *Derecho, poder y datos: aproximaciones críticas al derecho y las nuevas tecnologías*. Ediciones Uniandes. Recuperado de https://philpapers.org/archive/PEZAJD.pdf

Russell, S. y Norvig, P. (2010). *Artificial intelligence* (3ra ed.). Pearson.

Russell, S. y Norvig, P. (2020). *Artificial intelligence* (4ta ed.). Pearson.

Sandvig, C., Hamilton, K., Karahalios, K. y Langbort, C. (2014). An algorithm audit. En S. Peña, V. Eubanks y S. Barocas (Eds.), *Data and discrimination: collected essays* (pp 6-10). Open Technology Institute.

Stair, R. y Reynolds, G. (2010). *Principios de sistemas de información* (9na ed.) Cengase Learning.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25 (5/6), 701-721. https://doi.org/10.1023/A:1020867916902

Stalnaker, R. (2014). *Context*. Oxford University Press.

Suresh, H. y Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of EAAMO '21: Equity and access in algorithms, mechanisms, and optimization*, Estados Unidos, 1-9. https://doi.org/10.1145/3465416.3483305

Veale, M. y Binns, R. (2017). Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4 (2), 1-17. https://doi.org/10.1177/2053951717743530

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 63-80
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.562811

Cerrando una brecha: una reflexión multidisciplinar sobre la discriminación algorítmica*,**

Bridging a gap: A multidisciplinary reflection on algorithmic discrimination

*PILAR DELLUNDE****
*ORIOL PUJOL*****
*JORDI VITRIÀ******

Resumen. Este artículo aborda el concepto de discriminación algorítmica desde una perspectiva conjunta de la filosofía y la ciencia de la computación, con el propósito de establecer un marco de discusión común para avanzar en el despliegue de las inteligencias artificiales en las sociedades democráticas. Se presenta una definición no normativa de discriminación y se analiza y contextualiza el concepto de algoritmo usando un enfoque intencional, enmarcándolo en el proceso

Abstract. This article presents a joint reflection from philosophy and computer science on the concepts behind algorithmic discrimination with the aim of providing a common framework for discussion to advance the deployment of artificial intelligence in democratic societies. A non-normative definition of discrimination is presented, and the concept of algorithm is analyzed and contextualized using an intentional approach,

Recibido: 28-03-2023. Aceptado: 27-06-2023

* Esta publicación ha sido parcialmente financiada por los proyectos: 2021 SGR 01104 y 2021 SGR 00754 de la Generalitat de Catalunya y H2020-MSCA-RISE-2020 project MOSAIC (Grant Agreement 101007627).

** Todos los autores han participado de la misma forma en todo el proceso de elaboración del trabajo y han puesto sus nombres en orden alfabético de apellidos.

*** Es Catedrática de Lógica del Departament de Filosofia de la Universitat Autònoma de Barcelona. Sus líneas de investigación se centran en la lógica fuzzy, argumentación computacional y ética en el diseño de sistemas de inteligencia artificial. Ha publicado recientemente "An art painting style explainable classifier grounded on logical and commonsense reasoning" (Soft Computing, 2023) y "Probabilistic Argumentation: An Approach Based on Conditional Probability" (Lecture Notes in Computer Science 12678, 2021). Contacto: pilar.dellunde@uab.cat

**** Es Catedrático del Departament de Matemàtiques i Informàtica de la Universitat de Barcelona. Su línea de investigación principal trata sobre los fundamentos algorítmicos del aprendizaje automático y su impacto. Ha publicado recientemente "The Forgotten Human Autonomy in Machine Learning" (CEUR-WS, IAIL, 2022) y "Copying Machine Learning Classifiers" (IEEE Access 8, 160268-160284). Contacto: oriol_pujol@ub.edu

***** Es Catedrático del Departament de Matemàtiques i Informàtica de la Universitat de Barcelona. Sus líneas de investigación son aprendizaje automático, inferencia causal y aspectos éticos de la inteligencia artificial. Ha publicado recientemente "Estimand-Agnostic Causal Query Estimation with Deep Causal Graphs" (IEEE Access 10, 2022, 1370-71386), y "A survey on uncertainty estimation in deep learning classification systems from a Bayesian perspective" (ACM Computing Surveys, 54 (9), 2022, 1-35). Contacto: jordi.vitria@ub.edu

de toma de decisiones e identificando las fuentes de discriminación, así como los conceptos detrás de su cuantificación para terminar exponiendo algunos límites y desafíos.
Palabras clave: Discriminación, algoritmos, inteligencia artificial, métricas.

framing it in the decision-making process and identifying the sources of discrimination, as well as the concepts behind its quantification, ending by exposing some limits and challenges.
Keywords: Discrimination, algorithms, artificial intelligence, metrics.

1. Introducción

En los últimos capítulos de *Homo Deus*, Y. N. Harari reflexiona sobre el concepto de dataísmo (Harari, 2016, 821): «El dataísmo sostiene que el universo consiste en flujos de datos, y que el valor de cualquier fenómeno o entidad está determinado por su contribución al procesamiento de datos.» afirmando que «el trabajo de procesar los datos debe encomendarse a algoritmos electrónicos, cuya capacidad excede con mucho a la del cerebro humano.»

El debate al que Y. N. Harari contribuye tiene su origen, entre otros, en el artículo de D. Brooks publicado en *The New York Times*, en el que sostiene que existe una revolución de los datos:

> Ahora tenemos la capacidad de acumular enormes cantidades de datos. Esta capacidad lleva consigo un cierto presupuesto cultural —que todo lo mensurable debe ser medido; que los datos son lentes transparentes y fiables que nos permiten filtrar todo emocionalismo y toda ideología; que los datos nos ayudarán a hacer cosas significativas como predecir el futuro. (Brooks, 2013).

Muchas de las reflexiones filosóficas relevantes sobre el dataísmo prestan poca atención a lo que Y. N. Harari llama *algoritmos electrónicos,* o a la intervención humana en los procesos de decisión que utilizan dichos algoritmos. La trasparencia y fiabilidad de los datos a la que hace referencia D. Brooks y que recurrentemente encontramos en el imaginario popular, el discurso gubernamental, el márquetin mercantilista, e incluso, en ocasiones, en el argumentario filosófico, implica diversas nociones que promueven imágenes de los algoritmos que los presentan como objetivos y de propósito universal. Por ejemplo, el panel de la Unión Europea, en su estudio "A governance framework for algorithmic accountability and transparency", indica

> As with much of the digital economy, the use of algorithmic systems is characterised by the highly cross-border nature and global reach of the services that are built on these technologies. (EU Res Service, 2019)

o en las críticas a esta supuesta objetividad en (Guersenzvaig et al., 2022). El discurso sobre la objetividad tiene como finalidad última legitimar la confianza en el algoritmo bajo la afirmación de "un algoritmo objetivo es confiable". El concepto de objetividad científica entendido como fidelidad a los hechos/datos, ideal libre de valores, e independiente de sesgos es incompatible con el del algoritmo basado en IA. Si bien es cierto que el algoritmo

tiene un comportamiento determinista y por lo tanto libre del sesgo intrapersonal que se encuentra en la decisión humana; es decir, dado un mismo estímulo/situación un ser humano puede decidir sobre este de forma diametralmente opuesta en función de su estado mental y de su interocepción. Pero, salvo este comportamiento determinista, los algoritmos basados en datos no están libre de valores y sesgos.

Como se comentará, el tecnólogo diseña un producto con una intencionalidad y propósitos concretos, para conseguir unos objetivos, usualmente de forma conjunta con los especialistas del dominio y los agentes que pondrán el sistema en funcionamiento. Fuera de posibles pretensiones por parte de grandes corporaciones tecnológicas, raramente un tecnólogo puede pretender por su cuenta crear un sistema con vocación universal. La consciencia de la necesidad de acotar el uso del algoritmo a un contexto y entorno determinado viene marcada de salida por la propia accesibilidad a los datos, por su representatividad, por el conocimiento de los potenciales sesgos estructurales que estos pueden tener y por los desbalanceos potenciales correspondientes a sesgos estadísticos en el propio diseño experimental y la recogida de datos. Todos estos factores hacen que en la literatura técnica de aprendizaje automático se hable de conceptos que precisamente reconocen la limitación de uso y de objetivos y estudian formas de combatir estas limitaciones. Por ejemplo, conceptos como *transfer learning, domain adaptation* (Pratt, 1993), o *differential replication* (Unceta, 2020), son conceptos propios de las disciplinas de aprendizaje automático que asumen que el mundo real es variable y está en constante cambio y que la representación que el algoritmo tiene de éste es particular y puede contener sesgos no deseados. Los conceptos anteriormente mencionados recogen formas de adaptar y reaprovechar el conocimiento aprendido en el algoritmo cuando el dominio o las restricciones que imperan sobre el algoritmo ya sean normativas, operacionales, éticas, etc., convierten al algoritmo en no usable.

Las contribuciones originales de este artículo son producto de una reflexión conjunta desde las ciencias de la computación y la filosofía sobre la IA, centrado en la discriminación algorítmica, con el objetivo de avanzar de forma efectiva en un diseño más ético de estas tecnologías. Más concretamente: en la sección 2, revisamos el concepto de discriminación desde un punto de vista filosófico, proponiendo una el uso de una definición no normativa del concepto de discriminación, que permite aislarla de conceptos normativos o moralizantes y que puede fácilmente ser compartida y resultar útil en el campo de las ciencias de la computación. En la sección 3 proponemos un marco de análisis de la discriminación en la toma de decisiones que va más allá del "algoritmo", basado en el enfoque intencional, que define un marco teórico que también puede ser compartido entre la IA y la filosofía. Esto nos permite introducir la idea de "reducción al enfoque de diseño" como criterio para separar problemas de discriminación simples (originados por una discrepancia en los criterios normativos en un sistema deductivo) de problemas complejos (en los que los criterios normativos son implícitos y mediados por los datos y el proceso de construcción del sistema inductivo). Posteriormente, analizamos en detalle los aspectos propios de la toma de decisiones basada en IA e identificamos las fuentes de la discriminación, introduciendo el concepto de *legitimidad del uso de un sistema de IA*, como estadio previo a cualquier análisis ético del uso de la IA. En la sección 4, ampliamos la visión del *algoritmo como sistema* a *algoritmo dentro del proceso de la toma de decisiones*, e introducimos los conceptos fundamentales

y líneas de pensamiento filosófico que sirven de guía a la cuantificación de la discriminación. Así mismo, subrayamos las dificultades en este proceso e identificamos vías para uso efectivo. Finalmente, en la última sección exponemos algunos de los límites actuales y la necesidad de una reflexión multidisciplinar que genere un marco de discusión común para poder avanzar en el despliegue y uso de la IA en una sociedad democrática, teniendo en cuenta el punto de vista de todos los agentes implicados.

2. Una aproximación filosófica a la discriminación

En el artículo "Bias and Fairness in Machine Learning" (Mehrabi et al., 2021) de la serie *ACM Computing Surveys*, los autores presentan un estado del arte sobre los sesgos y la equidad en el aprendizaje automático, e introducen una taxonomía para clasificar las diferentes definiciones de equidad con las que los investigadores en aprendizaje automático han trabajado para tratar de evitar los sesgos en los sistemas de IA. En la parte final del artículo, se discuten retos y oportunidades para futuras investigaciones en este campo. Llama la atención el primero de estos retos que ellos definen como *synthesizing a definition of fairness*. Este desafío plantea varios elementos interesantes para la reflexión filosófica. En primer lugar, al plantear este reto no se tiene en cuenta que los conceptos de equidad y discriminación son complejos y dinámicos, y no pueden ser representados únicamente utilizando formalismos matemáticos, ya que su significado está condicionado por el contexto y el momento histórico.

En segundo lugar, aunque algunos autores, por ejemplo, (Simon et al., 2020) y (Seng et al., 2021), han destacado la necesidad de adoptar un enfoque más holístico en el diseño de estas tecnologías, cabe destacar que esta sigue siendo una demanda común (incluso en proyectos interdisciplinarios que emplean metodologías como el *Value Sensitive Design*), el pedir a los investigadores de humanidades y ciencias sociales que proporcionen definiciones que sirvan como base para introducir una noción precisa, universal y formal para su implementación. Esto implica desconocimiento del papel de las definiciones en estos ámbitos, y plantea la necesidad de resituar el papel de las humanidades y las ciencias sociales en el diseño de los sistemas de IA.

En esta sección presentamos la definición filosófica de discriminación del libro (Lippert-Rasmussen, 2014), una definición no moralizante (que no implica necesariamente que discriminar es incorrecto) llamada *discriminación grupal*. Esta definición representa un concepto más amplio de discriminación que el que encontramos en los artículos sobre discriminación algorítmica. Si bien en algunos artículos sobre discriminación algorítmica se citan contribuciones de K. Lippert-Rasmussen, por ejemplo, respecto a la discriminación estadística (Barocas, 2016), las reflexiones de (Lippert-Rasmussen, 2014) son aún poco conocidas en el ámbito de las ciencias de la computación. Sin pretender hacer una presentación exhaustiva, introducimos aquellos elementos básicos de la definición que consideramos que pueden ser útiles para futuras reflexiones interdisciplinares en este ámbito.

En (Lippert-Rasmussen, 2014) se presentan condiciones necesarias y suficientes para considerar que un acto, una política o una práctica es discriminatoria a nivel de grupo:

> X discriminates against Y in relation to Z by Φ-ing if, and only if,
>
> (i) there is a property, P, such that (X believes that) Y has P and (X believes that) Z does not have P,
>
> (ii) X treats Y worse than Z by Φ-ing,
>
> (iii) it is because (X believes that) Y has P and (X believes that) Z does not have P, that X treats Y worse than Z by Φ-ing,
>
> (iv) P is the property of being member of a certain socially salient group (to which Z does not belong), and
>
> (v) Φ-ing is a relevant type of act etc., and there are many acts etc. of this type, and this fact makes people with P (or some subgroup of these people) worse off relative to others, OR Φ-ing is a relevant type of act etc., and many acts etc. of this type would make people with P worse off relative to others, OR X's Φ-ing is motivated by animosity towards individuals with P or by the belief that individuals who have P are inferior or ought not to intermingle with others. (Lippert-Rasmussen, 2014, 44-45)

Hemos elegido esta definición porque no es moralizante, y precisamente por ello, nos permite analizar casos como el de los impuestos proporcionales, que tratan de manera diferente a los contribuyentes; la aplicación de este criterio proporcional representa un trato desventajoso para las personas con más capacidad económica, pero podemos considerar que esta discriminación no es ni incorrecta ni injusta. Una característica importante de la definición es que discriminar implica un trato desventajoso, hecho que debe distinguirse del trato que causa daños.

La creación de un marco conceptual para el debate sobre la discriminación algorítmica hace que un concepto que tenga en cuenta la pertenencia a un grupo sea más útil que una concepción como la de B. Eidelson que sostiene que «acts of discrimination are intrinsically wrong when and because they manifest a failure to show the discriminatees the respect that is due them as persons.» (Eidelson 2015: 7) Este relato prescinde explícitamente del requisito de pertenecer a un grupo socialmente destacado, y en su lugar solo requiere que el discriminador responda a alguna diferencia percibida de cualquier tipo entre la víctima y otras personas. Eidelson considera dos dimensiones de la personalidad (*personhood*): todas las personas 1) son iguales y tienen valor intrínseco y 2) son agentes autónomos. (Eidelson, 2015: 79), y la discriminación puede violar una o ambas de estas dimensiones. Este enfoque implica utilizar una definición moralizante de discriminación.

Para K. Lippert-Rasmussen es central la relación que se establece entre discriminación y trato diferencial en base a la pertenencia a un grupo socialmente destacado (*socially salient group)*. El autor entiende que un grupo es socialmente destacado si la percepción de la pertenencia a él es importante en la estructura de las interacciones sociales a través de una amplia gama de contextos. Ejemplos de grupos socialmente destacados pueden ser el conjunto de las mujeres o el de las personas inmigrantes en un país en un momento determinado de tiempo. En cambio, esta definición grupal nos permite no considerar, de manera general, la meritocracia como una discriminación de grupo.

La discriminación es esencialmente comparativa, ya que no se puede discriminar a nadie a menos que haya otras personas que reciban un trato mejor en comparación. Otra importante característica de la discriminación así definida es que es independiente de las propiedades reales de las personas, ya que no hay necesaria superposición entre las propiedades que convierten a alguien en objeto de discriminación y las propiedades que la persona realmente posee. Por ejemplo, un hombre podría ser víctima de discriminación contra las mujeres, por un error en la entrada de sus datos.

De especial relevancia para el estudio de las implicaciones éticas de los sistemas de toma de decisiones es el concepto de discriminación estadística, es decir, cuando se trata a personas de manera diferente sobre la base de generalizaciones estadísticas explícitas o implícitas sobre el grupo al que esta persona pertenece. En tanto que caso particular de la definición de discriminación que hemos introducido al principio de la sección, la discriminación estadística es esencialmente comparativa y se discrimina en función de la pertenencia a un grupo socialmente destacado, a excepción de casos como la discriminación genética. Lippert-Rasmussen añade una cláusula más a la definición original para definir este tipo de discriminación:

> (vi) It is because (X believes that) Y has P and (X believes that) Z has not, and because (X believes that) P is statistically relevant, that X treats Y worse than Z by Φ-ing. (Lippert-Rasmussen 2014: 81)

Pero ¿qué significa que un grupo socialmente destacado sea estadísticamente relevante? En (Lippert-Rasmussen 2014: 86) se considera que un grupo es estadísticamente relevante si la probabilidad de tener alguna otra característica (por ejemplo, solicitar permiso de paternidad, o poseer drogas ilegales) varía sobre la base de qué grupos uno es miembro.

A veces, la evidencia estadística disponible puede ser correcta y utilizada de una manera no sesgada. La discriminación estadística no necesariamente se basa en evidencias estadísticas insuficientes o falsas. Si bien el uso de información estadística a menudo puede ser selectivo (específicamente, a menudo se puede usar para tomar como objetivo a minorías), el uso de la información estadística *per se* no necesita ser selectivo. No todo tipo de discriminación estadística está relacionada con creencias sobre un estatus inferior de aquellos a quienes se discrimina. Por ejemplo, la discriminación estadística a nivel tributario tiene como objetivo evitar la evasión de impuestos por parte de los más ricos.

En esta sección hemos dejado fuera elementos importantes, como el análisis en profundidad de la definición de *socially salient group*, de discriminación indirecta (muy relevante porque es una de las más difíciles de detectar en el diseño de los sistemas de IA) y que el propio (Lippert-Rasmussen 2014: 54-74) construye sobre la definición aquí presentada, de discriminación económica, o de interseccionalidad. Nuestro objetivo no era una presentación exhaustiva, sino, sobre todo, hacer explícita la complejidad del debate sobre esta noción, y cuestionar la posibilidad de encontrar una única formalización operacional de este concepto.

3. Sobre los algoritmos y la discriminación

En el imaginario colectivo sobre la discriminación algorítmica existen varios conceptos (algoritmo, datos, sesgos, etc.) que predefinen las bases de la discusión y que desde nuestro punto de vista requieren de una mayor elaboración para convertirse en fundamentos sólidos de un diálogo fructífero entre la filosofía y las ciencias de la computación. En los siguientes apartados se hace un análisis crítico de su significado y se propone pasar del concepto de *algoritmo* al concepto de *sistema*.

3.1. Algoritmos, sistemas de inteligencia artificial y el enfoque intencional

La idea de discriminación se produce a un nivel de abstracción que requiere el concepto de agencia y desde este punto de vista creemos que el uso del término "algoritmo" no es especialmente adecuado, a causa de su significado reduccionista.

El concepto clásico de "algoritmo", entendido como una secuencia de instrucciones que sirven para llegar a solución de un problema, evoca un enfoque deductivo que no representa los actuales sistemas de inteligencia artificial (IA). Los sistemas de IA se sitúan en una categoría especial de sistemas que, siguiendo el marco teórico de D. Dennett (Dennett, 1987), deben ser entendidos desde un enfoque intencional y que tienen una naturaleza distinta de los algoritmos clásicos.

Según la propuesta de D. Dennett tenemos tres alternativas cuando queremos entender un sistema complejo. La primera alternativa, el enfoque físico, usa las leyes de la física a un determinado nivel de abstracción para modelizar el sistema a partir de sus constituyentes y de las interacciones que podemos observar. El comportamiento de un paraguas que sale volando a causa del viento estaría a este nivel. Desde este punto de vista "entender" el sistema quiere decir tener una cierta capacidad de predicción de su comportamiento usando exclusivamente las leyes fundamentales de la naturaleza.

La segunda alternativa, el enfoque de diseño, nos permite entender un sistema a partir de la asunción que ha sido diseñado con un propósito y que por lo tanto cabe esperar que su comportamiento se ajuste a este propósito. En este caso los aspectos ligados a la física del sistema pueden ser subsidiarios y por lo tanto no especialmente útiles para entender su comportamiento. Una silla es un claro ejemplo de sistema que debe ser interpretado con un enfoque de diseño, al igual que un reloj o un procesador de textos en un ordenador.

La tercera alternativa, o enfoque intencional, se aplica a aquellos sistemas que se entienden mejor como agentes racionales, a los que se puede suponer unas creencias, un propósito y hasta una cierta representación del mundo que les permite conseguir su propósito. Pertenecen a este nivel sistemas tan dispares como un termostato, una colonia de hormigas, un ser humano, una empresa o la misma sociedad en la que vivimos, pasando por sistemas artificiales complejos, como robots o sistemas de IA.

Situados en este marco de análisis, podríamos preguntarnos si las acciones de un termostato son potencialmente discriminatorias para aquellas personas con una sensación térmica fuera de los rangos que el termostato supone normales. Desde el punto de vista intencional el termostato tiene un propósito y unas creencias bien definidos, así como un comportamiento

que le permiten conseguir su propósito en la mayoría de los casos. Suponiendo pues este escenario, el análisis de este tipo de discriminación desde el enfoque intencional no tiene mucho recorrido porque sus propósitos, creencias y comportamiento se pueden traducir sin ambigüedades a una serie de decisiones de diseño. Esto permite reducir el problema de la discriminación, en el caso del termostato, a un problema de diseño y buscar una solución sin movernos de ese nivel de abstracción, en el cual los humanos nos sentimos especialmente cómodos desde hace siglos.

El caso de un algoritmo desarrollado por un programador y que tiene por objetivo codificar una serie de reglas usadas, por ejemplo, en un proceso burocrático de concesión de ayudas sociales (Johnson, 2022), se encuentra exactamente al mismo nivel que el termostato y también puede ser reducido a un problema de diseño, puesto que las cuestiones normativas que se pueden derivar de su definición y uso se reducen a dos escenarios: una discrepancia en el valor normativo de las reglas o la presencia de un error en la codificación de las mismas.

La reducción del problema de discriminación a un problema de diseño también se extiende a sistemas complejos como un automóvil o un avión, pero no a un sistema de IA basado en datos. En este último caso, la reducción a nivel de diseño no es posible a causa de la naturaleza de su proceso de creación y de sus operaciones (Zerilli, 2022), mucho más complejo que los programas y algoritmos convencionales.

En los siguientes subapartados repasaremos los elementos del proceso de creación, desarrollo y despliegue de un sistema de IA y su papel en la perpetuación o amplificación de la inequidad. Siguiendo con la nomenclatura clásica en temas de discriminación algorítmica, usaremos el término *sesgo* para denotar alguna causa potencial de discriminación (Fazelpour, 2020), aun cuando el término es ambiguo y puede usarse con connotaciones positivas o negativas a constructos tan distintos como un algoritmo, un conjunto de datos o un comportamiento.

3.2. Sistemas de IA basados en datos

El elemento principal de cualquier teoría de la (in)equidad de la IA es que los modelos de IA basados en datos construyen un sistema predictivo para la toma de decisiones de forma inductiva, a diferencia de los métodos clásicos, que se basan en un proceso deductivo. Es esta diferencia la que determina, en la mayoría de los casos, el uso de los enfoques de diseño o intencional para su comprensión. El proceso inductivo es complejo e involucra una gran variedad de elementos de naturaleza distinta, hecho que refuerza el concepto de *sistema* en substitución de la de *algoritmo*. Por otra parte, este proceso puede ser origen de sesgos que resulten en problemas de discriminación.

3.2.1. Legitimidad

El proceso de creación de estos sistemas se inicia a partir de una descripción genérica de un objetivo que, en el caso de predecir eventos o estados futuros sobre personas y tener consecuencias sobre el mundo real, debe estar sujeto a un análisis ético. Suponiendo que este objetivo es legítimo (Sternberger, 1968), la legitimidad del uso de un sistema de IA para con-

seguir tal objetivo se puede determinar a partir del nivel de precisión de sus predicciones, de sus potenciales efectos discriminatorios, de su eficacia respecto al objetivo y también a veces del nivel de transparencia (Lazar, 2022). A estas propiedades podríamos añadir una condición de prudencia, la irreductibilidad: que no exista una solución viable basada en algoritmos clásicos y que por tanto el problema de la discriminación no sea reducible al enfoque de diseño.

3.2.2. Los datos

La adquisición de datos para entrenar un algoritmo de IA es un proceso que requiere un análisis detallado que permita evaluar su representatividad, veracidad, estabilidad, etc. y así evitar o minimizar los sesgos de representación (Mehrabi, 2021) o de medida (Jacobs, 2021).

Un aspecto de especial importancia en el proceso de recogida de datos es que el mundo real, *el mundo tal y como es*, no se corresponde necesariamente con nuestras creencias, con *el mundo tal y como podría y debería ser*. Los datos que obtenemos pueden ser pues reflejo de situaciones o procesos que pueden ser éticamente deficientes desde nuestro sistema de valores y que por lo tanto pueden contaminar el sistema resultante. Este tipo de sesgo se llama sesgo estructural, sistémico o social (Mitchell, 2021).

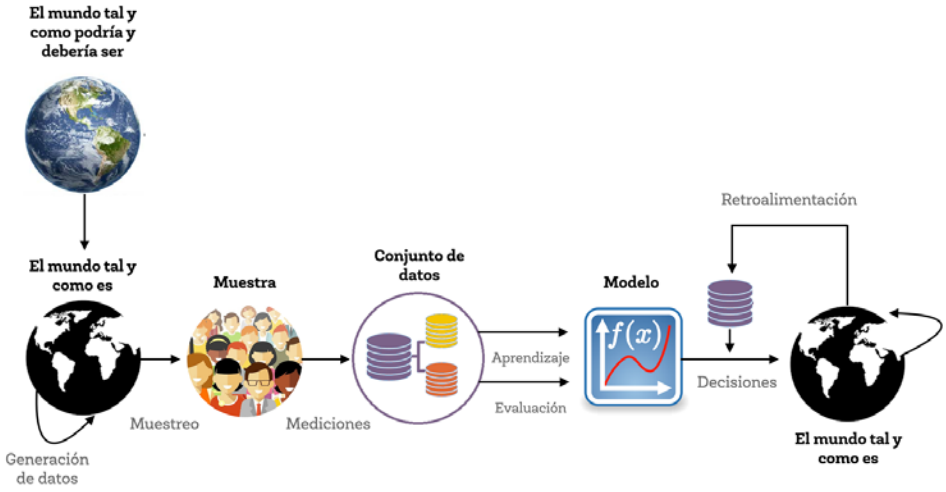

Figura 1. Proceso de creación de un sistema de IA

3.2.3. Los algoritmos

Dado un conjunto de datos que representa el fenómeno que se quiere modelizar, hace falta determinar una función objetivo, que pueda ser optimizada por un algoritmo de aprendizaje automático (Hardt, 2022). Esta función objetivo suele estar relacionada sólo de forma

indirecta con el objetivo general del sistema, puesto que éste no suele ser fácilmente expresable en términos predictivos. Por este motivo, si el objetivo general y la función objetivo no están alineadas se pueden introducir sesgos discriminatorios durante la predicción.

El producto final del algoritmo de aprendizaje es una serie de parámetros que definen un modelo de predicción. Estos modelos varían en su capacidad expresiva y también en su interpretabilidad, hecho que debe ser tenido en cuenta en aquellos ámbitos en los que la transparencia es necesaria (Rudin, 2019). También es importante señalar que en general los algoritmos de aprendizaje forman parte de librerías externas que han sido desarrolladas con objetivos genéricos y que su inspección no siempre está al alcance de los desarrolladores.

El algoritmo de aprendizaje también puede introducir otro tipo de sesgos que, aunque sean más sutiles que los originados en los datos, también pueden tener consecuencias. Este es el caso del sesgo inductivo (Mitchell, 1980) de un algoritmo de aprendizaje, entendido como el conjunto de suposiciones que se usan para maximizar la capacidad de generalización de un modelo.

3.2.4. El despliegue

El despliegue y aplicación de un modelo predictivo es uno de los aspectos menos estudiados, pero puede ser también el origen de alguna discriminación.

En primer lugar, se puede dar el llamado sesgo de despliegue, que se refiere a cualquier sesgo que surja cuando un sistema se usa o interpreta de manera inapropiada, muchas veces sin el conocimiento expreso de los diseñadores o desarrolladores. Este sesgo se podría dar, por ejemplo, en el caso de un modelo desarrollado para evaluar la probabilidad de cometer un nuevo delito y que fuera reusado para tomar decisiones sobre la duración de la sentencia de un preso.

También relacionado con el despliegue, podemos tener un sesgo de retroalimentación. Este tipo de sesgo se produce cuando un modelo en sí mismo influye, a través de su efecto en el mundo real, en la generación de datos que se utilizan para entrenarlo. Los algoritmos de recomendación son especialmente susceptibles de estar afectados por este sesgo (Baeza-Yates, 2018).

4. Sistemas algorítmicos en contexto

4.1. El proceso de decisión y el proceso de predicción

La consideración de algoritmo como sistema no es la única generalización que es necesario hacer cuando se habla de éstos. En la literatura técnica es común mezclar los conceptos de decisión y de predicción, que, aunque relacionados, deben considerarse aparte. El sistema algorítmico exclusivamente realiza predicciones. Como tal, define modelos que nos permiten establecer medidas de probabilidad, riesgo, o cuantificar la perdida en términos de la diferencia entre lo deseado y lo obtenido. Un proceso distinto es el de la toma de decisiones. En teoría de la decisión (Von Neumann, 1944) se plantea el estudio de cuál es la mejor decisión para un agente racional cuando se enfrenta a la selección de distintas opciones

bajo incertidumbre. La teoría de la decisión necesita establecer para cada posible acción la probabilidad de cada potencial respuesta, así como el valor asociado al par acción-respuesta. Y es este punto el que permite la unión entre la decisión y la predicción. El algoritmo de predicción puede alimentar las probabilidades de las respuestas. Sin embargo, queda libre un punto básico en esta teoría utilitarista, y es el valor asociado al par. Es decir, si tomo una decisión A y el resultado es B, ¿qué valor tiene este resultado? El valor del producto cartesiano entre acciones y resultados se recoge en una matriz denominada *matriz de decisión*. En la literatura de aprendizaje automático no se hace esta distinción entre predicción y decisión y pragmáticamente se considera que la decisión que se debe tomar corresponde a la que tiene mayor probabilidad. Y efectivamente esto es así cuando la matriz de decisión asigna la misma utilidad a todas las decisiones. En ocasiones se habla muy livianamente de estos conceptos introduciendo el coste de los errores de un sistema algorítmico cuando se introduce la *matriz de confusión*. Esta matriz cuantifica la cantidad de predicciones en función del valor esperado e introduce los conceptos de verdaderos y falsos, positivos y negativos. Recordemos que los falsos positivos (FP) y los falsos negativos (FN) corresponden a los errores de predicción realizados. Asociados a esta matriz aparecen métricas parciales de rendimiento que fundamentaran las métricas estadísticas de medida de la discriminación.

En este punto nos gustaría ampliar el discurso que considera que el proceso de decisión se compone exclusivamente de un algoritmo monolítico. Supongamos un algoritmo entrenado para la predicción de una enfermedad infecciosa grave. Un sistema que minimiza la cantidad de errores cometidos implícitamente asume que el valor utilitario de los FP y de los FN es el mismo. Sin embargo, el contexto nos indica que, en este caso, no deberían tener el mismo valor. Un FN significa que el sistema predice 'no enfermedad' a un paciente que sí padece la enfermedad. El riesgo que comporta tomar una decisión basada en este sistema es muy alto. Por lo tanto, sería deseable que el número de FN sea el menor posible, idealmente cero, aunque eso implique aumentar el número de FP. Con esto evitaríamos que pacientes con la enfermedad quedasen sin tratar, con el riesgo epidemiológico que comportaría. Pero ¿hasta qué punto es admisible aumentar los FP? Es admisible diagnosticar una enfermedad a un paciente sano, tanto en cuanto el sistema de diagnóstico y de decisión no termina ahí. Pruebas complementarias o nuevos sistemas de predicción usados a continuación, pueden focalizarse en esta población para poder distinguir los casos de forma más fina. La visión monolítica del algoritmo independiente del proceso de decisión nos oculta esta visión que debe considerarse en el uso de estos sistemas.

4.2. El algoritmo me discrimina

¿Cómo sé que estoy siendo discriminado? ¿Se puede medir la discriminación algorítmica? ¿Tiene sentido hablar de algoritmos que discriminan? Y si es así ¿qué consideraciones se han de tener en cuenta? ¿Qué aproximaciones a éste?

Las anteriores preguntas, que aplican a cualquier agente de toma de decisiones, ponen de manifiesto la necesidad de un instrumento que permita identificar esta percepción. En un intento de formalizar la discriminación, la comunidad científica que estudia la justicia algorítmica ha propuesto múltiples métricas para poder identificar, medir y así ayudar a mitigar

los efectos de la disparidad bajo distintas perspectivas (Verma, 2018). Estas métricas se pueden dividir en tres grandes bloques en función de la corriente filosófica que siguen: desde el punto de vista del igualitarismo encontramos un conjunto de métricas estadísticas. Siguiendo un enfoque de justicia individual en la línea aristoteliana (Libro V de la *Ética nicomáquea*) de los "casos similares" se establecen un conjunto de métricas individuales basadas en el concepto de similitud. Finalmente, subyacente al discurso sobre el trato e impacto dispar se encuentra la identificación de las causas que dan pie a las métricas causales.

4.2.1. Medidas estadísticas

Desde un punto de vista del igualitarismo encontramos un conjunto de métricas estadísticas. En estas definiciones, tal y como se promueve en la corriente igualitaria, se pretende ecualizar las oportunidades indistintamente del colectivo y de los elementos no controlables o elegibles de los individuos. En este punto entra el concepto de identidad grupal como elemento no controlable y por el que dos individuos que pertenecen a dos colectivos distintos deben tener la misma consideración.

Para definir las métricas vamos a considerar tres elementos: llamaremos T al valor real esperado, O al valor de la predicción y P a la propiedad que define el grupo destacado.

Para motivar las medidas consideremos el caso de un proceso de decisión algorítmico que se usa para contratar potenciales candidatos a un trabajo. Imaginemos que tenemos sospechas de que el proceso de decisión puede tener un sesgo potencialmente discriminatorio en función del sexo. Para identificar una potencial discriminación podríamos preguntarnos las siguientes preguntas: *¿Se contratan de la misma forma candidatos del sexo masculino y del sexo femenino? ¿Del conjunto de los candidatos que el algoritmo ha predicho que podían ser potencialmente contratados, el algoritmo hace diferencias entre los dos colectivos?* o *¿Sobre todos aquellos que merecen ser contratados, la tasa de contratación es la misma independientemente del colectivo al que pertenecen?* Las tres preguntas son pertinentes y dan lugar a métricas diferentes. La primera, conocida como *independencia*, hace referencia a si la predicción es independiente del atributo sensible, y se puede formalizar como sigue

$$p(O=1 \mid P=a) = p(O=1 \mid P=b).$$

Usualmente se usa en relación a la decisión ventajosa y se la denomina *paridad demográfica o estadística*, y se encuentra asociada a la doctrina de discriminación directa o trato dispar.

La segunda, conocida como *suficiencia*, está relacionada con el concepto de precisión de la matriz de confusión y se formaliza

$$p(T=1 \mid O=1, P=a) = p(T=1 \mid O=1, P=b).$$

Así formalizada se la denomina *paridad predictiva* y si se exige para las decisiones positivas y negativas se denomina *igualdad de exactitud de uso condicional*.

Finalmente, la tercera, conocida como *separabilidad*, se relaciona con la sensibilidad, y se formaliza,

$$p(O=1|\ T=1, P=a) = p(O=1|\ T=1, P=b),$$

A esta formalización se la conoce como *igualdad de oportunidades*. Al igual que en los anteriores casos si se exige para las dos decisiones se obtiene la *igualdad de posibilidades* (equalised odds). Estas dos últimas se asocian a las formas de discriminación indirecta o impacto dispar.

Existen diversos teoremas que nos demuestran el delicado equilibrio entre estas tres definiciones, que son mutuamente excluyentes (Chouldechova, 2016). Esto significa que las tres medidas no se pueden cumplir a la vez. Por lo que, bajo este conjunto de medidas, cualquier agente que realice una decisión y use una o más de estas medidas para justificar que no realiza un trato dispar se puede demostrar que ejerce un impacto dispar, y a la inversa. Esto nos indica que el uso del concepto discriminación grupal en genérico es limitante y no se puede considerar como un problema de optimización matemática con solución única.

4.2.2. Medidas individuales

Por otro lado, encontramos las medidas basadas en la justicia individual y la visión aristotélica de *tratar los casos similares de forma similar* (Aristoteles, *Nicho*.). En otras palabras, argumenta que las personas deben ser tratadas según sus características y circunstancias individuales, en lugar de su identidad grupal. La identificación de esta medida se formaliza en las siguientes tres desigualdades:

A. $d(\text{'Y'},\text{'Z'})<\text{tolerancia}$

B. $p(O=1|I=\text{'Y'}) > p(O=0|I=\text{'Y'})$

C. $p(O=1|I=\text{'Z'}) < p(O=0|I=\text{'Z'})$

Dados los casos I='Y' y I='Z' correspondientes a dos individuos distintos, las desigualdades (B) y (C) indican que la decisión tomada para 'Y' y para 'Z' es distinta. La desigualdad (A) nos indica que la diferencia entre 'Y' y 'Z' es pequeña.

Este enfoque no está exento de dificultades puesto que se ha de definir la distancia entre los individuos representados por sus descriptores. Así como en los anteriores casos se enfatiza la noción de característica diferencial, en este caso nos plantea la cuestión de bajo que parámetros dos individuos son comparables. Por otro lado, la noción de justicia individual también tiene asociada la noción de grupo puesto que se compara el individuo con un colectivo definido por similitud (Binns, 2020). Y aún más, se produce un efecto de estandarización y de perdida de la noción de individualidad.

4.2.3. Causalidad y justicia contrafactual

Subyacente al discurso de disparidad se encuentra el concepto de causalidad. Sin embargo, los algoritmos descritos hasta el momento son algoritmos que explotan las correlaciones estadísticas entre los datos y la variable a predecir. Se mezclan causas y efectos. El estudio de la causalidad de una acción requiere de una intervención en el mundo real. Sin embargo, ésta puede ser irrealizable físicamente, tener un impacto inaceptable o no ético. Los modelos causales permiten estudiar el efecto causal a partir de datos observacionales bajo ciertas condiciones. Permiten responder preguntas causales a dos niveles: a nivel poblacional a partir de intervenciones y a nivel individual a partir de contrafactuales. Esto permite establecer nuevas métricas de igualdad en la denominada *justicia contrafactual* (Carey 2022). Formalmente, se extenderían las métricas anteriores bajo la perspectiva de las intervenciones. Por poner un ejemplo usando la notación causal, el concepto de independencia se puede extender al campo causal de la siguiente forma:

$$p(O=1 \mid Do(P=a)) = p(O=1 \mid Do(P= b)).$$

Donde el operador Do(P=a) indica la intervención que hace que un determinado atributo P tome el valor *a* para toda la población. Esta expresión se lee de la siguiente forma: Queremos que la probabilidad de la predicción ventajosa sea la misma cuando forzamos que la característica P tome el valor *a*, que cuando forzamos que la característica P tome el valor *b*. Esto respondería a ¿Cuál es el efecto causal sobre la proporción de contrataciones si la negociación la realiza un hombre o una mujer?. A nivel contrafactual se plantean preguntas más complejas y que incluyen los hechos y el contrafactual, como, por ejemplo, ¿se habría contratado a un candidato si su sexo fuese femenino (contra) sabiendo que no se ha contratado y su sexo es masculino (factual)? Y por lo tanto sería análogo causal a las medidas individuales.

4.2.4. Usando las métricas

El uso efectivo del concepto de discriminación requiere de la selección deliberada e intencional de una métrica particular en función del establecimiento de las prioridades correspondientes al contexto de aplicación. Si nos centramos en las medidas estadísticas, diversas guías (Ruf 2021) (Binns 2020) nos permiten identificar como usarlas. Por ejemplo, en aquellas ocasiones que el uso del sistema pretenda mitigar una desigualdad sistémica propiciando acciones que protejan a grupos menos privilegiados asumiendo que pretende revertir un sesgo estructural se entiende que se busca establecer políticas que obvian la causalidad. En este caso métricas basadas en independencia como la paridad demográfica sería recomendable. Si, por el contrario, consideramos la ausencia de sesgos estructurales deberíamos usar métricas derivadas de las de suficiencia o de separabilidad. O si consideramos las métricas de justicia individual, las métricas individuales no ajustadas velarían por las disparidades debidas a las decisiones personales, mientras que las ajustadas por el grupo ayudarían a mitigar sesgos estructurales. Usualmente la realidad nos plantea situacio-

nes complejas que acostumbran a mezclar ambas visiones y pueden requerir una selección amplia de métricas.

Complementaria a la visión estadística que sólo nos explicita la relación entrada-salida, la visión causal nos permite considerar las interacciones causales entre los distintos atributos. Esta visión es mucho más rica puesto que nos permite identificar explícitamente los mecanismos asociados a la discriminación directa e indirecta. Y, por lo tanto, medirlos. En este contexto se habla de discriminación de un camino causal si existe un efecto causal entre la variable protegida y el resultado siguiendo el camino que las une en el grafo causal. Usando el concepto de contrafactual asociado a los efectos directos, indirectos y espurios, (Plecko, 2023) establece una potencial reconciliación entre la independencia y la suficiencia, rompiendo efectivamente el teorema de imposibilidad. Para ello introduce el concepto de *necesidad de negocio*. Este concepto es un paralelo a la anterior intencionalidad del sistema y que permitía el uso de las métricas estadísticas. En este caso, la necesidad de negocio requiere identificar que atributos son atributos que deben obedecer a conceptos de impacto directo y cuales a impacto indirecto. Con esta identificación, se pueden medir las distintas magnitudes y establecer el equilibrio adecuado entre ellas.

5. Conclusiones

La necesidad de conectar una tecnología, cada vez más cercana a las actividades propias de los humanos, con los aspectos epistemológicos y normativos derivados de su desarrollo y uso requieren un punto de partida común entre diversas disciplinas, libre de *apriorismos* y simplificaciones estériles. En esta dirección hemos propuesto el uso de un concepto no normativo de discriminación que permite aclarar algunos aspectos epistemológicos. También hemos usado la imposibilidad de reducir los algoritmos de IA a un enfoque de diseño como su elemento definitorio principal, libre de tecnicismos innecesarios. Finalmente, hemos repasado el estado del arte en las métricas de discriminación, haciendo especial hincapié en los problemas asociados a su uso y sus potenciales soluciones. El presente texto pretende realizar la conexión de todos estos conceptos entendiendo que, como limitación de éste, un tratamiento exhaustivo de muchas de las ideas presentadas requiere de una discusión en mayor profundidad.

Aunque el nivel de la reflexión sobre la relación entre datos, algoritmos y decisiones ha avanzado mucho desde las propuestas dataístas de la década pasada, algunos de los desafíos identificados siguen sin una solución evidente, al tiempo que se crean nuevos problemas. A continuación, hacemos un repaso a algunos de los desafíos desde un punto de vista epistemológico y normativo.

5.1. Limites epistemológicos

El hecho que un sistema predictivo preciso no constituye necesariamente una buena base para la toma de decisiones es una evidencia científica que ha impactado en la investigación en IA de forma reciente y con resultados desiguales. Dada una aplicación, es necesario evaluar es si el objetivo propuesto es de naturaleza puramente predictiva o tiene alguna carac-

terística intervencional que impida una toma de decisiones basada en datos observacionales (Fernández-Loría, 2022). Las técnicas de inferencia causal representan la mejor aproximación a este problema (Pearl, 2018), pero su viabilidad a gran escala está aún por demostrar.

En algunas situaciones, la discriminación puede ser compuesta y aditiva, es decir, puede estar basada en muchos pequeños actos de discriminación que desembocan en una consecuencia grave al cabo del tiempo. Las métricas de discriminación no son capaces de detectar estas situaciones, que sólo pueden ser detectadas con métodos cualitativos (Narayanan, 2022).

Las métricas basadas en grupos, en general, tienden a ignorar los méritos de cada individuo en el grupo. Algunas personas pueden ser mejores para una tarea determinada que otras personas del mismo grupo, lo que no se refleja en las definiciones de equidad basadas en grupos (Mittelstadt, 2023). Este problema puede dar lugar a dos comportamientos problemático: (a) la profecía autocumplida en la que, al elegir deliberadamente a los miembros menos calificados del grupo protegido, colaboramos en la construcción de un mal historial para el grupo, y (b) el tokenismo inverso, donde al no elegir a un miembro bien calificado del grupo no protegido, uno de los objetivos del sistema se convierte en crear refutaciones convincentes para los miembros del grupo protegido que tampoco son seleccionados.

5.2. Problemas normativos

Desde el punto de vista normativo hace falta avanzar en un concepto de legitimidad para el uso de los algoritmos de IA en la toma de decisiones que considere, tal y como lo hace en el campo de la filosofía política, las condiciones de legitimización a la vez que las consecuencias de la renuncia a su uso (Martin, 2022).

Los avances tecnológicos han abierto de nuevo la definición de grupo protegido y la necesidad de considerar a los llamados grupos algorítmicos (Wachter, 2022). Estos son los grupos creados a partir de técnicas de perfilado algorítmico, que no se correlacionan con grupos legalmente protegidos. La opacidad en su uso por parte de las grandes compañías tecnológicas abre la posibilidad de una nueva fuente de discriminación oculta que hace falta clarificar.

Por último, hace falta reconsiderar la relación entre las características usadas por los modelos predictivos y su función (Creel, 2022). Estos modelos pueden usar tres tipos de características: características legítimas, características protegidas, y características arbitrarias. Supongamos un proceso de decisión para la concesión de un préstamo. El sueldo podría estar entre las primeras, el género entre las segundas, y el número de ascensores en el edificio en el que se halla el domicilio habitual podría considerarse arbitraria. ¿En qué casos es ético el uso de características arbitrarias? ¿En qué casos no se deberían usar nunca características arbitrarias?

Referencias

Aristoteles (1984), *Nichomachean Ethics,* Princeton University Press, Vol.3.1131a10–b15

Baeza-Yates, R. (2018), "Bias on the web". Commun. ACM. 61, pp. 54–61.

Barocas, S.; Selbst, A. D. (2016), "Big Data's Disparate Impact", Cal. Law Review, Vol.104

Binns R. (2020), "On the apparent conflict between individual and group fairness", ACM Proceedings of Int. Conf. on Fairness Accountability and Transparency in Machine Learning.

Brooks, D. (2013), "Opinion | The Philosophy of Data". New York Times, https:// www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html

Carey, A.; Wu, X. (2022), "The Causal Fairness Field Guide: Perspectives from social and formal sciences", Frontiers in Big Data, Vol 5.

Chouldechova, A. (2017), "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". Big data, 5(2), pp. 153-163.

Creel, K.; Hellman D. (2022), "The algorithmic Leviathan: arbitrariness, fairness, and opportunity in algorithmic decision-making systems." Canadian Journal of Philosophy 52.1. pp. 26-43.

Dennett, D. C. (1987), *The intentional stance*. MIT Press.

Eidelson, B. (2015), *Discrimination and Disrespect*. Oxford University Press.

EU P Serv (2019) "A Governance Framework for algorithmic accountability and transparency". Recuperado de: "https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2019)624262"

Fazelpour, S.; Danks, D. (2020), "Algorithmic bias: Senses, sources, solutions." Philosophy Compass 16.8: e12760.

Fernández-Loría, C.; Foster P. (2022), "Causal decision making and causal effect estimation are not the same… and why it matters." INFORMS Journal on Data Science 1.1, pp. 4-16.

Guersenzvaig, A.; Casacuberta, D. (2022), "La quimera de la objetividad algorítmica: dificultades del aprendizaje automático en el desarrollo de una noción no normativa de salud", IUES ET SCIENTIA, Vol 8 N 1, pp. 35-56.

Harari, Y. N. (2015), *Homo Deus: A Brief History of Tomorrow*. Random House. Traducción al castellano de la editorial Debate.

Hardt, M.; Recht, B. (2022), *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press.

Jacobs, A. Z.; Wallach, H. (2021), "Measurement and fairness". In Proceedings of the 2021 ACM Conference on fairness, accountability, and transparency, pp. 375-385.

Johnson, R. A.; Zhang, S. (2022) "What is the Bureaucratic Counterfactual? Categorical versus Algorithmic Prioritization in US Social Policy". In 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1671-1682.

Lazar, S. (2022), "Legitimacy, Authority, and the Political Value of Explanations". arXiv preprint arXiv:2208.08628.

Lippert-Rasmussen, K. (2014), *Born Free and Equal?*. Oxford University Press.

Martin, K.; Waldman, A. (2022), "Are algorithmic decisions legitimate? The effect of process and outcomes on perceptions of legitimacy of AI decisions". Journal of Business Ethics, pp. 1-18.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. (2021), "A Survey on Bias and Fairness in Machine Learning". ACM Comput. Surv. 54, 6, Article 115, pp. 1-35.

Mitchell, T. M. (1980), "The need for biases in learning generalizations ".New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ., pp. 184-191.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; Lum, K. (2021), "Algorithmic fairness: Choices, assumptions, and definitions." Annual Review of Statistics and Its Application, 8, pp. 141-163.

Mittelstadt, B. D.; Allo, P.; Taddeo, M.; Watcher, S.; Floridi, L. (2016), "The ethics of algorithms: mapping the debate". Big Data & Society, pp. 1-26.

Mittelstadt, B., Wachter, S., Russell, C.s (2023), "The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default". Available at SSRN: https://ssrn.com/abstract=4331652

Narayanan, A. (2022), "The limits of the quantitative approach to discrimination." James Baldwin lecture [transcript], Princeton University.

Pearl, J.; Mackenzie, D. (2018), *The book of why: the new science of cause and effect*. Basic books.

Plecko, D; Bareinboin, E. (2023) "Reconciling predictive and statistical parity: A causal approach", arXiv:2306.05059v1.

Pratt, L. Y. (1993), "Discriminability-based transfer between neural networks" (PDF). NIPS Conference: Advances in Neural Information Processing Systems 5. Morgan Kaufmann Publishers. pp. 204–211.

Rudin C. (2019), "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". Nat. Mach. Intell. 1(5), pp. 206–15.

Ruf, B.; Detyniecki, M. (2021), "Towards the Right Kind of Fairness in AI", arXiv:2102.08453v7.

Seng, M.; Floridi, L.; Singh, J. (2021), "Formalising tradeoffs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics". AI & Society, 1:529–544.

Simon, J.; Wong, P.; Rieder, G. (2020), "Algorithmic bias and the Value Sensitive Design approach". Internet Policy Review, 9(4):1-16.

Sternberger, D. (1968), "Legitimacy" in International Encyclopedia of the Social Sciences (ed. D.L. Sills) New York: Macmillan, Vol. 9, p. 244.

Unceta, I. (2020), "Environmental Adaptation and Differential Replication in Machine Learning", Entropy (Basel). 3;22(10):1122.

Verma, S.; Rubin, J. (2018), "Fairness definitions explained". IEEE/ACM Int Workshop on Software Fainess,

von Neumann, J.; Morgenstern, O. (1944), *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Wachter, S. (2022), "The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law." arXiv preprint arXiv:2205.01166.

Zerilli, J. (2022), "Explaining Machine Learning Decisions". Philosophy of Science, 89(1), pp. 1-19. doi:10.1017/psa.2021.13

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 81-94
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.563231

Más allá de los datos: la transformación digital del museo tradicional*

Beyond Data: The Digital Transformation of the Traditional Museum

*ALGER SANS PINILLOS***
*VICENT COSTA****

Resumen. Este trabajo se centra en el museo virtual entendido como la transformación digital del museo tradicional. En primer lugar, se revisan las principales conceptualizaciones del museo virtual en la literatura especializada y, a partir de las mismas, se propone una definición básica. Asimismo, se presentan argumentos que muestran la insufi-

Abstract. This work focuses on the virtual museum, understood as the digital transformation of the traditional museum. First, the main conceptualizations of the virtual museum in the specialized literature are reviewed, and based on them, we propose a basic definition. Furthermore, we argue in favor of the insufficiency of the dataist

Recibido: 30/03/2023. Aceptado: 14/06/2023.

* Ambos autores han participado en todo el proceso de elaboración del trabajo y el orden de autoría corresponde alfabéticamente al nombre de los autores.

** Università degli Studi Umanistici di Pavia (UNIPV) <alger.sanspinillos@unipv.it>. Investigador postdoctoral en Filosofía de la Ciencia en la *Sezione di Filosofia* del *Dipartimento di Studi Umanistici* de la UNIPV. Sus líneas de investigación actuales son el razonamiento abductivo, la moralidad inherente de los espacios, los fundamentos biológicos y socioculturales de la creatividad y la implementación de la inferencia abductiva en didáctica de las ciencias naturales. Actualmente es miembro del *Computational Philosophy Laboratory* en el marco del proyecto *Ignorance in the Perspective of an Ecology of Cognition: Cognitive Niches, the Extended Mind, and Ignorance-Based Reasoning*. Publicaciones recientes: Sans Pinillos, A. (2022). Neglected Pragmatism: Discussing Abduction to Dissolute Classical Dichotomies. *Found Sci 27*, 1107–1125, https://doi.org/10.1007/s10699-021-09817-x; Magnani, L., Sans Pinillos, A., Arfini, S. (2022). Language: The "Ultimate Artifact" to Build, Develop, and Update Worldviews. *Topoi 41*, 461–470, https://doi.org/10.1007/s11245-021-09742-5. ORCID: https://orcid.org/0000-0002-8817-7286.

*** Instituto de Investigación en Inteligencia Artificial del Consejo Superior de Investigaciones Científicas (IIIA-CSIC) <vicent@iiia.csic.es>. Científico Titular del IIIA-CSIC, miembro del Institut d'Història de la Ciència (Universitat Autònoma de Barcelona) y del Grupo de Estudios Humanísticos sobre Ciencia y Tecnología (GEHUCT, 2021 SGR 00517). Esta publicación es parte del proyecto PID2022-139835NB-C21, financiado por MCIN/AEI/10.13039/501100011033/. Sus líneas de investigación actuales son las lógicas difusas para la inteligencia artificial explicable y la ética e inteligencia artificial. Publicaciones recientes: Costa, V., Alonso-Moral, J.M., Falomir, Z., & Dellunde, P. (2023). An art painting style explainable classifier grounded on logical and commonsense reasoning. *Soft Computing*, https://doi.org/10.1007/s00500-023-08258-x; Tarrés-Puertas, M.I., Costa, V., Pedreira Alvarez, M., Lemkow-Tovias, G., Rossell, J.M., & Dorado, A.D. (2023). Child–Robot Interactions Using Educational Robots: An Ethical and Inclusive Perspective. *Sensors* 23(3):1675, https://doi.org/10.3390/s23031675. ORCID: https://orcid.org/0000-0001-6352-7238.

ciencia de la perspectiva dataísta en el estudio de la injusticia epistémica relativa al museo virtual. Finalmente, se analiza la exclusión relativa a la participación y la interacción entre las visitantes de un museo (tanto virtual como físico) y la propia institución.
Palabras clave: museo virtual, museo digital, injusticia epistémica, dataísmo, irradiación.

perspective in the study of epistemic injustice associated with the virtual museum. Finally, the exclusion related to the participation and interaction between a museum's visitors and the institution is analyzed.
Keywords: virtual museum, digital museum, epistemic injustice, dataism, irradiation.

1. Introducción

La transformación digital ha influido (e influye, pues los procesos relacionados con este fenómeno siguen en desarrollo) en un abanico muy diverso de áreas donde se incluyen, entre otras muchas, la industria, la educación, el cuidado de la salud y, por supuesto, la cultura. De este modo, y en la medida en que esta adopción de las tecnologías digitales ha supuesto una disrupción respecto a nuestra forma de estar en el mundo, los nuevos espacios digitales y aquellos híbridos que derivan de la integración de estos últimos en los espacios físicos plantean cuestiones y problemáticas irresolubles por medio de conceptualizaciones *analógicas,* propias de una etapa anterior. Considérense, a modo de ilustración, los interrogantes relativos a la autoría de una obra artística que suscitan algunos sistemas de inteligencia artificial actuales (Epstein et al., 2020), la necesidad de legislar, teniendo en cuenta el nuevo escenario digital, aspectos relacionados con términos como, por ejemplo, el acoso sexual en el trabajo (Ramirez et al., 2023) o los retos que plantea la integración de sistemas de inteligencia artificial en el aprendizaje a la evaluación educativa, entre los que se podría incluir una reconceptualización del proceso de aprendizaje (Niu, 2022; Schiff, 2021). Asimismo, este paradigma actual no sólo permite sino que exige una ampliación de nuestra visión ética, siendo la relación humana con la tecnología uno de los retos principales (Casacuberta y Guersenzvaig, 2019).

Este artículo se centra en el museo virtual entendido como la transformación digital del museo tradicional, cuyo espacio es esencialmente físico (véase Sección 2). Esta transformación se da tanto respecto a los museos de arte como a los museos de historia social y cultural, si bien es menester mencionar que tal distinción no siempre es unívoca, e incluso algunas autoras han entendido el museo de arte como un museo desfuncionalizado (Groys, 2014). No obstante, este trabajo se acota a los museos de arte por su relevancia en el imaginario colectivo[1]. Tomando la hipótesis de que la virtualidad de este espacio posibilita un marco desde el cual combatir las injusticias epistémicas del museo tradicional, revisamos en la sección 2 las principales conceptualizaciones del museo virtual en la literatura especializada y, a partir de las mismas, proponemos una definición básica que las integra. Asimismo, se presentan argumentos sobre la insuficiencia de la perspectiva dataísta en el estudio de la injusticia epistémica mencionada. Por último, analizamos la exclusión relativa a la participación y la interacción entre las visitantes de un museo (tanto virtual como físico) y la propia institución.

1 Agradecemos a la revisora anónima 1 por hacernos ver la importancia de esta distinción (museo de arte y museo social y cultural) en nuestro proyecto. Sin embargo, teniendo en cuenta las limitaciones de extensión de este trabajo, un estudio similar relativo a los museos de historia social y cultural se desarrollará en un trabajo futuro.

2. El concepto de «museo virtual»

Este trabajo se centra en el museo virtual entendido como el resultado de la transformación digital aplicada al museo tradicional, concebido este último como la «institución sin ánimo de lucro, permanente y al servicio de la sociedad, que investiga, colecciona, conserva, interpreta y exhibe el patrimonio material e inmaterial» (ICOM, 2022), cuyo espacio principal es físico, si bien puede presentar elementos auxiliares virtuales (por ejemplo, un sitio web o una aplicación móvil). Ahora bien, esta definición no es unívoca, o siquiera totalmente consensuada, y cabe destacar que la heterogeneidad de propuestas y el desacuerdo son todavía mayores en lo que respecta al museo virtual. Ciertamente, tal y como señalan (Latham & Simmons, 2014), la conceptualización del museo virtual engloba desde un conjunto de obras u objetos digitalizados hasta una experiencia inmersiva, a través de la tecnología de la realidad virtual, que emula la experiencia de un museo tradicional. Asimismo, tampoco hay unanimidad ni en cuanto a la terminología del concepto (en ocasiones, se habla también de *museo en línea*, *museo digital*, *hipermuseo*, *web museo*, *museo electrónico* y *cibermuseo*), ni en cuanto a la primera aparición del término (Schweibenz, 2019). En cualquier caso, un análisis filosófico del museo virtual y de las relaciones que articulamos con él requiere, al margen de la disparidad mencionada, establecer una definición del concepto como punto de partida del estudio. En el presente digital, propiciado por el desarrollo de la inteligencia artificial, la ciencia de datos masivos y la tecnología de realidad virtual, determinar una definición tal se torna más complejo si cabe; pero, al igual que ocurre con otros conceptos *digitalizados*, la cotidianeidad alimenta la urgencia de abordar las problemáticas relacionadas con el mismo. Dedicamos, pues, el resto de la sección a ello.

En este artículo, solo usaremos la expresión *museo virtual*, y proponemos una definición *minimal* de museo virtual que trata de recoger los aspectos esenciales de la noción, aunque sin la pretensión última de establecer categorizaciones unívocas a partir de la misma. Recogiendo la definición general propuesta por el ICOM para el museo, diremos que un museo virtual es el espacio principalmente virtual (y, en particular, digital) donde se *investiga, colecciona, conserva, interpreta y exhibe el patrimonio material e inmaterial*. Entendemos que la virtualidad de un espacio viene dada por la disrupción de las propiedades, leyes o dimensiones espacio-temporales que se dan en el escenario físico (Rodríguez de las Heras, 2004), con lo que la naturaleza principal del espacio diferencia, en parte, ambos tipos de museo, el virtual y el tradicional. Asimismo, el espacio digital se caracteriza por crearse a partir de la tecnología y es virtual en la medida en que las leyes que lo gobiernan difieren de las propias del espacio físico o natural, si bien es menester recordar que no todo espacio virtual es digital (piénsese, por ejemplo, en el espacio creado en los sueños)[2]. En definitiva, mientras que la fisicalidad es una condición sine qua non para los museos tradicionales, el museo virtual requiere necesariamente un espacio digital. De este modo, con esta definición básica recogemos, por una parte, una variedad muy amplia de propuestas distintas que se encuentran en la literatura especializada (por ejemplo, (Hazan et al., 2014; Negri, 2012; Djindjian, 2009)). Por otra parte, a partir de esta definición se descartan algunas conceptualizaciones de museo virtual que, a nuestro juicio, corresponden a otras nociones

2 Nótese que el uso del término *museo virtual* viene dado por una adecuación a la notación mayoritaria en la literatura especializada, pues consideramos que la noción de *museo digital* sería más rigurosa, teniendo en cuenta la contextualización presentada en este artículo.

que si bien están relacionadas, no engloban la idea de museo virtual y, por lo tanto, no se prestan al análisis que proponemos.

En primer lugar, subrayamos que la condición necesaria de digitalidad no limita el museo virtual a un contenido independiente por completo del arte tangible, como sería el caso de un museo virtual cuyas exhibiciones se limitan a obras de arte digital y cuyo espacio se ciñe estrictamente al digital. Aunque algunas propuestas importantes (como, por ejemplo, la expuesta en (Sacher, 2017) o el Virtual Online Museum of Art[3]) han concebido los museos virtuales de este modo, en este trabajo hablaremos de museos estrictamente virtuales para referirnos a esta subclase del museo virtual que hemos definido y, en general, el museo estrictamente virtual no será objeto de estudio. De hecho, tal y como puede inferirse de los argumentos expuestos en la sección posterior, los museos estrictamente virtuales comparten solo algunas de las problemáticas asociadas al museo virtual y se prestan solo en ocasiones a nuestro modelo de análisis.

Asimismo, la definición de museo virtual no hace mención de ningún espacio físico, con lo que no se limita el museo virtual a un mero gemelo digital[4] de un museo tradicional. Más aún, en la medida en que se define el espacio virtual como principal, los gemelos digitales de museos tradicionales no se incluyen en nuestra categorización de museo virtual. Consideramos que esta omisión no es especialmente relevante, pues, por un lado, este tipo de museos, entendidos como gemelos virtuales y que en la actualidad integran sobre todo tecnologías de realidad virtual (Banfi et al., 2023), se contagian trivialmente de algunas de las problemáticas y cuestiones éticas y epistemológicas propias del museo tradicional y no presentan retos nuevos ni necesitan de análisis diferentes a los propios del museo tradicional (a fin de cuentas, idealmente, un gemelo digital de un museo tradicional sería indistinguible de este último). Por otra parte, no descartamos aquellos museos virtuales inspirados, asociados o relacionados con museos tradicionales (los cuales son de hecho los más comunes en la actualidad), es decir, los museos virtuales que mantienen algún tipo de relación con otros museos tradicionales. Aunque esta relación se ha analizado en profundidad, todavía restan por esclarecer muchos aspectos de la misma (Schweibenz, 2019; Chalmers et al., 2008; Bandelli, 1999) y, de cualquier modo, una reflexión sobre ella queda fuera del objetivo del presente trabajo.

Terminamos la sección con la discusión de dos consideraciones respecto a nuestra conceptualización del museo virtual. Por un lado, se consideran los museos virtuales en el contexto actual. Esta precisión es importante, pues hay un reto técnico en el diseño de estos espacios de importancia no solo en lo que concierne a los aspectos científico-tecnológicos, sino también en lo que respecta al análisis filosófico del concepto. Ciertamente, cumplir con algunas de las funciones indicadas en la definición de museo virtual requiere el uso de la tecnología puntera actual, de manera que muy a menudo integran en el diseño sistemas de inteligencia artificial y tecnologías relacionadas con el internet de las cosas. De este modo, el diseño del sistema (informático) del museo virtual, en la medida en que recoge diversas tecnologías como las mencionadas, no puede entenderse sin considerar la relación constitutiva que la ciencia y la ingeniería mantienen con la sociedad en la que se desarrollan (Wagner, 2022). Así pues, cuestiones sociales y culturales se integran en el propio diseño tecnológico

3 VOMA: https://www.voma.space/.

4 Un gemelo digital replica un objeto, proceso o sistema del mundo físico, y existen diferentes categorías de gemelos digitales (por ejemplo, el urbano, el cognitivo o el histórico). Para una revisión actual y detallada de este concepto, véase (Luther, 2023).

del museo virtual, mientras que, en muchas ocasiones, es la dimensión económica aquello que, vertebrando los aspectos sociales y culturales, prevalece en la toma de decisiones.

Por otro lado, es menester contextualizar las funciones del museo virtual recogidas en nuestra definición y propuestas por el ICOM, teniendo en cuenta la naturaleza digital de este tipo de museo. Por ejemplo, la conservación de obras artísticas en este contexto deviene una actividad cualitativamente distinta a la correspondiente en el espacio físico, o también *prestarse al servicio de la sociedad* adquiere matices distintos cuando se trata de un entorno digital, como se mostrará en la sección siguiente. Teniendo esto en cuenta, no es de extrañar que ya se le haya dado otros usos al museo virtual. Por ejemplo, en (Daviddi et al., 2022), los autores utilizan un museo virtual para realizar un experimento sobre la compleja relación entre la codificación de la memoria, su reactivación y la probabilidad de informar sobre recuerdos falaces. Ahora bien, también se han mantenido otros usos clásicos, aunque menos frecuentes, del museo tradicional. Por ejemplo, el de la divulgación científica (Banfi et al., 2023).

3. Justicia epistémica y museo virtual: las limitaciones de los datos

Debido al tipo de valor que gestiona y, en ocasiones, genera, el museo tradicional es una institución especialmente sensible a causar el tipo de injusticia identificada como epistémica, esto es, aquella que se da cuando la capacidad epistémica de una persona o un grupo de personas se pone en tela de juicio por el hecho de pertenecer a un determinado colectivo (c.f., Fricker 2007). En la sección anterior ya se han mencionado las funciones de conservación y preservación de los museos, tanto de los virtuales como de los tradicionales. Ahora bien, una parte importante del patrimonio cultural, especialmente aquel vinculado al arte, se acumula en los museos (sobre todo en los tradicionales) y tiende a invisibilizar las minorías y exaltar los valores elitistas de las sociedades. De esta forma, vertebrada por los prejuicios social y culturalmente arraigados en una comunidad, esta dimensión social en la base de la injusticia epistémica sugiere tomar asimismo la noción de injusticia social en nuestro análisis, entendida esta última como «la discrepancia entre lo que es y lo que debería ser» (Opotow, 2011).

En el caso de los museos, el elitismo se manifiesta en los productos culturales exhibidos, los cuales siguen, principalmente, la norma del gusto que considera y mezcla el valor estético con un nivel cultural elevado y de sofisticación. Con frecuencia, se ha tratado de legitimar esta discriminación aludiendo a una supuesta necesidad impreriosa: al fin y al cabo, las limitaciones propias de los catalógos y las restricciones espacio-temporales requieren una selección de obras que, necesariamente, excluirá al conjunto de las restantes[5]. En lo que concierne a los museos virtuales, podría creerse que la exposición de otro tipo de obras artísticas al margen de las canónicas se posibilita, justamente, gracias a la virtualidad de los mismos. Concretamente, bien gracias a la digitalización de la totalidad del catálogo del museo tradicional asociado o bien gracias a una digitalización más ambiciosa, basada en grandes bases de datos, podría argüirse que las limitaciones intrínsecas del museo tradicional desaparecen en el contexto de un museo virtual. Esta idea,

5 Si bien un estudio del museo tradicional queda fuera del trabajo presentado en este artículo, creemos conveniente mencionar que la experiencia atestigua la flaqueza de este argumento, pues generalmente los museos con grandes recursos mantiene esta homogeneidad en sus exposiciones principales, mientras que las exhibiciones al margen del canon imperante se suelen relegar a exposiciones temporales o a meras eventualidades.

por cierto, no es reciente, pues hace algo más de dos décadas, (Hertzum, 1998) entendió el museo virtual como una manera de quebrar las limitaciones espaciales del museo tradicional. En efecto, si bien puede considerarse razonable el hecho de que la flexibilidad y el uso de grandes bases de datos pueden contribuir a paliar las formas de exclusión perpetradas en el modelo de museo tradicional, es menester tener adoptar una perspectiva que no se limite a los postulados dataístas.

Indudablemente, a menudo el diseño de un museo virtual exige la digitalización de obras y objetos presentes en museos tradicionales, y el hecho de que con frecuencia esta digitalización no sea inmediata ni trivial[6] obliga a una toma de decisiones relevante, influenciada también por los intereses y valores predominantes en el imaginario de los responsables de la decisión. De esta forma, podría sugerirse una digitalización de *todas las obras existentes* para atenuar esta exclusión y, con ello, también el poder de quienes pueden decidir sobre dichos asuntos; sin embargo, la propuesta resultaría un bálsamo que, por sí solo, no devendría solución.

En efecto, por un lado, una base de datos con *todas las obras existentes* incluiría exclusivamente la totalidad de aquello que se concibiera como «obra» y cuya concepción ya estuviera aceptada previamente por los mismos criterios elitistas que se imponen en la mayoría de decisiones relativas a los museos tradicionales. En este sentido, no se daría, pues, ningún cambio significativo. Nótese además que esta línea argumentativa diverge la posición teórica del dataísmo, pues se defiende la insuficiencia del dato en un contexto como el de nuestro análisis, esto es, un enfoque donde la injusticia epistémica, irreducible a la mera dimensión de la episteme, determina el estudio. Se entiende aquí por dato una obra artística digitalizada sin ningún tipo de clasificación o etiquetado (por ejemplo, se consideraría que la digitalización facilitada por Wikipedia.org[7] del cuadro *Los Nenúfares* de Claude Monet es, efectivamente, un dato). Así, la invisibilización no se da necesariamente por el acto deliberado de suprimir la representación: las minorías invisibilizadas no lo son por su tamaño, sino por su reducida representación social. Sin ir más lejos, quienes toman las decisiones sobre el día a día de la ciudadanía son una minoría, que, ciertamente, no está invisibilizada en un sentido representativo.

Indudablemente, las personas que *visitan* un museo virtual han integrado en su realidad social los prejuicios que perpetúan la educación reglada y, en definitiva, el proceso de socialización, con lo que la posibilidad de contemplar *cualquier* obra en un entorno digital no evitará que solo sean de objeto de atención las obras más conocidas por el público general[8]. Es decir, incluso partiendo de una base de datos equilibrada y completa, la exclusión se manifestaría a posteriori, con las decisiones de las usuarias. Más allá de los datos, se necesita, por consiguiente, un diseño del museo virtual que visibilice los colectivos excluidos y no decaiga en el ardid de la suficiencia de los datos.

6 Piénsese, por ejemplo, en la digitalización de una escultura en 3D con el fin de integrarla en un museo virtual cuyo espacio digital se construya con técnicas de realidad virtual.

7 https://upload.wikimedia.org/wikipedia/commons/7/70/Los_nenúfares_(Monet).jpg.

8 Este análisis puede extrapolarse, al margen del impacto de la publicidad, a la cuestión de los datos de reproducciones de plataformas de música como Spotify, en la que un número muy reducido de autores (y, de entre ellos, un conjunto de canciones en particular) son quienes se escuchan mayoritariamente. Aún la persona que quiere huir de las tendencias populares, tiende a caer en su propia red de consumo habitual, reduciendo mucho el espectro de música escuchada, por el propio diseño del algoritmo de recomendación basado en el aprendizaje por refuerzo, esto es, el área del aprendizaje automático que busca maximizar una función de recompensa.

Por otro lado, al no ser las obras catalogadas como, por ejemplo, «obra de arte», se exacerbaría aún más la invisibilización pues, se incluirían directamente sin contar con el mínimo reconocimiento de ser expuesto en un museo, aunque con una etiqueta sesgada. Por ejemplo, las etiquetas de «primitivo» o «tribal» pueden contribuir a perpetuar la idea de que las culturas no occidentales están menos desarrolladas. Esto se hace palmario cuando vemos que la primitividad de una obra de arte religiosa africana no se considera en una obra de arte religiosa europea (incluso si esta última es cronológicamente anterior a la primera). Aún más, seguramente la obra de arte religiosa europea estará expuesta en un museo de arte (por ejemplo, el *Pantocrátor de San Clemente de Tahull*[9]), mientras que la africana se ubicará en un museo de etnografía (por ejemplo, la talla de madera *Osun, diosa del agua dulce y de los ríos, acompañada de asistentes*[10,11]). En el espacio virtual, esta circunstancia se manifestaría en la invisibilización del arte catalogado desde la exclusión porque su etiqueta no correspondería con la usada para el catálogo.

En definitiva, cuando los museos perpetúan estas dos formas de exclusión presentadas (es decir, el elitismo y la invisibilización), la *realidad social*, entendida como el resultado de la relación entre los mundos personales (microscópicos) y las estructuras sociales (macroscópicas) (Brewer, 2004), se restringe.

4. Exclusión tradicional: entre el museo físico y el museo virtual

La injusticia epistémica también se da en lo que concierne a la participación y relación que se establece entre las visitantes de un museo y la propia institución. En efecto, la invisibilización de la que se ha hablado en la sección anterior, potenciada por la falta de representación en las colecciones exhibidas, teje una realidad social donde ciertos grupos encuentran ajena la idea de acudir a un museo. Aunque, ciertamente, no sea aplicable en todos los casos (Rectanus, 2006), el modelo predominante de museo es aquel cuya función social de conservación, colección y exposición invita a sus visitantes al tipo de contemplación ociosa (Prior, 2006) y a la búsqueda de experiencias estéticas (Bell, 2017), o incluso reparadoras (Kaplan, et al., 1993), reservadas a quienes hayan tenido la oportunidad de aprender a apreciar y aprovechar este tipo de actividades.

Esta situación plantea una problemática sobre la responsabilidad de la exclusión. En efecto, las injusticias pueden perpetuarse involuntariamente, lo que, aplicado al caso que nos ocupa, significa que los museos pueden devenir espacios de exclusión por el simple uso que se hace de ellos. De este modo, en nuestro análisis ha de considerarse la dimensión ética que los espacios representan para las personas. El museo tradicional, como espacio construido, es un artefacto moral porque las acciones que se llevan a cabo en él pueden ser resignificadas y, con ello, tener un impacto social. Dicho de otra manera, no hay neutralidad, ni en los museos ni en las actividades que se desarrollan en ellos. Teniendo en cuenta que una estructura es incapaz de reaccionar ante las acciones, sería de esperar que son las visitantes de los museos

9 https://www.museunacional.cat/ca/search/content/pantocrator.

10 https://cataleg.museuetnologic.bcn.cat/fitxa/africa_meb/H422936/?lang=es&resultsetnav=5f02011e62195.

11 Hay, también, un conflicto entre lo pagano y lo religioso. Mientras que Osun es presentada como una diosa de un panteón mitológico, el dios cristiano es presentado como una "realidad" (Dios) distinta del pensamiento mágico de las creencias de otras civilizaciones.

y sus directoras quienes perpetúan la exclusión. No obstante, tal conclusión es falsa, pues los mecanismos de exclusión de estos museos no actúan directamente sobre las visitantes, sino que se dan a través de la influencia que ejerce la idea de museo compartida social y culturalmente.

En este sentido, se mantiene una relación axiológica con el museo donde se le atribuye un valor (por ejemplo, el valor atribuido al arte y a la experiencia estética), por lo que la institución deviene un mediador moral inerte, esto es, un distribuidor pasivo de la moralidad humana (Magnani y Bardone, 2007; Magnani, 2018) indirecto y sin capacidad para representar los valores que se le atribuyen (Sans Pinillos, 2023). La principal característica de los artefactos inertes es que su capacidad para distribuir valores reside en las actividades que se pueden realizar con o en él. Por lo tanto, será durante el desarrollo de estas actividades que habrá distribución axiológica (estética y moral).

Esta distribución axiológica puede asimismo analizarse, tal y como la revisora anónima 1 ha sugerido, desde un enfoque donde las actividades que se realizan con y en el museo se conciben como un sistema multi-agente. En este, la agencia de las conservadoras, entendida como la agencia humana tradicional de la filosofía de la acción (Ferrero, 2022; Lier, 2023), se encuentra con las agencias- colectiva e individual- de las visitantes, esto es, con sus biografías, expectativas, intenciones y demás. En cuanto a la interacción específica entre la usuaria y un museo digital, este enfoque basado en la agencia podría enriquecerse con la consideración de un marco donde la la conceptualización de la agencia no prioriza el agente humano respecto al no-humano (en este caso, el museo digital), como por ejemplo se propone en (Dattathrani, 2023).

Esta forma de distribución puede llamársele radiación: la forma como la estructura deviene moral a partir de la concepción previa que se tiene se tiene de ella y la influencia que ejerce en las acciones de las personas. La contraparte de la radiación es la irradiación axiológica, la cual propone condiciones que permiten a los agentes interactuar con el entorno a través de la influencia que la estructura ejerce. Así pues, la distribución axiológica surge por tres vías de radiación-irradiación (Sans Pinillos, 2023):

1. por la comprensión de las actividades que se realizan,
2. por la imitación del resto de los agentes, y
3. por la misma predisposición que ofrece el artefacto.

En este sentido, para esclarecer las causas de la exclusión y poder, así, proponer una línea de subsanación, es importante distinguir los roles de quienes participan en la radiación axiológica de los museos. Al respecto, tomamos la distinción cognitiva de Feyerabend entre las cuestiones de las participantes y las de las observadoras en el contexto de la filosofía de la ciencia (Feyerabend, 1978: 18). En breve, son las observadoras quienes analizan los resultados desde la perspectiva histórica (como mínimo, de la distancia temporal) de una investigación acabada; y las participantes son quienes realizan la investigación. En lo que respecta a los museos, la distinción de Feyerabend se aplica a las expertas y las visitantes. Por un lado, las participantes serían las visitantes, quienes acuden a los museos e interactúan como meras espectadoras con lo expuesto. Por otro lado, las observadoras corresponderían a las expertas, quienes montan las exposiciones, diseñan las instituciones y los edificios que luego se construyen, etc. En este sentido, las expertas no solo evalúan desde la perspectiva histórica, sino que la definen.

Esta influencia se refleja en los conocimientos que ponen en práctica, los cuales matizan y dirigen su empresa hasta alcanzar cierto compromiso con el bagaje profesional e histórico,

el cual, inevitablemente, se impregna de los valores socioculturales (Sans Pinillos, 2021: 337-338). En ese sentido, hay una tensión entre el comisariado de exposiciones y el campo de conocimiento que define su papel imprescindible, pues no habría una transferencia unívoca de su agencia en la aprehensión de la *agencia* de los enfoques curatoriales por parte de las visitantes. Por ello, incluso con la mejor de las reorganizaciones de los museos, la distribución axiológica museística se inicia en la misma noción de museo y se actualiza con su realización por parte de las expertas y uso que las visitantes hacen de ellos.

Así, tanto la realidad social, que refleja la sociedad y sus estructuras, como la relación entre la dimensión privada y pública del individuo son significantes epistémicos que inciden directamente en la memoria y el recuerdo a través de complejos procesos cognitivos. Estos procesos involucran la dimensión personal en la experiencia, constituyendo la memoria única de la presentación del agente al mundo en diferentes escenarios y perspectivas a través tanto de su conciencia autobiográfica como de su memoria autobiográfica (Nelson y Fivush, 2020). Cuando la biografía se construye sobre la exclusión, puede generarse desafección hacía los museos, malas expectativas y, en última instancia, un rechazo a visitarlos. Esta interacción, en cambio, es de una naturaleza distinta en el caso de los museos virtuales, y podría decirse que existe una crítica relativamente generalizada (Schweibenz, 2013) en lo que respecta a la interacción que puede darse en este tipo de museos.

Por una parte, si bien es cierto que este tipo de museos, los virtuales, cumplirían parcialmente los requisitos de la definición presentada en este artículo, creemos menester mencionar que, en un principio, se destacó la ausencia de cualquier tipo de interacción en los primeros museos virtuales, propios del siglo pasado (c.f., Huhtamo, 2010). El espacio virtual de estos museos originales es, así, un tipo especial de mediador moral inerte pues, más que distribuir, su función principal podría describirse en términos de *perpetuación* del marco axiológico vigente en el museo tradicional. Aunque la distinción entre visitante y experta sigue aplicándose, la ausencia de interacción podría contribuir a exacerbar este marco axiológico, precisamente por falta de experiencias que las confrontan. Así pues, la visión contemplativa (y tradicionalmente dominante) se vio reforzada con este tipo de espacios: en definitiva, el papel de la usuaria se reducía fundamentalmente a la contemplación de las obras digitalizadas, ya que la interacción en entornos virtuales todavía no se había desarrollado[12]. Retomando la teoría de la radiación-irradiación con la que se ha abordado la exclusión que pueden causar los museos tradicionales, podría decirse que, en los nuevos espacios virtuales, esta situación se confirma en términos de la ratificación de la autobiografía. En este contexto, la relación que se establece con el museo digital responde a modelos institucionalizados de interacción con el arte que, tradicionalmente, excluyen grupos de personas con escasa formación y minorías sociales o etnias invisibilizadas. Por lo tanto, dicha confirmación autobiográfica surge precisamente por la evidencia de la exclusión, en el caso de los colectivos invisibilizados, o por la ratificación y perpetuación de un criterio de arte específico, para la mayoría de las usuarias.

En la actualidad, si bien los museos virtuales permiten un tipo de interacción más complejo, podría entenderse que el uso de las bases de datos de las que se sirven los diseños

12 Al margen de esta dimensión contemplativa, en los museos virtuales primitivos podía darse un tipo de interacción reducida y, presumiblemente, marginal, a saber, la que se establecía mediante correo postal o llamada telefónica entre el servicio técnico y las usuarias.

(algunas de estas son, por ejemplo, *Art500k*[13], *WikiArt*[14] o la *Metropolitan Museum of Art Collections Database*[15]) refuerza esta visión contemplativa. Ahora bien, el museo virtual, tal y como se defiende en este trabajo, difiere de una mera base de datos que recopila obras artísticas digitales o digitalizadas, pues mientras que las bases de datos recopilan y, a lo sumo, relacionan datos concernientes a las obras de arte, el museo virtual habrá de cumplir con un número significativo de las funciones mencionadas en la definición (esto es, investigación, colección, conservación, interpretación y exhibición). De igual modo, podría argumentarse que, en los museos virtuales presentes, la contemplación se posibilita a partir de una navegación por el entorno digital (por ejemplo, en el entorno virtual de una aplicación para móviles) que, a su vez, implica una cierta participación de la usuaria que no es equiparable al acto mecánico de recorrer el espacio físico de un museo tradicional.

Por otra parte, cabe señalar que la posible participación de las usuarias en la mayoría de los museos virtuales no se limita a la contemplación. Ciertamente, las personas pueden decidir, al menos parcialmente, el orden y el detalle (por ejemplo, ampliando el zoom) con el que visualizar las obras artísticas, y en algunos de estos espacios podría darse también la opción de interactuar con el asistente virtual del museo. La interacción entre las visitantes del museo, empero, no suele integrarse en los museos virtuales, con lo que estos presentan una carencia respecto a los museos tradicionales. Mencionando algunos trabajos de Heath et al., Schweibenz escribe al respecto:

> «The major problem of information technology both in the museum and online is the scarcity of interaction between visitors. While physical visits allow interaction between the visitors any time, interactive computer exhibits in museums most often allow only an exclusive interaction between one visitor and technical device he or she uses instead of interaction between several visitors (Heath, Hindmarsh & Lehn 2002: 20f; Heath & Lehn 2003: 10).» (Schweibenz, 2013: 47).

Podría argüirse que el desarrollo tecnológico de estos espacios virtuales puede solventar en parte esta cuestión, en la medida en que se trataría de integrar sistemas para interactuar y relacionarse entre las visitantes en los museos virtuales. Sin embargo, esta interacción digital no sería equivalente a la que puede darse en un museo tradicional[16]. En cualquier caso, el diseño de un museo virtual en el que se considere la dimensión ética deberá incorporar este tipo de sistemas de comunicación entre las visitantes. Tomando la idea expuesta anteriormente de que las expertas responsables de la distribución axiológica museística tradicional están influenciadas por la noción social de museo, también las programadoras del museo virtual se ven influenciadas del mismo modo, aunque en este caso el efecto se plasma en el diseño de los museos virtuales.

A lo largo de la sección se ha reflexionado acerca de la relación e interacción que ciertos grupos sociales establecen con el museo virtual e, implícitamente, se ha asumido un acceso universal a las tecnologías del museo virtual. Esta presunción tiene una motivación teórica,

13 https://deepart.hkust.edu.hk/ART500K/art500k.html.

14 https://www.wikiart.org/.

15 https://www.metmuseum.org/art/collection/search?searchField=All&showOnly=openAccess&sortBy=relevance&pageSize=0.

16 Esta comparativa presupone la interacción entre personas sin diversidad funcional o capaces de comunicarse en el espacio de un museo tradicional.

pues es evidente que no se corresponde con el estado actual de los hechos. Por ello, el diseño de los museos virtuales debería tener en cuenta la brecha digital. La brecha digital hace referencia a la desigualdad social y económica generada por las variaciones en el uso de las tecnologías de la información y la comunicación (Gómez, 2019), e involucra preocupaciones tanto institucionales como individuales (Berdé, 2019). De hecho, la función integral y vital de la tecnología identifica la brecha digital como un factor crucial para determinar las desigualdades socioeconómicas (Hooft, 2018). Actualmente, la conceptualización de la brecha digital presenta tres categorías principales (Lythreatis et al., 2021), a saber, la brecha de acceso (entendida como la imposibilidad que tienen las personas a la hora de costear los dispositivos necesarios para acceder al recurso en cuestión), la de uso (relativa a la bisoñez en habilidades digitales que restringen el uso de la tecnología, la cual da lugar a un desaprovechamiento de los beneficios de la misma) y la de calidad (que concierne a la calidad de uso y, además, establece una distinción entre las personas que utilizan la tecnología en un nivel básico y aquellas que son capaces de apropiarse y hacer un uso experto y transformador de la misma).

En lo que respecta al museo virtual, cabe destacar que, aunque, en general, este es económicamente menos costoso que uno tradicional, hoy día el acceso a este tipo de espacio virtual no está garantizado. Asimismo, en relación a la brecha de uso, el diseño de un museo virtual debería integrar y dar cabida a todo tipo de usuarias, además de proveer a estas de las instrucciones necesarias para recorrer el espacio virtual. Finalmente, y en coherencia con la tesis defendida en esta sección, se debería promover un uso del museo virtual de calidad donde las usuarias puedan reconfigurar el espacio virtual y personalizar su experiencia en el mismo.

5. Conclusiones

En este artículo hemos revisado las principales conceptualizaciones del museo virtual en la literatura especializada para, a partir de las mismas, proponer en la sección 2 una definición básica que las integre. A continuación, se han presentado argumentos para establecer la insuficiencia de la perspectiva dataísta en el estudio de la injusticia epistémica relativa al museo virtual. Finalmente, hemos estudiado la exclusión relativa a la participación y la interacción entre las visitantes de un museo (tanto virtual como físico) y la propia institución.

A partir de lo expuesto en este trabajo, concluimos en primer lugar que el museo virtual, si bien abre nuevas oportunidades para la eliminación de la discriminación que se da en los museos tradicionales, necesita diseñarse al margen de los postulados dataístas. Así pues, se ha de combatir esta exclusión sin el criterio dataísta de integración de toda obra de arte en un catálogo universal y con una orientación que, en propio el diseño y en el uso de los museos digitales, tenga en cuenta las problemáticas expuestas en este artículo. En particular, propondríamos como punto de partida una estrategia basada en las siguientes tres vías conectadas entre sí:

1. visibilizar las obras representativas de los colectivos minoritarios resaltando su presencia;
2. redefinir las etiquetas con el fin de que las obras se expongan con el resto de las tradicionales;
3. dar voz a las usuarias para saber qué esperan de un museo, qué les gustaría visitar y de qué forma.

De este modo, los colectivos minoritarios se harían visibles y participarían en los diferentes procesos asociados a un museo (en este caso, virtual). Así pues, dada la relación constitutiva entre la cultura y la ciencia y tecnología que posibilitan los museos virtuales, esta estrategia para luchar contra la exclusión debe entenderse una empresa de constante refinamiento. El reto de eliminar un problema de exclusión es, por lo tanto, integrar esta estrategia en un proceso cambiante, y, en tal sentido, la estructura del museo virtual no puede ser inerte. En definitiva, son la participación en el diseño del museo virtual así como su posterior revisión los dos ejes desde los que habilitar un mecanismo social paralelo a los límites dataístas donde la perpetuación de la exclusión podrá combatirse de forma activa y constante.

Referencias bibliográficas

Bandelli, A. (1999). Virtual Spaces and Museums. *Journal of Museum Education*, 24 (12), 20-22.

Banfi, F., Pontisso, M., Paolillo, F.R., Roascio, S., Spallino, C., & Stanga, C. (2023). Interactive and Immersive Digital Representation for Virtual Museum: VR and AR for Semantic Enrichment of Museo Nazionale Romano, Antiquarium di Lucrezia Romana and Antiquarium di Villa Dei Quintili. *ISPRS International Journal of Geo-Information*, 12(2), 28. https://doi.org/10.3390/ijgi12020028

Bell, D. R. (2017). Aesthetic encounters and learning in the museum. *Educational Philosophy and Theory*, 49(8), 776-787. https://doi.org/10.1080/00131857.2016.1214899

Brewer, J. D. (2004). Imagining the sociological imagination: The biographical context of a sociological classic. *The British Journal of Sociology*, 55, 317-333. https://doi.org/10.1111/j.1468-4446.2004.00022.x

Berde, É. (2019). "Digital Divide and Robotics Divide" en Gu, D., Dupre, M., (eds.)., *Encyclopedia of Gerontology and Population Aging*, Switzerland: Springer, Cham https://doi.org/10.1111/j.1468-4446.2004.00022.x

Casacuberta, D., Guersenzvaig, A. (2019). Using Dreyfus' legacy to understand justice in algorithm-based processes. *AI & Soc*, 34, 313-319. https://doi.org/10.1007/s00146-018-0803-2

Chalmers, M., Galani. M. (2008). "Blurring Boundaries for Museum Visitors" en Marty, P.F., Burton Jones, K. (eds.)., *Museum Informatics. People, Information, and Technology in Museums*, New York: Routledge, 157-177.

Costanza-Chock, S. (2020). *Design justice: community-led practices to build the worlds we need*. Cambridge: The MIT Press.

Dattathrani, S. and De', R. (2023). The Concept of Agency in the Era of Artificial Intelligence: Dimensions and Degrees. *Information Systems Frontiers*, 25:29–54. https://doi.org/10.1007/s10796-022-10336-8

Daviddi, S., Mastroberardino, S., Jacques, P.L.St., Schacter, D.L., & Santangelo, V. (2022). Remembering a Virtual Museum Tour: Viewing Time, Memory Reactivation, and Memory Distortion. *Frontiers in Psychology*, 13, 869336. https://doi.org/10.3389/fpsyg.2022.869336

Djindjian, F. (2009). The virtual museum: an introduction. *Archeologia e Calcolatori*, 9-14.

Epstein, Z., Levine, S., Rand, D. G., & Rahwan, I. (2020). Who gets credit for AI-generated art? *iScience*, 23(9), 1-10. https://doi.org/10.1016/j.isci.2020.101785

Ferrero, L. (2022). The Routledge Handbook of Philosophy of Agency. Abingdon, Oxon; New York, NY: Routledge. https://doi.org/10.4324/9780429202131

Feyerabend, P. (1978). *Science in a free society*, London: Lowe & Brydone Ltd.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford: Oxford University Press.

Gómez, D.C. (2019). Technological capital and digital divide among young people: An intersectional approach. *Journal of Youth Studies*, 22, 941-958. https://doi.org/10.1080/13676261.2018.1559283

Boris, G. (2014) "On Art Activism" *e-Flux Journal* 56 (June 2014). Accessed May 25, 2023. https://www.e-flux.com/journal/56/60343/on-art-activism/.

Hazan, S., Hermon, S., Turra, R., Pedrazzi, G., Franchi, M., & Wallergard, M. (2014). What is a Virtual Museum? *V-MUST.net - D 3.1b*, Recuperado de http://www.v-must.net/library/documents.

Heath Christian, Hindmarsh Jon, Lehn Dirk vom (2002). Video-based Field Studies in Museums and Galleries. *Visitor Studies Today!*, 5(3), 15-17, 20-23.

Heath Christian, Lehn Dirk vom (2003). "Displacing the Object. Mobile Technologies and Interpretative Resources" en *International Conference on Hypermedia and Interactivity in Museums (ICHIM)* en Paris, Septiembre 8-12, Paris: ICHIM, 1-15.

Hertzum, M. (1998). A Review of Museum Web Sites. *Search of User-Centred Design. Archives and Museum Informatics*, 12(1), 127-138. https://doi.org/10.1023/A:1009009104685

Hooft Graafland, J. (2018). New Technologies and 21st Century CHILDREN: Recent Trends and Outcomes. *OECD Education Working Papers*, 179.

Huhtamo, E. (2002). "On the origins of the virtual museum" en Ross P. (ed.)., *Museums in a digital age*, London: Routledge, 121-135.

ICOM (2022). Acta de la Asamblea General Extraordinaria del ICOM, celebrada en Praga el 24 de agosto de 2022. Recuperado de https://icom.museum/es/recursos/normas-y-directrices/definicion-del-museo/.

Kaplan, S., Bardwell, L. V., & Slakter, D. B. (1993). The museum as a restorative environment. *Environment and Behavior*, 25(6), 725-742. https://doi.org/10.1177/0013916593256004

Latham, K.F., Simmons, J.E. (2014). *Foundations of Museum Studies. Evolving Systems of Knowledge*, Santa Barbara, CA: Libraries Unlimited.

van Lier, M. (2023). Introducing a four-fold way to conceptualize artificial agency. *Synthese* 201, 85. https://doi.org/10.1007/s11229-023-04083-9

Luther, W., Baloian, N., Biella, D., & Sacher, D. (2023). Digital twins and enabling technologies in museums and cultural heritage: An overview. *Sensors*, 23(3), 1583. https://doi.org/10.3390/s23031583

Lythreatis, S.; Singh, S.K.; El-Kassar, A.-N. (2021). The digital divide: A review and future research agenda. *Technol. Forecast. Social Change*, 175, 121359. https://doi.org/10.1016/j.techfore.2021.121359

Magnani, L., Bardone, E. (2007). Distributed Morality. Externalizing Ethical Knowledge in Technological Artifacts. *Foundations of Science*, 13(1), 99–108. https://doi.org/10.1007/s10699-007-9116-5

Magnani, L. (2018). The urgent need of a naturalized logic. *Philosophies*, 3(44). https://doi.org/10.3390/philosophies3040044

Mintz, A. (1998). "Media and Museums: A Museum Perspective" en Thomas, S. and A. Mintz (eds.)., *The Virtual and the Real. Media in the Museum,* Washington, DC: American Association of Museums, 19-34.

Negri, M. (2012). The Virtual Museum, a shift in meaning en Nicholls, A., M. Pereira, and M. Sani (eds.)., *The Virtual Museum. The Learning Museum Network Project*. Report 1, 12-19. http://online.ibc.regione.emiliaromagna.it/I/libri/pdf/LEM_report1_theVirtualMuseum.pdf.

Nelson, K., & Fivush, R. (2020). The Development of Autobiographical Memory, Autobiographical Narratives, and Autobiographical Consciousness. *Psychological Reports*, 123(1), 71-96. https://doi.org/10.1177/0033294119852574

Niu, P. (2022). An artificial intelligence method for comprehensive evaluation of preschool education quality. *Frontiers in Psychology*, 13, 1-12. https://doi.org/10.3389/fpsyg.2022.843865

Opotow, S. (2011). "Social Injustice" en Christie, D. J. (sd.)., *The Encyclopedia of Peace Psychology*. https://doi.org/10.1002/9780470672532.wbepp256

Ramirez, E. J., Jennett, S., Tan, J., Campbell, S., & Gupta, R. (2023). XR Embodiment and the Changing Nature of Sexual Harassment. *Societies*, 13(2), 36. https://doi.org/10.3390/soc13020036

Prior, N. (2006). "Postmodern restructurings" en Macdonald, S. (ed.)., *A Companion to Museum Studies*, Hoboken: John Wiley & Sons, Ltd, 509-534.

Rectanus, M. W. (2006). "Globalization: Incorporating the museum" en Macdonald, S. (ed.)., *A companion to museum studies*, Hoboken: John Wiley & Sons, Ltd, 381-397.

Rodríguez de las Heras, A. (2004). Espacio digital. Espacio virtual. *Debats*, 84, 63-67.

Sacher, D. A. (2017). *Generative Approach to Virtual Museums Using a New Metadata Format: A Curators', Visitors' and Software Engineers' Perspective*, Berlin: Logos.

Sans Pinillos, A. (2021). "Distribución abductiva de los valores culturales: el proyecto de Alejandría" en Estany, A. y Gensollen, M. (eds.)., *Diseño institucional e innovaciones democráticas*, Aguascalientes: Universidad Autónoma de Barcelona-Universidad Autónoma de Aguascalientes, 333-352.

Sans Pinillos, A. (2023). "Abductive Irradiation of Cultural Values in Shared Spaces: The Case of Social Education Through Public Libraries" en Magnani, L. (eds)., *Handbook of Abductive Cognition*, Switzerland: Springer, Cham, 1147–1171. https://doi.org/10.1007/978-3-030-68436-5_51-1

Schiff, D. (2021). Out of the laboratory and into the classroom: the future of artificial intelligence in education. *AI & Society*, 36, 331-348. https://doi.org/10.1007/s00146-020-01109-7

Schweibenz, W. (2019). The virtual museum: An overview of its origins, concepts, and terminology. *The Museum Review*, 4(1), 1-29. https://doi.org/10.1515/tmr-2019-0001

Schweibenz, W. (2013). Museum Exhibitions - The Real and the Virtual Ones: An Account of a Complex Relation. *Uncommon Culture*, 3(5/6), 38-52. https://journals.uic.edu/ojs/index.php/UC/article/view/4715

Wagner, A. (2022). Superando las «dos culturas». Retos filosóficos más allá de la dicotomía entre ciencia y cultura. Pensamiento. *Revista De Investigación E Información Filosófica*, 78(298 S. Esp), 573-593. https://doi.org/10.14422/pen.v78.i298.y2022.017

SECCIÓN 2

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 97-98
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.577651

Ética aplicada para una Inteligencia Artificial confiable

Applied Ethics for Trutstworthy Artificial Intelligence

ELSA GONZÁLEZ- ESTEBAN[*]
DOMINGO GARCÍA-MARZÁ[**]

La Inteligencia Artificial –en adelante, IA– se ha convertido en el centro de atención de teóricos y profesionales de diferentes ámbitos de la actividad humana. El interés no sólo recae en profesionales informáticos. De hecho, su abordaje y desarrollo se encuentra en el centro de las políticas públicas y económicas de todos los países. Esto se debe a la supuesta capacidad que la IA muestra para analizar, predecir, mejorar y dar respuesta a los múltiples problemas y conflictos que subyacen a cada ámbito de actividad y, cuya no utilización parece que podría limitar su progreso.

En el ámbito político y económico, la aplicación de algoritmos de IA en los diferentes procesos de las instituciones, organizaciones y empresas está permitiendo mejorar notablemente en la optimización de los recursos, la erradicación de la corrupción, el nepotismo y la desafección, la captación de talento, el acceso a ingentes cantidades de información, la asignación ágil de financiación, la precisión y control administrativo, y el diagnóstico de enfermedades, entre otras. Sin embargo, estos modelos no están exentos de fuertes críticas y grandes dudas tanto internas como externas. Las cada vez mayores evidencias de su falta de objetividad, su alta opacidad, su relativismo conductual, su arbitraje sesgado, etc., muestran la exigencia de que sean los propios afectados los que definan las expectativas de la IA a corto y medio plazo para evitar las consecuencias negativas que está produciendo o puede llegar a producir. En definitiva, muestran la necesidad de la acción y supervisión humanas, así como de una reflexión ética profunda de la misma capaz de orientarla.

En este contexto esta sección monográfica: «ética aplicada para una Inteligencia artificial confiable» ha querido servir de espacio para la reflexión ética, con un énfasis en las propuestas ético-discursivas, sobre la nueva situación creada por la creciente penetración y colonización de la IA en todos los órdenes de la vida humana.

En concreto las contribuciones presentes en la sección aportan relevantes propuestas y recomendaciones en tres líneas de investigación desde una perspectiva ético-crítica: (i) Los presupuestos conceptuales en materia de ética y principios éticos utilizados en el contexto de

* Profesora Titular de Filosofía Moral, Universitat Jaume I – esteban@uji.es
** Catedrático de Filosofía Moral, Universitat Jaume I – garmar@uji.es

la Inteligencia Artificial. (ii) Las directrices y normativas internacionales y nacionales que se están proponiendo para regularla desde sistemas de regulación jurídica o de autorregulación. (iii) La profundización desde la ética aplicada sobre los desafíos éticos a los que se están enfrentando las diferentes esferas de la vida humana y sus profesiones: los diseños e investigaciones en IA, la deliberación pública, el periodismo, el derecho y la actividad académica.

En definitiva, esta sección presenta aportaciones que permiten concretar qué significa y cómo se puede garantizar la acción y supervisión humanas en el diseño, uso y aplicación de la inteligencia artificial en diferentes ámbitos donde la utilización de la IA requiere del conocimiento ético experto. Durante tres años, 2020-2023, más de 20 investigadoras e investigadores han abordado estas líneas de trabajo en un seminario bimensual con equipos multidisciplinares y los textos que aquí se presentan se han nutrido de las actividades de investigación del proyecto coordinado de investigación del Plan Estatal I+D+I "Ética discursiva y democracia ante los retos de la Inteligencia Artificial" (EDDERIA) [PID2019-109078RB-C21 y PID2019-109078RB-C22 financiado por MCIN/ AEI /10.13039/501100011033] de la Universitat Jaume I y la Universitat de València y de Proyecto PROMETEO *y en las actividades del grupo de investigación de excelencia PROMETEO CIPROM/2021/072, financiado por Conselleria d'Innovació, Universitats, Ciència i Societat Digital de la Generalitat Valenciana*.

La generación de espacios de reflexión crítica y de avance del conocimiento en ámbitos muldisciplinares que requieren de un diálogo interdisciplinar, como es el propio de las éticas aplicadas, es vital, por lo que queremos agradecer a la Revista Daimon y a su Consejo de Edición, en la persona de Emilio Martínez Navarro la generación de tales espacios así como el apoyo y ayuda en que este trabajo pueda ver la luz. Y que lo haga en una publicación en abierto de calidad donde la revisión por pares y el cuidado de la calidad de los textos es la prioridad. Mención especial requieren en este momento los más de veinte autores y autoras que enviaron sus trabajos para revisión, así como los revisores y las revisoras especialistas en las diferentes temáticas abordadas que han participado en este proceso.

Gracias a este trabajo colectivo, la sección aporta un camino para avanzar en la definición de la confiabilidad en la Inteligencia Artificial, desde la perspectiva de los propios afectados, una concepción ético-discursiva. Los diferentes estudios perfilan y delimitan con rigor y precisión qué se puede entender por una IA confiable en el ámbito de la deliberación político-pública, el periodismo, el derecho, la actividad académica. Además, realizan abordajes críticos capaces de ofrecer luz acerca de cómo se puede definir en tales terrenos la confiabilidad desde las expectativas de los afectos por la IA a corto y medio plazo para evitar las consecuencias negativas que está produciendo o puede llegar a producir el diseño, producción y uso de la IA.

Castellón de la Plana, 13 de Julio de 2023

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 99-114
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.562821

Ética digital discursiva: de la explicabilidad a la participación*

Discursive digital ethics: From explicability to participation

*DOMINGO GARCÍA-MARZÁ***

Resumen. el presente artículo tiene como objetivo presentar los rasgos básicos de una ética digital dialógica a partir de una lectura crítica del documento elaborado por el grupo independiente de expertos de alto nivel para la Comisión Europea *Ethics Guidelines for Trustworthy AI* (High-level expert Group on Artificial Intelligence, 2019). Una ética digital que tiene en el diálogo y acuerdo posible de todos los agentes implicados y afectados por la realidad digital su horizonte normativo de actuación, su criterio de justicia. La finalidad es mostrar que, en su esfuerzo por generar una voluntad común y una gobernanza europeas ante la actual revolución industrial, la participación de todas las partes implicadas no solo es recomendable sino moralmente exigible. El reconocimiento de la igual dignidad que implica una Inteligencia Artificial centrada en las personas no es ni siquiera pensable sin el horizonte de una participación igual. Sin ella, la confianza no puede generarse ni garantizarse. Como pretendemos mostrar, en este objetivo juega un papel decisivo el nuevo principio de explicabilidad, principio que posee un valor moral y no solo instrumental.

Abstract. This article is intended to present a proposal for dialogic digital ethics on a critical reading of the European Commission's independent high-level expert group's document *Ethics Guidelines for Trustworthy AI* (2019). These would be digital ethics with a normative horizon for action and criteria for justice based on dialogue and possible agreement between all agents involved and affected by the digital reality. The aim is to show that the participation of all parties involved is not merely advisable but morally required as part of the Commission's effort to generate common European willingness and governance to deal with the fourth industrial revolution now going on. The acknowledgement of equal dignity implied by people-centric artificial intelligence (AI) is utterly unthinkable without this possibility of equal participation. Without it, trust cannot be generated or guaranteed. As we intend to show, the new principle of explicability plays a decisive role in this objective as a principle with a moral as well as an instrumental.

Recibido: 28/03/2023. Aceptado: 04/06/2023.

* Este trabajo se enmarca dentro de los objetivos del Proyecto de Investigación Científica y Desarrollo Tecnológico «Ética aplicada y confiabilidad para una Inteligencia Artificial» [PID2019- 109078RB-C21] financiado por el Ministerio de Ciencia e Innovación, así como en las actividades del grupo de investigación de excelencia CIPROM/2021/072 de la Comunitat Valenciana.

** Catedrático de Ética y Filosofía Política en la Universitat Jaume I de Castellón. Es autor, entre otros, de los siguientes libros: *Public reason and Applieds Ethics* (junto con A. Cortina y J. Conill (eds.) Londres, 2008); *Ética y Filosofia Política. Homenaje a Adela Cortina* (junto con J. Félix Lozano; E. Martinez y J. C. Siurana, Madrid, 2018). Autor de numerosos artículos sobre la relación entre ética, política y economía y su aplicación en el diseño institucional. Los resultados de estas investigaciones se han plasmado en diversas instituciones públicas y privadas. Es co-director del Programa interuniversitario de Doctorado «Ética y Democracia» en la Universitat Jaume I y patrono de la Fundación ETNOR.

Con este fin este trabajo se estructura en tres partes. En primer lugar, se argumentará la propuesta de una ética digital dialógica encargada, como ética aplicada, de explicitar las bases éticas que subyacen a la confianza en la Inteligencia artificial, en sus decisiones, prácticas e instituciones. Desde este marco ético, en segundo lugar, se analizarán las Directrices Europeas y se propondrá revisar la consideración de la inclusión y participación de todas las partes implicadas no solo como un requisito recomendable sino como una exigencia moral, destacando la necesidad de justificar y potenciar una participación real y efectiva. Una justificación que se realizará, ya en el tercer punto, desde el principio de explicabilidad como principio moral, siguiente el camino del principio kantiano de publicidad. La finalidad es avanzar, desde esta propuesta de una ética digital discursiva, un diseño institucional capaz de responder de esta exigencia moral de la participación libre e igual de todos los afectados e implicados. El *principio de explicabilidad* se convierte así en un principio básico para garantizar este saber moral para la toma de decisiones y la creación de espacios de confianza "dentro" de las instituciones que conforman el sistema socio-técnico de la Inteligencia artificial.
Palabras clave: Ética, Ética aplicada, Ética digital, Inteligencia Artificial, recursos morales, participación, diseño institucional, infraestructura ética.

To this purpose, this study is structured in three parts. First, it will argue the proposal of a dialogic digital ethics in charge, as applied ethics, of making explicit the ethical bases that underlie the trust in Artificial Intelligence, in its decisions, practices and institutions. From this ethical framework, secondly, the European Guidelines will be analyzed, highlighting the need to justify and enhance the participation of all parties involved. This justification will be based on the Kantian principle of publicity. Finally, from this proposal of a discursive digital ethics, an ethical infrastructure capable of integrating all institutional design and all algorithmic development with this moral requirement of free and equal participation of all those affected and involved will be proposed. The principle of explicability thus becomes a basic principle to guarantee this moral knowledge for decision-making and the creation of spaces of trust "within" the institutions that make up the socio-technical system of Artificial Intelligence.
Keywords: Ethics, applied ethics, Digital ethics, artificial intelligence, moral resources, institutional design, participation, ethical infrastructure.

1. Hacia una ética digital dialógica como ética aplicada

Estas directrices éticas, aunque llegan con retraso dada la aceleración tecnológica actual, siempre son bienvenidas en su función de orientar la toma de decisiones individuales e institucionales y para influir en el desarrollo legislativo. Mientras que nuestros gobernantes insisten y repiten por doquier que todo avance tecnológico debe ir acompañado de una visión ética, de una base humanista, de un proyecto de progreso, de un desarrollo innovador e inclusivo, etc., las estrategias, las políticas públicas, las grandes tecnológicas, los contratos público-privados, etc., la realidad, en suma, lo desmiente. Un caso práctico nos puede servir de ejemplo. Lo encontramos en la *Estrategia de Inteligencia Artificial de la Comunidad Valenciana*, donde podemos leer: "los beneficios de aplicar la IA se reflejarán en una mejor toma de decisiones al tener en cuenta todas las posibles variables y contrastar un mayor número de datos. Al estar basadas en algoritmos que no se ven afectados por subjetividades personales, las decisiones son más objetivas. También son más rápidas porque la capacidad de cómputo supera a la capacidad de la inteligencia humana (Generalitat Valenciana, 2018: 5).

Si no se denuncian estos sesgos de superioridad, neutralidad y objetivismo, que parecen acompañar a la comprensión de la actual transformación digital, si no se desvelan estos prejuicios, una ética digital —el análisis de lo correcto o incorrecto en el ámbito tecnológico y digital— solo podrá aplicarse después de aparecido el problema, tras las consecuencias de las decisiones y acciones ya tomadas desde esta racionalidad tecnológica reducida a los datos y macrodatos, a los algoritmos y a la inteligencia artificial. Y ya será, como la experiencia nos muestra en el día a día, demasiado tarde (Apel, 1988; Rehg, 2015; Yuste et al., 2017).

Esta negativa a reconocer cualquier intervención humana, como si los datos estuvieran "ahí fuera" esperando ser descubiertos, como si los algoritmos no fueran de facto fruto de nuestra interpretación de la realidad, de nuestros intereses de conocimiento, no solo demuestra una clara torpeza para comprender una situación, sino también una clara intencionalidad para *cosificar*, y, por tanto, *obstruir* toda posibilidad de diseño y gobernanza éticos (García-Marzá, 2022). Esta denuncia es la primera tarea de una ética digital dialógica, pero no la única como a continuación veremos.

Una ética digital, centrada en las bases morales que subyacen a la confianza que depositamos en las diferentes prácticas digitales y sus respectivas tecnologías, debe comenzar por criticar estos prejuicios que, una vez más, reaparecen vinculados al dominio tecnológico que la ciencia presupone, ya sea como cientificismo, objetivismo, positivismo, etc. y que hoy son recuperados por las neurociencias y la Inteligencia Artificial (García-Marzá, 2019). El punto de partida para una ética digital no puede ser otro que la superación de este obstáculo, de este dogmatismo epistemológico que impide abordar la parte humana que define y gestiona todo tipo de tecnologías. Si la decisión algorítmica es más justa que la decisión humana, tanto la autonomía moral como la política desaparecen. Debemos mostrar de nuevo, insistir sin cansarnos, que la definición, la identificación y selección, de un dato depende siempre de determinados intereses y, por lo tanto, cuando se utilizan para la construcción de los algoritmos, está ya cargados de valores (Mittelstad et. al.2016; Floridi, 2019).

De ahí que la Comisión Europea, en su informe *Generar confianza en la IA centrada en el ser humano* (COM.2019), afirme que "La IA trae retos, ya que permite a las máquinas "aprender" y tomar decisiones y ejecutarlas sin intervención humana. Ahora bien, las decisiones adoptadas por algoritmos pueden dar datos incompletos y, por tanto, no fiables, que puede ser manipulados, sesgados o simplemente estar equivocados" (COM.2019: 2). Como el resto de tecnologías, las digitales son un instrumento, un medio, para conseguir determinados fines, no un fin en sí mismo. Como veremos, el significado de un dato, por qué un dato lo es y otro no, siempre es una decisión humana. Los algoritmos dependen de una realidad construida desde un interés determinado. Desde este interés se construye el algoritmo que reunirá, integrará y dará sentido a los datos. Las redes y el internet de las cosas son la fuente de los macrodatos, los algoritmos ordenan estos datos (García-Marzá; Calvo, 2022).

El dictamen 4/2015 del European Data Protection Supervisor titulado *Hacia una nueva Ética Digital. Datos, Dignidad y Tecnología* (2015), apuesta por estimular un debate abierto y documentado sobre la definición, justificación y aplicación de una *nueva ética digital* (2015: 5). Un debate en el que participen la sociedad civil, los diseñadores, las empresas, los académicos, las autoridades reguladoras, etc. Una ética que permita "mejorar los beneficios de la tecnología para la sociedad y la economía por vías que refuercen los derechos y las libertades de las personas físicas" (European Data Protection Supervisor, 2015: 5).

Perfecto, podríamos pensar, pero si no se concreta esta participación, si no se establecen los mecanismos y procedimientos institucionales que la posibiliten, solo son vanas palabras. La duda es siempre la misma: ¿no estamos ante un nuevo intento para encubrir las injusticias provocadas por las nuevas tecnologías digitales y las grandes empresas tecnológicas, dueñas hoy por hoy de la globalización, con la piel de cordero de la dignidad humana? De hecho, este peligro de caer en un *ethics washing* está claramente explícito en el reciente informe de la UNESCO (2022).

Como ética aplicada, una ética digital dialógica tiene como objetivo explicitar y gestionar las bases éticas de la confianza depositada en la llamada revolución digital. Se concibe como una ética aplicada porque su objetivo no se detiene en la explicitación del saber moral que utilizamos en todo proceso de digitalización, desde la determinación de los datos y su integración a través de los algoritmos, hasta el aprendizaje autónomo y la Inteligencia Artificial, pasando por la hiperconectividad proporcionada por el internet de las cosas, las máquinas y robots, así como los problemas derivados de la computación en nube. El bien primario que aporta la práctica digital son precisamente los datos como significados atribuidos a los signos, como interpretaciones de lo dado, de la realidad. Los datos no recogen la realidad, lo dado, sino aquello que nos interesa de la misma y no lo hacen desde la lógica de la causa efecto, sino desde la lógica de la correlación, de las tendencias y patrones (García-Marzá et al., 2004). La neutralidad queda fuera de esta lógica, pues siempre se trata de una "elección" definir un hecho dado como un dato, para después afirmar que este es, por ejemplo, el comportamiento humano.

Podemos hablar también de ética de la inteligencia artificial, de ética algorítmica, de ética de datos, etc., pero en todos estos casos la preocupación y el interés es el mismo: dar razón de la significación y el valor que tiene hoy en día, en contextos globales sin orden jurídico global, la dimensión ética. Una dimensión que se caracteriza, no lo olvidemos, por su pretensión de universalidad. De hecho, hoy en día, la ética digital, con sus valores, principios, directrices, etc., ya interviene en el mundo de la tecnología mucho más que cualquier otra fuerza, pues la percepción y valoración de lo que es moralmente bueno, correcto o justo, influye en la opinión pública, en lo socialmente aceptable o preferible y en lo políticamente factible, y por tanto, en última instancia, en lo legalmente exigible. Por desgracia, este poder de intervención solo suele aparecer con las consecuencias negativas producidas o esperadas, esto es, cuando el mal está hecho.

La ética digital se entiende como una ética aplicada cuyo ámbito de acción es la práctica digital, sus procesos y tecnologías; sus sistemas de toma de decisiones, así como los marcos institucionales en los que se producen —empresas, centros de investigación, etc. Floridi se refiere a la *infoesfera* como conjunto de prácticas conducidas por y dependientes de los datos, ocupándose la ética digital del alcance y las reglas que permiten las interacciones en esta nueva esfera digital. El problema ya no es tanto la innovación digital, como su gobernanza (Floridi,2018; Calvo, 2021).

2. La generación de confianza: una revisión del marco europeo para una IA confiable

Al igual que ocurre con el resto de las éticas aplicadas, la generación y el mantenimiento de la confianza en la esfera digital requiere tres pasos básicos derivados de la imposibilidad

de trasladar automáticamente los principios éticos a la práctica, pues no dejan de ser obligaciones morales abstractas (García-Marzá, 2004). De ahí que necesitemos diferenciar tres niveles en el camino que va de la obligación moral a su realización práctica:

- *Nivel de justificación:* si queremos hablar de validez moral y, por tanto, de normatividad, de lo justo o correcto, debemos fundamentar los principios morales con los que explicitamos nuestro saber moral. La fundamentación debe apelar a razones que justifiquen la universalidad de los principios, esto es, que garanticen la igual dignidad que nos define como personas. Desde este punto de vista, los principios éticos son condiciones de posibilidad de esta igual dignidad.
- *Nivel de realización:* este saber moral debe transformarse en recursos, capacidades y competencias que, dentro de las instituciones, *todo* ser humano tiene a su disposición a la hora de relacionarse con los demás. Estos recursos morales solo aparecen cuando consideramos a los demás como iguales, como interlocutores válidos, cuando actuamos desde el reconocimiento recíproco siguiendo un interés que es de todos, no solo de unos cuantos o de uno solo (García-Marzá, 2004).
- *Nivel de concreción organizativa:* se requiere tanto una cultura como una infraestructura ética para que estos recursos morales puedan ser utilizados en todas las fases que supone la construcción del espacio digital y no solo, como veremos, cuando el problema ya ha aparecido. En este sentido es importante centrarnos en el diseño institucional, tanto de los procesos digitales como de las organizaciones que se encargan de la investigación y de su producción. La confianza en las organizaciones que realizan investigación e innovación, así como las encargadas de su financiación, depende de las razones que posean para sostener su credibilidad y su reputación. Unas razones que exigen la presencia de espacios de participación, de espacios capaces de generar confianza, en el interior de las mismas, en su estructura organizativa. De ahí que hablemos de un diseño institucional para la aplicación y el desarrollo de una ética digital dialógica, de la necesidad de una infraestructura ética (García-Marzá, 2017).

La misma idea de que no es posible pasar directamente de los principios éticos a las prácticas digitales necesitadas de regulación la encontramos en la propuesta *Ethics Guidelines for Trustworthy AI* (High-level expert Group on Artificial Intelligence, 2019) promovida por la Comisión Europea y elaborada por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial. Una propuesta que no deja lugar a dudas sobre el papel que se espera de la ética: "nuestra visión consiste en establecer la ética como pilar fundamental para garantizar y expandir una IA confiable" (par.13). Como ya hemos mencionado, la experiencia nos previene ante el peligro de que estas directrices queden solo en palabras, que no sean capaces de aplicarse, con rigor y efectividad, a una revolución digital que ya lleva tiempo en marcha. Se habla mucho, pero se hace poco o nada. Con lo que el resultado no es la confianza sino al contrario, la suspicacia y la desconfianza. Hablamos de *confianza*, cuando creemos tener razones para esperar algo, cuando pensamos que tal o cual expectativa va a cumplirse. La *fiabilidad*, por su parte, nos lleva a la probabilidad del buen funcionamiento de algo. Ambas remiten a una *confiabilidad* como capacidad de las personas e instituciones de ser dignas de confianza. La voluntad de una parte de depender de otra no es casual, ni se basa en

intuiciones arbitrarias y ajenas a la experiencia. Confiar no es un sentimiento o una creencia irracional, más bien son las razones las que nos disponen a confiar, las que nos dan ánimo, empuje y aliento para actuar, nos motivan a la acción (García-Marzá, 2004). La Estrategia Europea y su Plan Coordinado dejan bien claro que: "la confianza es un requisito previo para garantizar una IA centrada en el ser humano". La Inteligencia Artificial, nos recuerda la Comisión, no es un fin en sí mismo, sino un medio que debe servir para "mejorar" la vida de las personas. Estas son las expectativas que tenemos y en las que la práctica digital apoya su legitimidad y su credibilidad social.

El marco normativo desarrollado en este documento coincide con los niveles anteriormente descritos que deben recorrer toda ética aplicada para pasar de la validez moral a la práctica, de los principios éticos a la realidad. En suma, para la realización del "deber ser". Ninguno de estos tres niveles puede obviarse porque cada uno de ellos implica una normatividad distinta. El documento que venimos analizando recoge estos tres niveles:

1. Los *fundamentos* de una IA fiable se encuentran en la declaración de derechos humanos y en la Carta Europea de Derechos Humanos. Son principios éticos basados en los derechos fundamentales y, en último lugar, en el reconocimiento de la igual dignidad de todas las personas. Son los siguientes: *autonomía* (respeto de la dignidad humana y reconocimiento del otro como interlocutor válido); *no maleficiencia (*prevención del daño y protección de la dignidad*)*; *justicia* (inclusión y distribución justa de costes y beneficios) y *explicabilidad* (transparencia, información y participación en la toma de decisiones). La finalidad de estos principios éticos es inspirar y guiar la lógica del desarrollo, utilización y aplicación de los sistemas de IA y que, por tanto, definen su responsabilidad, aquello de lo que deben dar razón ante la sociedad.
2. La *realización* de una IA fiable requiere siete requisitos básicos que definen las condiciones para su ejecución a lo largo del ciclo de vida de los sistemas de IA: acción y supervisión humanas; solidez técnica y seguridad; gestión de la privacidad y datos; transparencia; diversidad, no discriminación y equidad; bienestar social y ambiental y rendición de cuentas.
3. La *aplicación* concreta requiere a su vez de un análisis de las condiciones y características propias de los contextos específicos, de la *evaluación* de las posibilidades reales de acción.

No es posible entrar en este artículo en los entresijos de esta arquitectónica. Solo afirmar que nos encontramos ante unas directrices donde, por primera vez en las directrices europeas, la ética juega un papel central tanto en la definición como en la gobernanza de una Inteligencia Artificial centrada en el ser humano. La Comisión es consciente de que la confianza buscada depende tanto del cumplimiento legal como de las condiciones de una tecnología robusta. Pero también, y en especial, depende del consentimiento, voluntario e informado, de todos los agentes y procesos que forman parte del contexto socio-técnico, del acuerdo de todos los agentes individuales e institucionales que participan en la generación y en su gestión. Textualmente:

> Para hacer una IA fiable es preciso garantizar la inclusión y la diversidad a lo largo de todo el ciclo de vida de los sistemas de IA. Hay que tener en cuenta a todos los afectados y garantizar su participación en todo el proceso, también es necesario garantizar la igualdad de acceso mediante procesos de diseño inclusivos, sin olvidar la igualdad de trato (par.79).

Esta participación viene exigida desde el principio de autonomía como reconocimiento de la igual dignidad de todas las personas. Una ética digital dialógica tiene precisamente en esta participación libre e igual, en la deliberación y búsqueda de acuerdos, el pilar central para la generación de confianza. Autores como Karl O. Apel y Jürgen Habermas han desarrollado las bases de una ética discursiva, mientras que en la actualidad se desarrollan, en una segunda ola, las éticas aplicadas derivadas de esta exigencia moral de la participación como requisito para el desarrollo de la autonomía y como sostén del valor intrínseco de la dignidad (Cortina; Conill; García.Marzá, 2008). El fundamento moral de la necesidad de este diálogo y posible acuerdo se encuentra en el reconocimiento del valor intrínseco de dignidad de todas las personas implicadas en la realidad digital. Desde este horizonte de actuación, el documento falla al no concretar las posibilidades y procedimientos para esta participación. Como veremos a continuación, encontramos lagunas en el texto que limitan la confianza que pretenden generar.

El valor moral de las directrices que estamos analizando y, con él, su fuerza vinculante, su obligación y su capacidad de convertirse en recursos morales disponibles por todas las partes implicadas, no desaparece en el segundo y tercer nivel. Más bien en estos niveles de adecuación y concreción los principios éticos del primer nivel deben integrarse con las posibilidades de realización, con los límites y potencialidades que cada realidad concreta nos ofrece. La moralidad, y por lo tanto, la exigibilidad de las decisiones, acciones o instituciones, no desaparece con la aplicación de los principios, como bien muestra nuestra capacidad de valorar moralmente los resultados alcanzados y en los que, como veremos en el siguiente apartado, se apoyan las bases éticas de la confianza.

Sin embargo, en el documento europeo no se aprecia bien *esta trazabilidad moral*. En mi opinión, tres cuestiones básicas deberían replantearse para poder explicitar y gestionar estas directrices éticas para una IA confiable.

En primer lugar, es evidente que la fundamentación de estos cuatro principios no puede estar en la recopilación y en la mayor o menor coherencia entre las actuales directrices internacionales, como afirma Floridi (2018: 696). La validez moral, y, por lo tanto, su obligatoriedad, no dependen de un análisis empírico, de una mera comparación entre diferentes propuestas actualmente existentes. Antes bien, las directrices coinciden porque son conceptos que recogen las experiencias históricas de la protección de la dignidad humana, concretadas en los Derechos Humanos y en la Carta Europea. Los principios éticos descritos en el primer nivel pueden considerarse como condiciones de posibilidad de la realización de la dignidad humana en la esfera digital y en sus prácticas y procedimientos.

En segundo lugar, la semejanza con los principios de la bioética no deriva solo del hecho que la bioética es la que más se parece a la ética digital al "tratar ecológicamente con nuevas formas de agentes, pacientes y entornos" (Floridi, 2019). Más bien habría que insistir en la profunda, y muchas veces insondable, asimetría de poder como rasgo básico compartido

en ambos ámbitos. Asimetría que se da entre aquellos que tienen la capacidad de definir y gestionar los datos y quienes van a sufrir las consecuencias de su conversión en algoritmos, en una nueva realidad. Al igual que en su momento ocurrió con la investigación con seres humanos, los cuatro principios bioéticos se adaptan bien a los nuevos retos éticos que plantea la inteligencia artificial, pues también aquí se trata de convertir a los pacientes en agentes, en ciudadanos digitales. Lo que, a mi juicio, no es comprensible es la razón por la que desparece el principio de beneficencia cuando, de hecho, las palabras "mejora"," bienestar", "calidad de vida", etc. brotan por doquier en el documento.

En tercer lugar, a lo largo del documento la participación se reduce a la información, transparencia y, a lo sumo, monitorización. Se trata, generalmente de una relación unidireccional. Cuando se habla de acción humana se entiende que "los usuarios deberían ser capaces de tomar decisiones autónomas con conocimiento de causa en relación con los sistemas de IA. Se les deberían proporcionar los conocimientos y herramientas necesarios para comprender los sistemas de IA e interactuar con ellos de manera satisfactoria y, siempre que resulte posible, permitírseles evaluar por sí mismos o cuestionar el sistema. Los sistemas de IA deberían ayudar a las personas a tomar mejores decisiones y con mayor conocimiento de causa de conformidad con sus objetivos" (par.64).

La tesis principal del presente artículo se centra en argumentar que el principio de explicabilidad puede responder a gran parte de estas preguntas, rellenar estos vacíos, siempre y cuando no pierda su carácter moral, esto es, no se reduzca a ser una mera estrategia ante el pragmatismo de lo realmente posible (Apel, 1988).

La razón de añadir un nuevo principio a los ya utilizados por la bioética no debemos buscarla solo en la complejidad de los macrodatos, del internet de las cosas o de los algoritmos, ni en la desproporción entre quienes construyen los algoritmos y las máquinas de decisión y aprendizaje y los que van a sufrir las consecuencias. Esta asimetría no nos lleva solo a la necesidad de comprender y rendir cuentas de los procesos de toma de decisiones de la IA, como algunos autores creen (Floridi et. al. 2018: 699), sino también a restituir la falta de reciprocidad y de reconocimiento recíproco a través de la *participación libre e igual y, con ella, la posibilidad de influir de forma efectiva, de discutir las posibilidades de acción, de limitar aquellos desarrollos tecnológicos que no tengan claras las consecuencias según el principio bien conocido de precaución*. Una necesidad que no es meramente estratégica, que no deriva solo de la complejidad, opacidad e ininteligibilidad de las prácticas digitales, sino de la moralidad de un principio que nos permite el paso de lo deseable a lo posible, pues es capaz de vincular internamente principios, requisitos y contextos.

En el documento que venimos analizando se presenta el principio de explicabilidad como "crucial" para que los usuarios confíen en los sistemas de IA y para mantener esta confianza. Esto significa:

> (…) que los procesos han de ser transparentes, que es preciso comunicar abiertamente las capacidades y la finalidad de los sistemas de IA y que las decisiones deben poder explicarse —en la medida de lo posible— a las partes que se vean afectadas por ellas de manera directa o indirecta. Sin esta información, no es posible impugnar adecuadamente una decisión (par.53).

En las diferentes directrices internacionales actualmente existentes este principio se expresa en diferentes términos, con mayor o menor identidad semántica: transparencia, responsabilidad, inteligibilidad, comprensibilidad, trazabilidad, comunicabilidad, apertura, claridad, etc. (Jobin, 2019; Mittelstad et. al., 2016). El principio abarca el conocimiento de cómo funciona el algoritmo y también quién es el responsable de su funcionamiento. De esta forma, este nuevo principio se convierte en el hilo con el que se cose la coherencia con el resto de principios, pues se trata que el ciudadano conozca y comprenda la realidad digital en la que está inmerso (Floridi; Cowls, 2019: 12).

Sin embargo, de la definición del principio de explicabilidad que encontramos en las Directrices ni se sigue, ni se exige, la participación de todos los afectados en la que se apoya su valor moral. Parece más bien que estemos ante un mero proceso informativo y, por lo tanto, vertical. Leemos sobre transparencia, sobre explicación y rendición de cuentas de las decisiones tomadas, sobre quién y cómo ha decidido actuar, etc. Características totalmente necesarias, pero totalmente insuficientes, pues no exigen la presencia y el acuerdo de todas las partes implicadas. Podemos pensar que mejor es poco que nada, pero de nuevo lo que está en duda es el valor moral del principio de explicabilidad y, por tanto, su fuerza. Este aspecto sale a la luz solo con seguir leyendo el documento de las directrices:

> No siempre resulta posible explicar por qué un modelo ha generado un resultado o una decisión particular (ni qué combinación de factores contribuyeron a ello). Esos casos, que se denominan algoritmos de «caja negra», requieren especial atención. En tales circunstancias, puede ser necesario adoptar otras medidas relacionadas con la explicabilidad (por ejemplo, la trazabilidad, la auditabilidad y la comunicación transparente sobre las prestaciones del sistema), siempre y cuando el sistema en su conjunto respete los derechos fundamentales (par.53).

Por lo visto, y así se afirma en el texto, el grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado eerróneo o inadecuado. Pero, en mi opinión, tal afirmación no se sostiene desde el punto de vista moral puesto que nada sabemos sobre quién decide las consecuencias posibles en el momento de la recogida de datos o sobre quién define los datos que alimentarán al sistema, por poner dos ejemplos. Como bien sabemos, la autonomía depende de la posibilidad de participar en todo aquello que nos afecta. En esa capacidad se apoya nuestra dignidad y el valor moral de este principio. Valor moral que desaparece si para nada se tiene en cuenta la exigencia de una inclusión libre e igual de todos los posibles afectados. Según el texto anterior, el grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado erróneo o inadecuado. ¿Quién decide si el contexto es adecuado o no? ¿quién la gravedad de las consecuencias? Sin una respuesta, los tres principios éticos anteriores al de explicabilidad caen fácilmente en una especie de "justifica-lo-todo".

En mi opinión, la trazabilidad y la auditabilidad, al igual que la claridad y la inteligibilidad, forman parte efectivamente de todo proceso de transparencia. Pero la transparencia es solo la mitad de la explicabilidad. Hace falta la participación que nos permita deliberar acerca de las preguntas que hemos ido dejando sin respuesta, que nos posibilite, por ejemplo,

alcanzar acuerdos acerca de la conveniencia o no de soportar ciertas desventajas. Sin la aceptación libre y voluntaria, no hay autonomía. Sin autonomía no tenemos ninguna base moral para apoyar el reconocimiento de la igual dignidad de las personas. Este es el fundamento moral del principio de explicabilidad y su potencialidad para el diseño institucional del que a continuación nos ocuparemos.

3. Del principio kantiano de publicidad al principio de explicabilidad. La aportación al diseño institucional

A primera vista puede parecer una ingenuidad exigir la inclusión de *todos* los posibles implicados en las decisiones que les afecten cuando hablamos de miles de millones de personas, de procesos complejos, muchas veces incomprensibles, de consecuencias que nadie esperaba o preveía, etc. Pero tal candidez desaparece cuando recordamos los tres pasos necesarios que van de la exigencia moral a la realidad, del deber ser de los principios éticos a la pretensión de justicia de cada situación concreta, a la realidad de los contextos sociotécnicos, sociales y económicos.

De ahí que debamos mostrar que la validez moral de los principios éticos presentados, la exigencia de universalidad y, por lo tanto, su valor de convicción, su fuerza, sigue estando cuando descendemos al nivel de realización, a los requisitos necesarios para posibilitar y garantizar su aplicación, siempre mediada a través de las condiciones empíricas de realización. Este es, a nuestro juicio, el papel clave que juega el principio de explicabilidad, pues nos permite hablar de una *estrategia moral*, de una mejor o peor aproximación a la idea de la participación y posible acuerdo de todas las partes implicadas (Apel, 1988; García-Marzá, 2019). En mi opinión, una comparación del principio de explicabilidad con el *principio de publicidad* introducido por Kant en la *Paz Perpetua* nos permitirá entender este funcionamiento de la ética digital que, como ética aplicada, debe dar razón del "deber ser" incrustado en la realidad de toda práctica digital.

Con el *principio de publicidad* Kant se propone mediar entre los principios éticos y la práctica, entre lo deseable y lo posible, sin que esta mediación pierda su valor moral (García-Marzá, 2012). Quizás refrescar algunas cuestiones de su propuesta nos sirva para aclarar el valor moral del principio de explicabilidad y la exigencia de participación qu7e conlleva. Cuando Kant nos avisa del peligro que encierran aquellas instituciones donde "todos sin ser todos deciden" nos previene de las consecuencias de anular la diferencia entre la exigencia moral, el segundo *todos* de la frase, y su concreción práctica, el primer todos. Esta distancia entre los principios éticos y su posible realización práctica nunca puede recorrerse del todo, por más justificada que esté la estructura participativa de la empresa, la universidad, el parlamento, el laboratorio, etc. Es precisamente el recorrido, más o menos largo, entre el "todos moral" y el "todos fáctico" donde se asienta toda perspectiva crítica y desde donde se construyen las razones que apoyan la confianza (García-Marzá, 2004).

Esta es la idea básica que desarrollará Kant con el principio de publicidad, cuya primera definición nos habla de la injusticia o inmoralidad de las decisiones tomadas y es bien sencillo: *"Todas las acciones referidas al derecho de otros hombres cuya máxima no es compatible con la publicidad, son injustas"* (ZeF, VIII, 381). Aplicado a nuestro campo, si

los algoritmos deben ser secretos es, de forma clara y tajante, porque son injustos. Podemos buscar eximentes en los derechos de propiedad o en las patentes, pero la validez moral es proporcionalmente inversa a la opacidad. Si nuestra decisión o acción, si los presupuestos, por ejemplo, el origen de los datos, no pueden hacerse públicos es porque son injustos. Desde este principio, todo elemento de los sistemas de Inteligencia Artificial que no pueda ni comunicarse, ni explicarse, es inmoral. Esto significa que deben descartarse. Los algoritmos secretos no pueden ser éticos.

Pero a continuación Kant también nos ofrece una definición positiva del principio de publicidad, aunque ya no tan tajante como la negativa. Dice así: *"Todas las máximas que necesitan de la publicidad (para no fracasar en sus propósitos) concuerdan con el derecho y la política a la vez"* (*ZeF,* VIII, 386).

Este segundo principio ya no exige sólo hacer visibles y conocidas las razones, decisiones, procedimientos, etc., sino que se refiere también a una especie de unión o coincidencia de todos los afectados: la posibilidad de un conocimiento público conlleva, por así decirlo, la necesidad de la aceptación o aprobación de los demás, de otra forma podría fracasar en sus propósitos. De su consentimiento, en definitiva. Aquí público adquiere un segundo significado: ya no se opone sólo a secreto, también parece oponerse a cerrado pues requiere de la aprobación del público para que se cumpla. Necesitar de la publicidad significa que los algoritmos, y demás elementos de los sistemas de IA, no podrían tener éxito, ser eficaces, sin que fuesen públicos. Sencillamente por la desconfianza que generan.

Por supuesto, este acuerdo posible no puede ser anticipado por una o por varias de las partes implicadas, por ejemplo, programadores, empresas, agencias gubernamentales, etc. Si bien no podemos anticipar el acuerdo, sí que existen mecanismos para saber si nos alejamos o nos acercamos a la idea del consenso posible de todos los implicados. Al final de la obra, Kant nos propone un paso más en esta relación entre los principios y su realización, en la explicitación del potencial crítico que encierra el principio de publicidad (García-Marzá, 2012). Como si de una tercera formulación se tratara, Kant equipara la publicidad con la eliminación de toda desconfianza. Textualmente: "Si sólo mediante la publicidad puede lograrse este fin, es decir, mediante la eliminación de toda desconfianza respecto a las máximas, éstas tienen que estar en concordancia con el derecho del público, pues sólo en el derecho es posible la unión de los fines de todos". (ZeF, VIII, 386) ¿No es esta la confianza que buscan las directrices éticas?

La justificación moral y no solo instrumental del principio de explicabilidad deriva, por tanto, del reconocimiento y respeto de la autonomía de todos aquellos afectados o implicados por la regulación o legislación. Kant remite esta justificación a la razón pública, definida como una facultad "donde todos tienen voz". En esta participación de todos los implicados y afectados, personas e instituciones, se basa la justificación moral del principio de explicabilidad. Las tres formulaciones derivan del reconocimiento recíproco de todos aquellos implicados en la regulación institucional, del carácter insustituible de la voluntad libre en la que se asienta la dignidad de las personas. Esta voluntad es la que exige nuestro consentimiento o acuerdo, nuestra libre aceptación. Solo sobre esta posible conformidad es posible generar y garantizar la confianza. Desde esta justificación moral, la aplicación del principio de explicabilidad, entendido como la suma de transparencia y participación *siempre* debe ser posible.

Sin este principio, capaz de mediar entre el *todos moral* y el *todos pragmático,* el resto de principios deja de tener sentido y nos quedamos sin un criterio de lo que es correcto o incorrecto. Un criterio que debe aplicarse, como a continuación veremos, desde el inicio, desde la definición de los datos y el diseño de los algoritmos. En esta tensión inherente al principio de explicabilidad queda integrada la *responsabilidad*, la capacidad de dar razones de lo que hacemos o dejamos de hacer ante los afectados. Más aún, la responsabilidad se convierte siempre en corresponsabilidad. De ahí la generación de confianza.

Veamos cómo se desarrolla este principio de explicabilidad en los requisitos que, según el documento que comentamos, la aplicación de una IA fiable exige. Solo cuando entramos en el terreno de los requisitos o condiciones para la aplicación de los principios nos encontramos con la acción y la supervisión humanas dividida a su vez en tres niveles:

1) *Participación humana*: capacidad de que intervengan los seres humanos en todos los ciclos de decisión del sistema, algo que en muchos casos no es posible ni deseable.
2) *Control humano*: capacidad de que intervengan seres humanos durante el ciclo de diseño del sistema y en el seguimiento de su funcionamiento.
3) *Mando humano*: capacidad de supervisar la actividad global del sistema de IA, incluidos sus efectos económicos, social, jurídicos y éticos, así como la capacidad de decidir cuándo y cómo utilizar el sistema en una situación determinada (par. 65)

Sin embargo, a la hora de concretar estas exigencias morales derivadas de nuestra autonomía, se afirma rotundamente que la intervención de todos los seres humanos *"no es posible ni deseable*". Lo primero es evidente, lo segundo es una afirmación que rompe con la misma justificación moral que se pretende. De esta forma, todo el sistema de una Inteligencia Artificial confiable pierde valor moral y, por lo tanto, fuerza y eficacia. La participación pasa de considerarse condición de posibilidad de la confianza, de ser moralmente exigible, a ser solo "recomendable" consultar a las partes interesadas que pueden ser afectadas directa o indirectamente por el sistema. Más aún, si la participación de todos los afectados resulta, incluso, "indeseable". ¿Cómo confiar en un sistema que no depende de nosotros? ¿Cómo gestionar, incluso conocer, los límites de la tecnología? ¿Cómo definir lo posible y lo imposible?

El paso del "*todos"* reflejado en los principios éticos (nivel 1) al "*todos"* pragmático en las diferentes prácticas e instituciones (niveles 2 y 3) es una de las dimensiones más importantes de la reflexión ética y la clave para convertir el saber moral en un recurso moral, finalidad última de toda ética aplicada (García-Marzá, 2004). Solo así nos acercaremos a las bases éticas de la confianza en la Inteligencia Artificial.

Para esta motivación, para generar confianza, no es suficiente con una declaración de buenas intenciones por parte de los profesionales o de sus organizaciones. Desde el principio de explicabilidad como principio ético toda gestión de la información que pretenda validez moral debe pasar, en cada situación concreta, por hacer públicos los esfuerzos realizados. No se trata sólo de una disposición a la sinceridad, sino de que esta disposición adquiera el rango de un compromiso público, en el doble sentido de trasparencia y de participación que ya hemos analizado. Con esta idea trabajan las teorías del diseño institucional al remitir la capacidad de producir confianza a este "potencial de justificación discursiva" (Goodin, 2008).

Diseñar parece un término pretencioso y arriesgado, pero esta primera impresión desaparece cuando nos percatamos que su raíz etimológica designare nos indica la tarea de señalar qué institucionalización d ellos requisitos es la más adecuada a un contexto particular. Si bien diseñar o rediseñar son actividades intencionales, deben entenderse siempre como aportaciones a una deliberación pública acerca de qué infraestructura ética es la más adecuada para que nuestras organizaciones generen confianza. Es decir, acerca de cómo sostener y desarrollar la credibilidad y la reputación de nuestra organización. Dicho de otro modo, para responder de esta justificación pública no basta con la buena voluntad del profesional, sino que debemos contar con procesos y estructuras organizativas que permitan y potencien las directrices éticas señaladas.

Para anclar estas bases éticas de la confianza necesitamos tanto la transparencia, trazabilidad e inteligibilidad de la información, como la posibilidad de que los grupos de interés o sus representantes puedan participar desde la declaración de utilidad, hasta el cálculo de resultados, pasando por el mismo diseño. No hay transparencia sin posibilidad de participación, sin poder decidir, por ejemplo, de qué se informa o cómo calculamos las consecuencias y para quién son. Por así decirlo, el riesgo moral, la posibilidad de que otros sufran las consecuencias de mis decisiones o prácticas, de que no ocurra lo que esperábamos, es directamente proporcional a la participación. No hay autonomía sin posibilidad de ser incluido en las decisiones que acabarán afectándonos. Y esta participación, como muy bien resalta el documento, en todo el ciclo de vida del algoritmo, por supuesto también en su creación. No solo debemos acompañar a la tecnología, debemos adelantarnos, ir siempre un paso por delante (Etzioni 2017; Dignum, 2018).

De acuerdo con estas exigencias, esta es *la propuesta de una ética digital dialógica*, un diseño capaz de responder y de facilitar la aplicación del principio de explicabilidad, un diseño que no abandone las decisiones en una expertocracia irresponsable, debería adquirir la forma de una infraestructura ética con cuatro elementos básicos, adaptables a cada situación particular y a cada estructura organizativa concreta:

1. *Códigos ético y de conducta* El primer paso en esta generación de confianza lo constituye la elaboración y publicación de los códigos éticos y de conducta. Se trata de documentos formales donde encontramos una declaración explícita de los valores que deben orientar la conducta de empleados y directivos, propiciando así las buenas prácticas y marcando el carácter y la personalidad de la organización. Su función es, por lo tanto, doble: - desde el punto de vista interno, formalizar los valores y criterios de decisión que definen la cultura organizativa; desde el punto de vista externo, gestionar la reputación de la organización. No sólo nos presenta los valores que definen el carácter o ética de la organización, sino también los compromisos que está dispuesta a asumir para crear esta voluntad común y las conductas necesarias para su realización. Unos códigos que deben incluir en su seno su compromiso con los sistemas internos de cumplimiento y con las auditorías externas (García-Marzá, 2017).
2. *Comité de ética*. Se concibe como un *espacio de participación* y diálogo de los diferentes grupos de interés en el interior mismo de la organización, encargado del seguimiento y control del programa de ética y cumplimiento, así como del impulso

de las directrices éticas y sus diferentes procedimientos. Su función es triple: asesorar en temas relacionados con la interpretación y aplicación del código ético; resolver las notificaciones de sugerencias, alertas y denuncias realizadas a través de la línea ética; promover la información y formación de los empleados y directivos en el programa de ética y cumplimiento. La confianza en el comité dependerá, a su vez, de la confianza que sean capaces de generar sus componentes, como muy bien ha mostrado el "fiasco" del comité de ética de Google.

3. *Línea ética:* la participación buscada no puede limitarse a un pequeño comité que, aunque aporte la presencia y voz de los grupos de interés internos y externos, no sustituye a la voz de todos. Debemos establecer canales de comunicación que permitan la participación de todo aquel que quiera hacerlo, siempre centrada en el cumplimiento de los compromisos éticos adoptados. La comunicación no debe limitarse a la denuncia de malas prácticas. También, y en especial, debe potenciar una cultura ética a través de la implicación de los empleados en la formación y el desarrollo, en la gestión, en suma, de los valores éticos. Esta participación debe incluir a los grupos *internos* (investigadores, desarrolladores, directivos, trabajadores, etc.), como *externos* (compañías de la competencia, agencias gubernamentales, consumidores, organizaciones de la sociedad civil, etc.

Con estos tres instrumentos de gestión de la ética, estamos ante diferentes pasos progresivos para la generación de confianza en todo el tejido socio-técnico de la realidad digital. Ahora bien, la existencia y el funcionamiento de esta infraestructura ética tiene, a su vez, que ser verificada externamente. Este es el papel de la auditoría ética, de la evaluación de la IA fiable que nos proporciona el documento (par.112).

4. La *auditoría ética*. La auditabilidad se refiere en el documento a la capacidad de un sistema de IA de someterse a la evaluación de sus algoritmos, datos y procesos de diseño. De ahí que forme un elemento fundamental para el seguimiento de la participación (Buruk et. al, 2020).

Como conclusión, al diferenciar claramente entre tres dimensiones básicas para una IA confiable: legal, ética y robusta, las directrices se refieren también, por consiguiente, al *riesgo moral*, a los riesgos derivados del incumplimiento de los principios éticos y de sus requisitos de aplicación. No se trata de sancionar, sino de crear una cultura donde la transparencia y la participación dificulten las malas prácticas y reconozcan y potencien las buenas. La gestión de la confianza es inversamente proporcional a este riesgo moral, a la desconfianza que produce no saber si la organización va a cumplir o no con lo que se espera de ella. Este es el principal objetivo del principio de explicabilidad y la justificación moral de la exigencia de participación que le es inherente.

Referencias

Apel, K.O., (1988). *Diskurs und Verantwortung das Problem des Übergangs zur postkonventionellen Moral,* Frankfurt: Suhrkamp

Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, *23*(3), 387-399.

Calvo, P (2021). El gobierno ético de los datos masivos. *Dilemata. Revista internacional de* éticas aplicadas, (34), 31-49

Cortina, A., Conill,J; García-Marzá (eds.) (2008). *Public reason and applied ethics: The ways of practical reason in a pluralist society*. Londres: Ashgate Publishing.

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics Inf Technol* (20), 1–3.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, *21*(4), 403-418.

European Data Protection Supervisor (2015). *Hacia una nueva ética digital. Datos, dignidad y tecnología* (Dictamen 4/2015). ESPD. Recuperado de: https://edps.europa.eu/sites/edp/files/publication/15-09-11_data_ethics_es.pdf.

COM(2019) 168 final (2019). *Generar confianza en la inteligencia artificial centrada en el ser humano*. Bruselas: Comisión Europea.

High-level expert Group on Artificial Intelligence (2019). *Ethics Guidelines for Trustworthy AI. Brussels. European Commission*. Recuperado de: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.

Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology*, *31*(1), 1-8.

Floridi, L. Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, *32*(2), 185-193.

García-Marzá, D. (2004). Ética empresarial: del dialogo a la confian*z*a. Madrid: Trotta

García-Marzá, D. (2012). Kant´s Principle of Publicity, *Kant-Studien. Philosophische Zeitschrift der Kant-Gesellsschaft*, (103), 96-113.

García-Marzá, D. (2017). From ethical codes to ethical auditing: An ethical infrastructure for social responsibility communication. *El profesional de la información*, *26*(2), 268-276.

García-Marzá, D. (2019). Repensar la democracia. Estrategia moral y perspectiva crítica en KO Apel. *Daimon Revista Internacional de Filosofía*, (78), 75-89.

García-Marzá, D. & Calvo, P. (2022). Democracia algorítmica: ¿un nuevo cambio estructural de la opinión pública?. *Isegoría*, (67), e17-e17, 1-16. https://doi.org/10.3989/isegoria.2022.67.17

Generalitat Valenciana (2018). *Estrategia de Inteligencia Artificial de la Comunitat Valenciana*, Valencia: GVA.

Goodin, R. E. (Ed.) (1998). *The theory of institutional design*. Cambridge: Cambridge University Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kant, I. (1795). *Zum ewigen Frieden. Ein philosophischer Entwurf*, (ZeF), AA, VIII.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679

Rehg, W. (2015). Discourse ethics for computer ethics: a heuristic for engaged dialogical reflection. *Ethics and Information Technology*, *17*, 27-39.

UNESCO (2022). *Recomendación sobre la ética de la Inteligencia ArtificiaL de UNESCO*. Montevideo, Uruguay: UNESCO.

Yuste, R., Goering, S., Arcas, B. A. Y., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., Friesen, P., Gallant, J. L., Huggins, J. E., Illes, J., Kellmeyer, P., Klein, E., Marblestone, A. H., Mitchell, C., Parens, E., Pham, M. Q., Rubel, A., Sadato, N., . . . Wolpaw, J. R. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, 551(7679), 159-163. https://doi.org/10.1038/551159a

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 115-130
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.562371

Ética discursiva e inteligencia artificial. ¿Favorece la inteligencia artificial la razón pública?*

Discourse ethics and artificial intelligence. Does artificial intelligence favor public reason?

*JESÚS CONILL SANCHO***

Resumen. Este artículo muestra que la ética del discurso en versión de la *ethica cordis* contribuye a mantener una actitud crítica de discernimiento ante las tecnologías de la inteligencia artificial. Propone distinguir entre opinión pública y razón pública, para destacar el sentido crítico del uso público de la razón en la línea de Kant, Rawls, Habermas y Cortina. También se propone afrontar las dificultades para ejercer la razón pública en la era digital: impacto de la inteligencia artificial en la comunicación, hiperconectitividad, datificación, "panóptico digital" y la espiral del silencio, que ponen en peligro el uso de la razón y la intimidad personal.
Palabras clave: Ética, Inteligencia, Razón, Crítica, Opinión pública, Datificación

Abstract. This article shows that discourse ethics in version of *ethica cordis* contributes to maintain a critical attitude of discernment towards artificial intelligence technologies. It proposes to distinguish between public opinion and public reason, to highlight the critical sense of public use of reason in the line of Kant, Rawls, Habermas and Cortina. It also aims to address the difficulties in exercising public reason in the digital age: impact of artificial intelligence on communication, hyperconnectivity, datafication, "digital panopticon", and the spiral of silence, that endanger the use of reason and personal intimacy.
Keywords: Ethics, Intelligence, Reason, Criticism, Public Opinion, Datafication

---

Recibido: 24/ 03/2023. Aceptado: 04/06/2023.

* Este estudio se inserta en el Proyecto de Investigación Científica y Desarrollo "Ética cordial y Democracia ante los retos de la Inteligencia Artificial" PID2019-109078RB-C22 financiado por MCIN/ AEI /10.13039/501100011033 y en las actividades del grupo de investigación de excelencia PROMETEO/2018/121 y del "Programa Prometeo 2022 para grupos de investigación de excelencia, CIPROM/2021/072"de la Generalitat Valenciana. Agradezco en especial las sugerentes intervenciones de César Ortega, Javier Gracia, Juan Carlos Siurana y Pedro Pérez-Zafrilla, así como las de Domingo García-Marzá, Ramón Feenstra y Patrici Calvo.

** Catedrático de Filosofía Moral y Política (Universidad de Valencia). Sus libros más r recientes son: *Intimidad corporal y persona humana. De Nietzsche a Ortega y Zubiri* (2019) y *Nietzsche frente a Habermas. Genealogías de la razón* (2021), ambos en la editorial Tecnos. Contacto: Jesus.Conill@uv.es

1. Ética discursiva en su versión cordial

La ética discursiva propuesta por Karl Otto Apel (1985; Siurana, 2003) y Jürgen Habermas (1985; García-Marzá, 1992; Ortega, 2021) se inscribe en un proyecto ilustrado de modernidad crítica, en el que el *télos* ínsito al lenguaje y la comunicación orienta la acción humana. Articulada en dos partes, una de fundamentación y otra de aplicación, la ética discursiva se constituye a través de una transformación hermenéutico-comunicativa de la ética kantiana en una ética de la responsabilidad solidaria, capaz de afrontar las consecuencias planetarias del desarrollo científico-técnico en el medio de la intersubjetividad con unos mínimos de justicia (Cortina, 1985, 1986, 2007 y 2021).

Este proyecto mantiene el sentido eleuteronómico y emancipador de la razón moral moderna, más allá del uso instrumental y estratégico de la razón. La autonomía personal, expresada reflexivamente a través de las pretensiones de validez en los actos de habla, y la capacidad argumentativa en la interacción comunicativa abren un nuevo mundo, en el que pueda escucharse la voz de todos los posibles afectados. Libertad e intersubjetividad conforman la nueva perspectiva en la que nos reconocemos y respetamos mutuamente como personas en la modalidad de interlocutores válidos, pues "somos lo que somos gracias a nuestra relación con otros" (Mead, 1972: 307).

En la versión de la ética discursiva que ofrece la *ethica cordis*, la idea kantiana de persona moral se transforma en la de interlocutor válido, un ser dotado de "competencia comunicativa", es decir, capaz de defender en la interacción comunicativa la exigencia de atenerse a las pretensiones racio-cordiales de validez argumentativa mediante el diálogo (Cortina 1985 y 2007; Martínez-Navarro, 2022). Se trata del giro pragmático-lingüístico (comunicativo) de la filosofía transcendental, que se presenta en forma de pragmática transcendental y/o universal, en la que se mantiene un criterio de "validez" y de "incondicionalidad", irreductible a las vigencias que se imponen fácticamente.

Esta "ética discursiva", también denominada "dialógica" y "comunicativa", tiene como principio moral el ideal dialógico en el medio de la comunicación. En su fundamentación, pues, no basta la razón instrumental y estratégica, sino que recurre a una nueva figura de la razón práctica, que no es la aristotélica ni tampoco estrictamente la kantiana, aunque tenga en cuenta sus ineludibles aportaciones, sino a una razón práctica comunicativa, cuya "fuerza vinculante" se presupone en el uso del lenguaje y se descubre mediante reflexión reconstructiva, incluso a través de una peculiar genealogía de la razón (Habermas, 2019; Conill, 2021).

La razón comunicativa, como toda forma de razón y de saber, está regida por un peculiar interés. Pero, a diferencia de la razón técnica que está orientada por el interés de dominio, la razón comunicativa busca el entendimiento mutuo y, en la medida en que pone en marcha su dinamismo crítico, se complementa con el interés por la emancipación o liberación humana. Sin tener en cuenta la función de estos intereses no se comprendería el sentido de las diversas formas de la razón en la interacción humana.

La interacción humana de la que parte la ética discursiva es la del *factum* argumentativo, en el que pueden expresarse y explicitarse intersubjetivamente las presupuestas pretensiones de validez (sentido, inteligibilidad, verdad, corrección, veracidad, autenticidad) que al menos están siempre latentes de modo virtual como condición de posibilidad en los juegos lingüísti-

cos de la comunicación humana (Apel, 2017). Estas pretensiones de validez son condiciones universales de la constitución y validez del discurso humano. De lo contrario, no es posible coordinar la acción humana intersubjetivamente y orientarla hacia el entendimiento mutuo, sino que la acción estaría regida únicamente por el interés técnico y estratégico de dominio.

Si la interacción humana sólo se rigiera por la razón instrumental y estratégica, la moral sólo se compondría de imperativos hipotéticos y no contaríamos con ninguna incondicionalidad como base de la exigencia de un posible imperativo categórico. En la versión cordial de la ética discursiva se refuerza el valor de la acción comunicativa, cuya racionalidad tiene primacía sobre la técnica y estratégica, porque en ella cabe descubrir un "momento de *incondicionalidad*".

La acción comunicativa es aquella en que los actores se coordinan orientados por las pretensiones de validez. Pues, aunque incluso la acción lingüística puede utilizarse de modo instrumental y estratégico, el *télos* inherente al lenguaje humano es el entendimiento mutuo, de tal manera que el modo primordial y "originario" de usar el lenguaje es el que pretende lograr el acuerdo. Por tanto, según Adela Cortina, hay un uso del lenguaje más valioso que otro, hay una jerarquía axiológica entre los usos del lenguaje, porque "el *télos* del lenguaje, inherente a la razón comunicativa, muestra la primacía de ésta frente a la estratégica" y por eso en el orden del discurso y de la auténtica comunicación ha de prevalecer la "fuerza del mejor argumento" en busca de un consenso racional, anticipado contrafácticamente como idea regulativa (Cortina, 1985, 1990 y 2007).

En la comprensión pragmático-lingüística del uso de la razón, en el intercambio discursivo de las razones prácticas, la fuerza sin violencia del mejor argumento sólo puede entrar en juego si los participantes adoptan la "perspectiva del nosotros", desde la que juzguen "*imparcialmente*", sin perder de vista la dimensión de la "validez del deber" [*Sollgeltung*], qué importa en interés de todos los posibles afectados (Habermas, 2019: 784). El "nosotros" inclusivo no se restringe a una comunidad particular, sino que se extiende contrafácticamente a todas las personas; la razón práctica en la ética discursiva se orienta por esta idea de justicia comunicativa, adoptando una perspectiva del nosotros que trasciende todos los límites sociales y comunidades locales (Habermas, 2019: 786).

La ética discursiva sigue siendo una ética universalista como la kantiana, pero ahora el principio de universalización incorpora el consecuencialismo de la satisfacción de los intereses de los afectados, reformulando así el imperativo categórico mediante una razón dialógica. La "voluntad racional", es decir, lo que "todos podrían querer", como criterio para legitimar las normas morales, se traslada al orden del diálogo entre todos los afectados por la norma, al menos de modo virtual: "todos los seres capaces de comunicación lingüística deben ser reconocidos como personas, puesto que en todas sus acciones y expresiones son interlocutores virtuales" (Apel, 1985: II, 380).

En esta última formulación, el concepto de persona moral se sitúa en el orden dialógico, en la interacción comunicativa constituida por el reconocimiento recíproco entre interlocutores virtuales con competencia comunicativa y, por tanto, capaces de plantear pretensiones de validez y defenderlas discursivamente. Lo decisivo desde la perspectiva de la ética discursiva es mantener la distinción crítica entre "vigencia" (facticidad) y "validez".

Pero la validez procedimental requiere el presupuesto de alguna valoración fundamental (dignidad humana, acuerdo racional), alguna elección de valor; en último término, es nece-

saria una opción personal, pues "la *realización práctica de la razón* a través de la voluntad (buena) siempre necesita un compromiso que no puede demostrarse" (Apel, 1985: II, 392), pero que puede someterse a una constante consideración crítica como expresión de la libertad racio-cordial.

La ética discursiva, pues, fomentará una "actitud humana", configuradora de un êthos en la vida personal y compartida socialmente, cultivando la voluntad de verdad (el interés por la verdad) y de diálogo (la deliberación) sobre las necesidades e intereses de los afectados. El procedimiento de formación discursiva de la voluntad tiene en cuenta la libertad de las personas y su interrelación intersubjetiva en la vida social, orientada por la comunidad ideal de comunicación. Aquí la universalización adquiere un carácter dialógico a través del reconocimiento recíproco en la praxis comunicativa, en una argumentación polifónica donde se vive la tensión entre la incondicionalidad de las pretensiones de validez y la facticidad de los contextos concretos. Por eso, prosiguiendo el espíritu ilustrado la ética discursiva aboga por una razón comunicativa en forma de "publicidad razonante" en la vida social y sus instituciones (incluso de la democracia mediante la formación deliberativa de la voluntad) (Habermas, 2019: II, 765-766).

Se plantea entonces la cuestión de los procesos de aprendizaje correspondientes a la moral, pues en principio somos capaces de aprender también con respecto a la dimensión práctico-moral, de modo especial a través de la función educativa de los procesos comunicativos (Gozálvez, 2012; Gracia, 2018. Pues igual que en la dimensión biológica tenemos un triple cerebro a partir del más primitivo, que no podemos erradicar, también (¡o más!) tenemos en el orden sociocultural formas tribales de pensamiento que nos es imposible erradicar y difícil superar. No obstante, según Habermas, es constatable que la razón práctica ha dejado también huellas de procesos de aprendizaje y de progreso moral; que la razón comunicativa tiene un potencial cognitivo, es decir, que se puede aprender en el orden moral a través de la adopción recíproca de las perspectivas de los otros en busca del interés universalizable, al menos reconociendo al otro como persona igualmente libre (Habermas, 2019: II, 789 y ss.). Se presenta aquí la autocomprensión de la libertad racional del hombre primordialmente en virtud de su capacidad sociocognitiva de comunicación, porque "nadie puede ser autónomo [libre] por sí solo" (Habermas, 2019: II, 806).

2. ¿Razón pública en vez de opinión pública?

La ética de la razón cordial se basa en la ética discursiva, pero renovada y hasta transformada a través de la integración de nuevos componentes como las emociones y los sentimientos, los valores, las virtudes, la voluntad racio-cordial, la deliberación, la inteligencia sentiente (reforzada por las neurociencias) y el poder de la libertad en virtud de la intimidad corporal en el medio de la vida social (Conill, 2019). Uno de los escenarios más relevantes de esta nueva versión de la ética discursiva es la compleja intersubjetividad comunicativa (con todos sus ingredientes) a través del ejercicio de la razón pública. Así es como los ciudadanos ilustrados y razonables ponen de manifiesto su pluralismo y conviven con discrepancias, pero cooperando unidos por una razón cordial.

En el contexto moderno lo que llamamos "razón pública" quedó bien determinado en principio por la noción kantiana del "uso público de la razón" (Kant, 1968, 1983, 1985, 1981), que ha sido actualizado de modo significativo por Rawls (1971/1972, 1978, 1993, 1996, 2001, 1999: 573-611), Habermas (1981, 1987, 2019, 2022) e incluso por Sen (2010: caps. 15-17). La razón pública, que es expresión de la libertad efectiva y de la capacidad crítica, prosigue ahora el impulso ilustrado en su versión comunicativa.

Ahora bien, la razón pública no debería confundirse con la opinión pública, que tiene un carácter ambiguo. Por una parte, ésta constituye una presión social, ahora tecnológicamente reforzada; pero también lleva incorporado un potencial sentido crítico. Ya en la filosofía clásica antigua, Aristóteles distingue entre opinión y elección; ésta va conformando el êthos, no así la opinión. No bastan el apetito, el impulso, el deseo y la opinión, sino que hace falta cultivar el *orthós lógos*, la *recta ratio*, para orientar la acción como es debido —la buena *prâxis*— y configurar una vida virtuosa, potencialmente feliz. En los términos de una filosofía crítica (moderna y contemporánea), la voz de la razón comunicativa ha de conformar la "voluntad general" —la voluntad común—, al llevar ínsita —incoada— la pretensión de universalidad frente a los dogmatismos. No se reduce a la mera opinión, ni a la suma de opiniones, sino que constituye una vía crítica al estar orientada por una comunidad ideal de comunicación. Se instaura así el tribunal de la razón, que tiene vigor en la esfera pública universal, mostrando la efectividad de la razón en la historia, pues lo racional tiene que hacerse efectivo. Kant y Hegel, "*Moralität*" y "*Sittlichkeit*" han de articularse en la vida social mediante la razón pública[1].

Por su parte, la versión rawlsiana de la idea de razón pública considera su ejercicio como un deber de civilidad, pero se restringe a lo política y jurídicamente razonable (Conill, 2022). En cambio, la versión habermasiana amplía el sentido de la razón pública al conjunto de la sociedad civil, a todos los ámbitos de la cultura, como espacios de comunicación social abiertos a la deliberación pública. La visión ampliada de razón pública incorpora a todos los ciudadanos y se entiende como razonabilidad abierta, más allá de lo estrictamente político y jurídico, a los contenidos de las diversas concepciones del mundo en una sociedad postsecular (Habermas, 1981, 2009, 2019; Conill, 2021: caps. 1-3).

Habermas confía avanzar en la racionalización de la vida social y política mediante la aplicación del principio de publicidad (Habermas, 2009: 15)[2]. Lo explica ya en *Strukturwandel der Öffentlichkeit* y así lo reafirma posteriormente desarrollando un concepto normativo de publicidad y de política deliberativa (Habermas, 2022). No obstante, a pesar de esta aportación de un concepto normativo de publicidad, su uso ampliado comporta una cierta ambigüedad, ya que los promotores de la publicidad son las libres asociaciones que forman la red de comunicación a partir del entrelazamiento de las diversas manifestaciones de la opinión pública (*Öffentlichkeiten*). Tales asociaciones autónomas están especializadas en la creación y difusión de las convicciones prácticas, por tanto, en descubrir temas de relevancia social general, en contribuir a solucionar problemas, en interpretar valores y en producir

1 Agradezco a Juan Carlos Siurana sus intervenciones sobre la necesidad de aclarar este punto, que concilia las aportaciones de Kant y Hegel en la ética discursiva a través de la intersubjetividad.

2 Habermas utiliza en estos contextos indistintamente los términos "Publizität" y "Öffentllichkeit".

buenas razones o desvalorizar otras (Habermas, 2009: 62-63); son las instituciones de la libertad crítica.

Precisamente el modelo deliberativo de democracia explicará la fuerza legitimadora del proceder democrático con la ayuda del carácter racional de la formación de la opinión y de la voluntad (Habermas, 2009: 87). Este modelo parece ser un ejemplo especialmente convincente y eficaz para afrontar el profundo foso entre lo normativo y lo empírico en la ciencia política, intentando responder al problema de cómo ajustar —hacer compatible— un concepto normativo como el de "política deliberativa" con nuestra imagen realista de la sociedad mediática (Habermas, 2009: 87-88). Pues uno de los elementos del marco institucional de las democracias modernas y que constituyen el núcleo normativo de los Estados de derecho democráticos es el de una opinión pública [Öffentlichkeit] independiente que en tanto que esfera de libre formación de la opinión y de la voluntad une el Estado y la Sociedad Civil (Habermas, 2009: 89).

Pero Habermas, aunque es consciente de las crecientes dificultades que está generando la revolución de los nuevos medios de comunicación, no renuncia a su propósito de que el paradigma deliberativo no fracase en el intento de conectar sus "fuertes ideas normativas" con la actual complejidad social (Habermas, 2009: 92). El modelo deliberativo espera que su apuesta por la racionalización mejore la calidad de las decisiones, dirigiendo su mirada a las funciones cognitivas de la formación de la opinión y de la voluntad, así como a la busca cooperativa de la solución de problemas (Habermas, 2009: 95). Pretende así aprovechar el potencial racional de la deliberación y del discurso en el contexto de la comunicación de masas[3]. Lo que ocurre es que las meras opiniones están sometidas a las transacciones de la oferta y la demanda, pero no están regidas por la meta de encontrar soluciones legítimas a los problemas en litigio (Habermas, 2009: 109). Pues los participantes en la presunta comunicación de masas son espectadores y consumidores, que no hacen uso de la auténtica "razón pública", que siempre ha de llevar incorporado un sentido crítico y normativo (Marina, 2023).

En realidad, hace falta una publicidad crítica, no sólo nacional, sino europea y global o mundial. Un camino apropiado sería transnacionalizar las publicidades —opiniones públicas— existentes y hacerlas reflexivas, a fin de que la opinión pública masificada no quede sometida a la manipulación y la indoctrinación, sino que sirva a la ilustración (Habermas, 2009: 112 y ss., 137-138, 346).

3. El sentido crítico del uso público de la razón

El uso público de la razón es expresión del único derecho innato, que es la libertad; no es sólo una aplicación de las exigencias racionales al mundo político-moral, sino el núcleo mismo de la filosofía crítica, porque la crítica consiste en la posibilidad de que los ciudadanos libres presenten sus objeciones (Cortina, 2021; Conill, 2022). Con ello, el uso público-crítico de la razón se convierte en un presupuesto de la argumentación misma. Ya al comienzo de la *Crítica de la Razón pura* destaca Kant la dimensión comunicativa de la

3 Me parece que es imposible una "comunicación de masas", porque las masas no se comunican, sino que las relaciones entre ellas son de otro género.

autocomprensión humana, en la *Crítica del Juicio* reconoce el carácter nuclear de la comunicabilidad y, en el II Apéndice de *La paz perpetua*, el *Principio de Publicidad* adquiere carácter jurídico-político (Kant, 1968, 1983, 1985, 1981, García-Marzá, 2012).

El ámbito de los ciudadanos que pueden presentar sus objeciones a las propuestas y argumentaciones de la razón común humana no se circunscribe a una parte de la humanidad, la que comparte las mismas bases culturales, como si fueran los únicos capaces de comprender tales propuestas y argumentaciones, sino que el derecho a presentar objeciones es un derecho de cualquier ser humano, de modo que no se puede impedir a nadie que lo ejerza, ya que es un derecho de la humanidad, es decir, un derecho de rango cosmopolita (Cortina, 2021). Así lo confirma con claridad el siguiente texto de la *Crítica de la razón pura*: "También forma parte de esta libertad el exponer a pública consideración los propios pensamientos y las dudas que no es capaz de resolver uno mismo… […]. Esto entra ya en el derecho originario de la razón humana, la cual no reconoce más juez que la misma razón humana común, donde todos tienen voz" (Kant, 1983: A 751-752/B 779-780).

La referencia a la metáfora del tribunal de la razón y a la esfera de la opinión pública universal apunta a una estrecha relación entre la crítica de la razón y el cosmopolitismo (Bösch, 2007: 480). En esta línea interpretativa, Adela Cortina precisa que, al haber un derecho originario de la razón humana que no reconoce más juez que la misma razón, donde todos tienen voz, es necesaria una república mundial como condición de posibilidad del uso crítico de la razón y de la superación del dogmatismo.

El propósito de Kant con los textos citados de la "Doctrina Trascendental del Método" de la *Crítica de la razón pura* es profundizar en el proceso de Ilustración (*Aufklärung*), que expone claramente en *Beantwortung der Frage: Was ist Aufklärung?* (Kant, 1968: VIII, 33-42) y que exige fomentar un uso público de la razón. De ahí que la filosofía crítica kantiana se haya interpretado como una *paideia* (Munzel, 2012: XXI; Cortina 2021), con el propósito de cultivar la libertad interna. No obstante, todo ello supone procesos sociales de aprendizaje, como muestra Habermas reiteradamente en *Auch eine Geschichte der Philosophie*.

La razón crítica necesita del uso público de la razón, que todos los hombres se consideren "partícipes potenciales de una república mundial" (Andaluz, 2018: 438; Cortina, 2021). Una sociedad de ciudadanos del mundo es un requisito indispensable para llevar adelante el proceso de ilustración de la humanidad. Para que la razón crítica sea viable se requiere tomar en serio el mandato de la razón de avanzar hacia una sociedad cosmopolita (Cortina, 2021). La razón crítica reclama una esfera pública donde todos puedan tener voz, y la perspectiva del participante exige que el horizonte político de los ciudadanos se amplíe "para posibilitar una formación política común de la voluntad por encima de las fronteras nacionales y una acción política común en el nivel transnacional" (Habermas, 2019: II, 800).

¿Puede ejercerse la razón pública en el medio tecnológico de la IA?

Hemos visto que la ética discursiva surge precisamente desde la experiencia comunicativa y confía en la significación de una racionalidad comunicativa en el espacio público.

Pero ¿favorece la IA el desarrollo de la razón comunicativa y la razón pública, o constituye un nuevo obstáculo e incluso una peligrosa amenaza?

La ética discursiva, también denominada "comunicativa" o "del discurso", se desarrolla a partir del medio de la comunicación y presupone una noción de razón comunicativa operante en el espacio público, a diferencia de la razón técnica, instrumental y estratégica. En ese sentido, una de las pretensiones de la ética discursiva es contribuir a formar la propia conciencia y la voluntad, para que los ciudadanos lleguen a ser capaces de debatir y dialogar personalmente y en público.

No obstante, es palmaria la relevancia mundial del impacto de la IA en todos los órdenes de la vida. Nuestro mundo vital está siendo alterado —colonizado— por las tecnologías de la IA (en los ámbitos de la salud, el trabajo, la educación, la economía, las finanzas, la Administración), y muy especialmente se está transformando la interacción comunicativa. Pero en esta nueva realidad social los imperantes medios tecnológicos están reprimiendo la libertad que quería expresarse a través de la opinión pública y no permiten, o al menos dificultan gravemente, el ejercicio de la razón pública.

Los cambios sociales y culturales han conducido a una situación en que los ciudadanos se encuentran atrapados en entornos digitales y mediáticos que distorsionan la realidad y están dominados por los algoritmos diseñados para las redes sociales, creando lo que se han llamado "cámaras de eco" (que reducen el pluralismo a la insistente repetición de lo mismo), de modo que cada cual sólo escucha lo que ya comparte, impidiendo la comunidad de diálogo entre posiciones diversas. Cuando lo que se necesita es promocionar una ciudadanía madura e ilustrada también a través de los medios digitales y mediáticos, fomentando diálogos auténticamente argumentativos. Sólo si se logra una ciudadanía digital y mediática se podrá avanzar en la mejora de la calidad de la democracia, en la medida en que el universalismo moral y la justicia cordial inspiren la educación en la responsabilidad que exigen las nuevas condiciones sociales y los nuevos medios tecnológicos que han invadido la sociedad y el ejercicio de las profesiones.

La hiperconectividad digital no favorece el diálogo y la reflexión

Una interpretación de la era digital es la que considera que en ella emerge una nueva figura del ser humano como *homo poieticus* (Floridi/Sanders, 2003) en el seno de la cuarta revolución tecnológica de carácter digital. Pero también hay quienes han insistido (Foucault, 1975) en que la era digital hace peligrar la libertad por las sociedades modernas, debido al creciente control por medio de un cierto "panóptico digital". Todo se convierte en información y peligra la interioridad personal. La nueva tecnología digital posibilita una mayor dominación mediante el control del comportamiento cada vez más mecanizado (vaciado de significación vital propia) y, por tanto, más fácil de cuantificar y hasta de predecir (aprovechando el negocio de los *Big Data*).

Nos habríamos convertido en "organismos informacionales" (*inforgs*), que comparten el entorno informacional o "infoesfera" con otros seres que son artefactos inteligentes. Habría que reinterpretar la posición del hombre en la nueva realidad, en la que todo ha de estar funcionalizado y, por tanto, peligra la dimensión experiencial. Lo decisivo es

que las condiciones digitales no asfixien la experiencia del significado de la vida. Pues lo que está en juego es la forma de vida, si ésta se reduce primordialmente a producción y consumo de datos. Pues la coerción de la inmediatez y la brevedad de los mensajes está impidiendo la comunicación serena y la experiencia reflexiva, la disposición para meditar o pensar con serenidad y calma. Se necesitan espacios y tiempos de concentración, para encontrarse consigo mismo y orientarse en la vida, no dejarse invadir ni alterar constantemente por la hiperconexión digital. Pues se puede vivir hiperconectado, pero descentrado y desorientado. La calidad de vida no se mide por los instrumentos digitales que alteran continuamente la vida.

En el reino de los *infORGs* y del ciberespacio se suele confundir la comunicación con la mera información y la experiencia significativa con los meros datos; pero la comunicación no se identifica con la información. La sobrecarga de información impide pensar, reflexionar con calma y sosiego, formarse la conciencia moral, porque falta la necesaria concentración por exceso de alteración en el trepidante ritmo de la vida. La hiperconexión descentra y produce desorientación vital. Vivimos con más medios que nunca (Ortega y Gasset, 2005), con cada vez más sofisticados instrumentos digitales, pero sintiendo un deterioro de la auténtica calidad de vida, que no consiste en más bienestar, y sin haber logrado ser más y mejor humanos. ¿Se está produciendo realmente una mejor humanización universal de la vida compartida? ¿O está creciendo la desigualdad entre las personas por una nueva brecha tecnológica? (O'Neil, 2018).

Datificación de la esfera pública

La denominada "datificación" de la esfera pública no está siendo el mejor camino para ejercer una razón pública que incorpora la crítica humanizadora. En la nueva sociedad de la información algunos están "fascinados" por el creciente poder de las tecnologías de la IA en virtud de la hiperconectividad y los medios de codificación algorítmica (Calvo, 2019a). Pero la posibilidad efectiva de mantenerse permanentemente conectados con todo produce un flujo torrencial de datos, que constituye el nuevo elemento en el que se desarrolla la nueva forma de comunicación entendida como interacción (intercambio) de datos. Como bien se pregunta Patrici Calvo en relación con las democracias actuales, ¿se solucionan así mejor (con más imparcialidad, honestidad, integridad, comprensión y justicia) los problemas a los que nos enfrentamos en la vida real?

En su respuesta, Calvo señala varias deficiencias relevantes, de entre las que entresacamos las siguientes: 1) el carácter problemático de la presunta "objetividad" algorítmica, que se nutre del flujo masivo de datos proporcionados por una sociedad hiperconectada, pero que no puede ofrecer la determinación del "bien común" mediante la regla de la mayoría, ni garantiza la transparencia ni la fiabilidad, sino que introduce una nueva opacidad; 2) la imposible "neutralidad", debido a que hasta los modelos matemáticos incorporan sesgos (prejuicios) de diverso género, que reproducen los de cada sociedad, dependen de ciertas ideologías y de los diseñadores o programadores; 3) la nueva exclusión social, provocada por el silenciamiento de las personas más desfavorecidas de la sociedad o de los grupos que

carecen del poder suficiente para lograr relevancia pública (Cortina, 2022a, 2022b), porque están sometidos a alguna forma de aporofobia (Cortina, 2017).

Este nuevo mundo, de la interacción comunicativa ha generado un nuevo tipo de economía digital y de ciencia social computacional, a partir de una concepción renovada de la naturaleza humana en los términos siguientes: no somos individuos racionales, sino que "somos producto de nuestras redes sociales" (Moreno, 2021/2), seres influenciables mediante incentivos y presión social y, por tanto, predecibles y controlables.

Por otra parte, mediante la datificación se impone una concepción utilitarista de la vida y de la sociedad, reforzada por los nuevos instrumentos tecnológicos y el tipo de conocimiento que ofrecen, orientado primordialmente por el cálculo. De hecho, los procesos de "datatificación" o "datificación" van invadiendo incluso el ámbito moral produciendo una "etificación" (Calvo, 2019b). Se relega o anula el diálogo y la reflexión crítica regida por las pretensiones de validez, en favor de la agregación de las opiniones, preferencias y hábitos. Por tanto, se desconsidera o elimina la razón comunicativa en favor del cálculo y la matematización de los datos, que sirve para predecir el comportamiento y alimentar la razón instrumental y estratégica. Se elimina así la crítica recíproca entre los que participan en la esfera pública, porque en el fondo se presupone una concepción de la sociedad orientada por un individualismo cuantitativista.

Enfoque datificado de la ética versus ética discursiva

Por consiguiente, el enfoque datificado de la ética constituye un reto actual para la ética discursiva, en la medida que pretende ofrecer una mejor respuesta a qué es lo justo y lo bueno a través del análisis y procesamiento algorítmico de los datos masivos sobre lo moral procedentes de los individuos hiperconectados (Calvo, 2019a, 2019b).

Según Calvo, este nuevo enfoque posibilitado por el uso de las tecnologías de la inteligencia artificial constituye un nuevo intento de colonización sistémica del mundo de la vida, que distorsiona el saber moral e impide el entendimiento mutuo, porque más bien sirve para fomentar la polarización social, como ha destacado Pedro Pérez-Zafrilla (Pérez-Zafrilla, 2021a). Por tanto, desde la ética discursiva de Apel, Habermas, Cortina y García-Marzá, Patrici Calvo se propone criticar y superar esta datificación de lo moral, que en realidad conduce a una nueva versión del utilitarismo; pues la etificación consiste en recopilar, cuantificar, procesar y gestionar datos masivos sobre opiniones, preferencias y conductas de la ciudadanía hiperconectada para, desde el criterio del mayor bien para el mayor número, establecer qué es lo moralmente válido. En cambio, desde la ética discursiva hay que restablecer el sentido del diálogo de los afectados para la consecución de un acuerdo racional, la reflexión crítica ateniéndose a las pretensiones de validez en el medio de la interacción comunicativa. La ética no puede convertirse en una ciencia predictiva de la conducta y estar a merced de la razón instrumental y estratégica. Y la ética discursiva no debe difuminarse en función de la datificación de lo moral mediante los algoritmos, pues supondría un regreso al nivel convencional de lo moral, en el que el peso de las vigencias sociales hace desaparecer el diálogo reflexivo y la deliberación crítica.

El poder represor de la opinión pública

Una de las concepciones más lúcidas de la opinión pública es la que considera que lo que en ella se expresa son las opiniones y conductas que pueden mostrarse en público sin miedo al aislamiento, porque lo que más tememos es el aislamiento, dado que necesitamos a los otros para vivir a gusto, siendo acogidos por los demás. Como decía Tocqueville, "la gente teme al aislamiento más que al error" (Noelle-Neumann, 1982: 25). Todo lo contrario de lo que hemos visto que proponía Kant en relación con el sentido crítico del "uso público de la razón": "cada uno de los [ciudadanos libres] tiene que poder exponer sin temor sus objeciones e incluso su veto" (Kant, 1983: A 738-739 B 766-767).

Esta teoría de la espiral del silencio muestra el enorme poder de la autocensura (Cortina, 2022a, 2022b). A diferencia de los procedimientos brutos y agresivos de la represión manifiesta y violenta, el mecanismo más sutil y eficaz para silenciar determinadas propuestas en la vida pública, que tiene su raíz y está entrañado en la naturaleza de nuestro ser social, funciona a través de esa compleja realidad que es la *opinión pública* (Pérez-Zafrilla, 2021b).

"Hoy se puede demostrar que, aunque la gente vea claramente que algo no es correcto, se mantendrá callada si la opinión pública (opiniones y conductas que pueden mostrarse en público sin temor al aislamiento) y, por ello, el consenso sobre lo que constituye el buen gusto y la opinión moralmente correcta, se manifiesta en contra" (Noelle-Neumann, 1982: 14). El paréntesis que aclara qué sea la opinión pública es sumamente expresivo: la constituyen las opiniones y conductas que pueden mostrarse en público sin temor al aislamiento.

Desde la noción expuesta de "razón pública" hay que criticar el silencio a que se ven sometidos los disidentes y no claudicar a la presión social de la opinión pública. Porque la tendencia que sentimos a imitar y asimilarnos a los demás no proviene de la pretensión de verdad, sino de otro motivo más fuerte: evitar el aislamiento. Una tendencia tan poderosa, que ha hecho que se comprenda su vigencia social: "Quizá no simpaticemos con la naturaleza social del hombre, pero tenemos que intentar comprenderlo para no ser injustos con la gente que se mueve con la multitud" (Noelle-Neumann, 1982: 14).

Según Elisabeth Noelle-Neumann, la espiral del silencio es un proceso en que las observaciones realizadas en unos u otros contextos incitan a unas gentes a expresar sus opiniones y a otras, a tragárselas, a mantenerse en silencio, hasta que en un proceso en espiral un punto de vista domina la vida pública (Noelle-Neumann, 1982: 22). Pero no domina la vida pública ese punto de vista porque sea el más verdadero, sino que triunfa porque en todas las sociedades, también las oficialmente democráticas y en apariencia tolerantes, funciona la autocensura de aquellas opiniones que no van a ser bien acogidas. Por supuesto, en las totalitarias la autocensura va de suyo, excepto en el caso de disidentes valerosos, que suelen pagar su osadía, pero en todas las sociedades funciona la espiral del silencio, lo cual constituye un grave obstáculo para el auténtico pluralismo, la deliberación y la democracia.

Se podría decir que de igual modo que las democracias en los últimos tiempos no mueren tanto por aparatosos golpes de estado, sino por el paulatino deterioro de las instituciones y porque pierden fuerza unas reglas "morales" que la comunidad aceptaba y respetaba (Levitsky y Ziblatt, 2018; Cortina, 2021: 46), asimismo hay ideas valiosas que desaparecen

de la vida social no porque dejen de ser convincentes tras un debate abierto, sino porque las silencian quienes temen al aislamiento más que al error.

Y actualmente habría que añadir: porque temen al aislamiento y al linchamiento público más que a la mentira. En tiempos de la llamada "posverdad" este riesgo es mayor si cabe que antes, porque la posverdad puede caracterizarse como "una mentira emotiva" (Nicolás, 2019). Según Wikipedia, se trata de un "neologismo que describe la distorsión deliberada de una realidad con el fin de crear y modelar la opinión pública e influir en las actitudes sociales, en la que los hechos objetivos tienen menos influencia que las apelaciones a las emociones y a las creencias personales".

Las noticias falsas tienen mucho más impacto que las verdaderas, pues al generar más interacción atraen más la atención. Y como los algoritmos registran y valoran la interacción de los seguidores, produciendo entre otros fenómenos el dominio de los influyentes o influenciadores (Siurana, 2021), se pone en peligro la auténtica comunidad de comunicación.

El poder de los medios de comunicación llega actualmente a sustituir la experiencia directa y vivida personalmente, de tal manera que se produce un fenómeno muy llamativo, que ya advirtió en su momento Maquiavelo, aunque ahora haya que aplicarlo a la nueva situación en la que vivimos actualmente por la influencia inmediata de las tecnologías y redes de comunicación masiva: "los hombres, en general, juzgan más por los ojos que por las manos, que a todos es dado ver, pero tocar a pocos. Todos ven lo que pareces, pero pocos palpan lo que eres y esos pocos no se atreven a oponerse a la opinión de la mayoría" (Maquiavelo, 1985: 140 y 141; 1987: 103).

El comportamiento de la mayoría de la gente en su medio social se rige por las opiniones que se percibe que van ganando terreno y se convierten en dominantes. De tal modo que los que confían en la victoria se expresan en público, pero los perdedores tienden a callarse (Noelle-Neumann, 1982: 27, 40 y 44). Y la cuestión se ha agravado por el funcionamiento de las redes sociales, que transmiten rápidamente ("viralizan") las "sentencias" sobre lo que es aceptable por la opinión pública a través de nuevos mecanismos inquisitoriales como el movimiento *Woke* (un pensamiento rigorista sobre lo que es lícito pensar) o el castigo llamado "cancelación", que consiste en atacar a determinadas personas con el objeto de destruir su reputación y de provocar su muerte social. Aquí ya no se teme sólo al aislamiento, sino hasta incluso a la pérdida del trabajo profesional y de los propios medios de vida. Sigue siendo verdad, como decía Nietzsche, que "nos las arreglamos mejor con nuestra mala conciencia que con nuestra mala reputación" (Nietzsche, 1986: § 52; Conill, 2016: 806-807). Se está imponiendo por estas nuevas vías el miedo a la mala reputación y a la pérdida de estatus en la vida social. Cada vez dependemos más del beneplácito de los demás, expresado a través de unos medios de comunicación de masas que se convierten en potentes instrumentos de control social y que merman la libertad de las personas.

¿Sobrevivirá la razón pública al poder de la IA?

Igual que Nathanïel Persily preguntaba si la democracia puede sobrevivir a internet, tenemos que preguntarnos si lo que ha significado la razón pública podrá sobrevivir al poder que emerge del uso de la inteligencia artificial.

La nueva situación va en contra de la tradición de la opinión pública que surge desde el siglo XVIII y que confiaba en que la humanidad había iniciado un proceso de ilustración, en virtud del cual las personas pueden y deben atreverse a servirse de la propia razón: "¡Atrévete a servirte de tu propia razón!" (Kant). Pues de la libertad personal, según Kant, "forma parte […] el exponer a pública consideración los propios pensamientos" y no reconocer "más juez que la misma razón humana común, donde todos tienen voz" (Kant, 1983: A 751-752 B 779-780).

Existen dos concepciones y dos prácticas de la opinión pública que están funcionando en la esfera pública de las sociedades modernas y contemporáneas. Una es la concepción psico-sociológica: "Lo que importa no es la calidad de los argumentos, sino cuál de los […] bandos tiene la fuerza suficiente como para amenazar al contrario con el aislamiento, el rechazo y el ostracismo" (Noelle-Neumann, 1982: 288). Según Noelle-Neumann, es la presión social la que tiene realmente fuerza para cambiar los puntos de vista y funciona como control social, porque afecta a todos. Según esta concepción de la opinión pública, incluso los que conocen la realidad no se atreven a contradecir la opinión mayoritaria. La vigencia social se convierte en norma, en tribunal evaluador y juzgador de las opiniones y las conductas, en la medida en que ejerce un poder de control y represión.

En cambio, la concepción considerada "normativa" cree haber logrado una instancia crítica que no se somete a las vigencias sociales y tribales, ni a los mecanismos tecnológicos que las refuerzan, sino que confía en el poder específico de la razón crítica y comunicativa, expresada a través de las pretensiones de validez en un espacio de razones públicas. Esta línea, que se nutre de la herencia kantiana es la que desarrolla la ética discursiva en sus diversas versiones, tanto por parte de Apel, Habermas, Cortina y García-Marzá, como por algunos representantes de las nuevas generaciones de la teoría crítica de la Escuela de Frankfurt. Por ejemplo, atendiendo a la sugerencia de César Ortega en una sesión del grupo de investigación, podría aprovecharse el "principio de justificación" propuesto por Reiner Forst (Ortega, 2021: 377-384) para reforzar el espacio potencialmente crítico de las "razones" frente a las relaciones de dominación, que siguen impidiendo el derecho básico a la justificación, es decir, a ser tenido en cuenta en cada ámbito de la vida social ofreciendo y recibiendo razones. Forst instaura esta "exigencia incondicional" de que cada persona sea respetada como alguien que "merece razones" como un principio de reciprocidad universal, que en el tema que nos ocupa podría aplicarse al ámbito del ciberespacio comunicativo creado por las tecnologías de la inteligencia artificial.

La necesidad de una ética en el mundo digital ha conducido ya a formular unos principios éticos, como en la bioética (no maleficencia, beneficencia, autonomía y justicia), pero a los que se añade un principio decisivo en el nuevo contexto, como es el de explicabilidad o de trazabilidad, en el que han insistido especialmente Adela Cortina (2019) y Domingo García-Marza (García-Marzá/Calvo, 2022)[4]. Según este decisivo principio, los afectados

4 Intervenciones en el Curso virtual *Ética de la ciencia: transparencia y explicabilidad*, Universidad Jaume I de Castellón, Escuela de Doctorado, Facultad de Ciencias de la Salud, 28 de septiembre-1 de octubre de 2021 y en el Coloquio Internacional sobre "Ética discursiva e IA" (*XV Coloquio Latinoamericano sobre Ética del Discurso y IX Coloquio de la Red Internacional de Ética del Discurso*, organizado por la Fundación ICALA, Río Cuarto, Argentina, 11-12 de noviembre de 2021) y en el Congreso de Filosofía en la UJI, 6-9 de Abril de 2022.

tienen que poder controlar el uso de sus datos y conocer los algoritmos que los manejan, pues también los sistemas de IA operan con sesgos, que incluso son más invisibles (O'Neil, 2018), porque vienen a ser cajas negras. Por eso, los afectados por el mundo digital tienen que poder comprender los algoritmos que manejan sus datos, conocer la trazabilidad, quién los construye y con qué criterios y objetivos, es decir, cumpliendo las exigencias de la ética discursiva, aplicadas al poder de estas nuevas tecnologías de la IA.

Referencias

Andaluz, A. (2018), "La hospitalidad en el cosmopolitismo kantiano", en García-Marzá, D., Lozano, J.F., Martínez Navarro, E. y Siurana, J.C. (comps.), *Ética y filosofía política*, Madrid: Tecnos, 435-445.

Apel, K.-O. (1985), *Transformación de la filosofía*, Madrid: Taurus, 2 vols.

Apel, K.-O. (2017), *Transzendentale Reflexion und Geschichte*, herausgegeben und mit einem Nachwort von Smail Rapic, Berlin: Suhrkamp.

Bösch, M. (2007). Globale Vernunft. Zum Kosmopolitismus der Kantschen Vernunftkritik. *Kant-Studien*, vol. 98, nº 4, 473-486.

Calvo, P. (2019a), Democracia algorítmica: consideraciones éticas sobre la dataficación de la esfera pública. *Revista del CLAD Reforma y Democracia*, nº 74, 5-30.

Calvo, P. (2019b), Etificación, la transformación digital de lo moral. *KRITERION*, Belo Horizonte, nº 144, Dez., 671-688.

Conill, J. (2016), La intimidad corporal y sus bases neurobiológicas. *Pensamiento*, 273, 789-807.

Conill, J. (2019), *Intimidad corporal y persona humana*, Madrid: Tecnos.

Conill, J. (2021), *Nietzsche frente a Habermas. Genealogías de la razón*, Madrid: Tecnos.

Conill, J. (2022), "Razón pública e inteligencia artificial", en Pereira, G. y Pérez-Zafrilla, P. (eds.), *Actualidad de John Rawls en el siglo XXI*, Granada: Comares, 37-47.

Cortina, A. (1985), *Razón comunicativa y responsabilidad solidaria*, Salamanca: Sígueme.

Cortina, A. (1986), *Ética mínima*, Madrid: Tecnos; 2020 [18ª edición, corregida y aumentada].

Cortina, A. (1990), *Ética sin moral*, Madrid: Tecnos.

Cortina, A. (2007), *Ética de la razón cordial*, Oviedo: Nobel.

Cortina, A. (2017), *Aporofobia*, Barcelona: Paidós.

Cortina, A. (2019). Ética de la inteligencia artificial, *Anales de la Real Academia de Ciencias Morales y Políticas*, 96, 379-394. Sesión del 7 de mayo de 2019.

Cortina, A. (2021), *Ética cosmopolita*, Barcelona: Paidós.

Cortina, A. (2022a), La espiral del silencio y la presunta moralización de la sociedad. *Anales de la Real Academia de Ciencias Morales y Políticas*, 99, 419-430. Sesión del 31 de mayo de 2022.

Cortina, A. (2022b), Autocensura: destruyendo la democracia. *El País*, 8 de junio de 2022.

Floridi, L./Sanders, J.J. (2003). Internet Ethics: The Constructionist Values of Homo Poieticus. In R. Cavalier (ed.), *The Impact of the Internet on Our Moral Lives*. New York, SUNY,195-214.

Foucault, M. (1975), *Vigilar y castigar*, México: Siglo XXI.
García-Marza, D. (1992), *Ética de la justicia*, Madrid: Tecnos.
García-Marzá, D. (2012), Kant´s Principle of Publicity, *Kant-Studien*, 103 (1), 96-113.
García-Marzá, D. / Calvo, P. (2022), Democracia algorítmica: ¿un nuevo cambio estructural de la opinion pública?", *Isegoría*, 67, https://doi.org/10.3989/isegoria.2022.67.17
Gozálvez, Vicent (2012), *Ciudadanía mediática. Una mirada educativa*, Madrid: Dykinson.
Gracia, J. (2018), *El desafío ético de la educación*, Madrid: Dykinson.
Habermas, J. (1981). *Historia y crítica de la opinión pública*. Barcelona: Gustavo Gili, 2ª edición. Prólogo de Toni Domènech.
Habermas, J. (1985), *Conciencia moral y acción comunicativa*, Barcelona: Península.
Habermas, J. (1987). *Teoría de la acción comunicativa*. Madrid: Taurus.
Habermas, J. (2009), *Philosophische Texte*, Band 4, Studienausgabe, Frankfurt: Suhrkamp.
Habermas, J. (2006), *Entre naturalismo y religión*, Barcelona: Paidós.
Habermas, J. (2019), *Auch eine Geschichte der Philosophie*, Berlin: Suhrkamp, 2 vols.
Habermas, J. (2022), *Ein neuer Strukturwandel der Öffentlichkeit und deliberative Politik*, Berlin: Suhrkamp.
Kant, I. (1968). *Was ist Aufklärung?* Kants Werke. Akademie Textausgabe, VIII, Berlin: Walter de Gruyter.
Kant, I. (1981). *Crítica del Juicio*. Madrid: Espasa-Calpe.
Kant, I. (1983). *Crítica de la razón pura*. Madrid: Alfaguara, 2ª edición.
Kant, I. (1985). *La paz perpetua*. Madrid: Tecnos.
Levitsky, S. y Ziblatt, D. (2018). *Cómo mueren las democracia*. Buenos Aires: Ariel.
Maquiavelo, N. (1985), *El Príncipe*, Madrid: Cátedra.
Maquiavelo, N. (1987), *Discursos sobre la primera década de Tito Livio*, I. 25, Madrid: Alianza.
Marina, J.A. (2023), "¿Es de fiar la opinión pública?"
https://www.joseantoniomarina.net/categoria-blog/destacado/es-de-fiar-la-opinion-publica/
https://ethic.es/2023/01/es-de-fiar-la-opinion-publica-democracia/
Martínez-Navarro, E. (2022), La aportación de Rawls a la Ética de la razón cordial desde su teoría de los sentimientos morales, en Pereira, G. y Pérez-Zafrilla, P. (eds.), *Actualidad de John Rawls en el siglo XXI*, Granada: Comares, 75-92.
Mead, G.H. (1972), *Espíritu, persona y sociedad*, Buenos Aires: Paidós.
Moreno, J.A. (2021/2), "Una distopía digital: Alex Portland", *Acontecimiento*, 139, 3-5.
Munzel, F. (2012), *Kant's Conception of Pedagogy*, Evanston: Northwestern University Press.
Nicolás, J.A. (2019), Posverdad: cartografía de un fenómeno complejo. *Diálogo Filosófico*, 105, 302-340.
Nietzsche, F. (1986), *El Gay saber*, Madrid: Espasa-Calpe.
Noelle-Neumann, E. (1982), *La espiral del silencio. Opinión pública: nuestra piel social*. Barcelona: Paidós.
O'Neil, C. (2018), *Armas de destrucción matemática*, Capitán Swing.
Ortega, C. (2021), *Habermas ante el siglo XXI*, Madrid: Tecnos.
Ortega y Gasset, J. (2005), La rebelión de las masas, *Obras completas*, tomo IV, Madrid: Taurus.

Pérez-Zafrilla, P. (2021a), Polarización artificial: cómo los discursos expresivos inflaman la percepción de polarización política en internet. *RECERCA*, 26 (2), 1-23.
Pérez-Zafrilla, P. (2021b), "Bases neuroéticas de la corrección política. Una aproximación desde la teoría de la espiral del silencio de Elisabeth Noelle-Neumann", en López-Orellana, R. y Suárez-Ruiz, J. (eds.), *Filosofía posdarwiniana*, Londres: College Publications, 471-499.
Rawls, J. (1971/1972). *A Theory of Justice*. Oxford University Press.
Rawls, J. (1978). *Teoría de la justicia*. México: Fondo de Cultura Económica.
Rawls, J. (1993). *Political Liberalism*. New York: Columbia University Press.
Rawls, J. (1996). *El liberalismo político*. Barcelona: Crítica.
Rawls, J. (2001). *El derecho de gentes*. Barcelona: Paidós.
Rawls, J. (1999). *Collected Papers*. Harvard University Press.
Sen, A. (2010), *La idea de la justicia*, Madrid: Taurus.
Siurana, J.C. (2003), *Una brújula para la vida moral*, Granada: Comares.
Siurana, J.C. (2021), Ética para *influencers*, Madrid: Plaza y Valdés.

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 131-145
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.557391

Exigencias éticas para un periodismo responsable en el contexto de la inteligencia artificial*

Ethical demands for responsable journalism in the context of artificial inteligence

*ELSA GONZÁLEZ-ESTEBAN***
*ROSANA SANAHUJA-SANAHUJA****

Resumen. La irrupción de la utilización generalizada de inteligencia artificial en el ámbito de la comunicación y en concreto en el periodismo muestra un resultado claroscuro que cabe analizar desde una perspectiva crítica. Este artículo aborda desde un perspectiva crítica la revolución que la presencia creciente de la inteligencia artificial está provocando tanto los métodos como en los resultados periodísticos, afectando a sus garan-

Abstract. The irruption of the widespread use of artificial intelligence in the field of communication and specifically in journalism shows a chiaroscuro result that should be analyzed from a critical perspective. The paper takes a critical look at the revolution that the growing presence of artificial intelligence is causing in journalistic methods and results, affecting its guarantees of quality and excellence. The text argues that

Recibido: 14/02/2023. Aceptado: 04/06/2023.

* Este estudio se inserta en el Proyecto de Investigación Científica y Desarrollo "Ética aplicada y confiabilidad para una Inteligencia Artificial" PID2019-109078RB-C21 financiado por MCIN/ AEI /10.13039/501100011033, y en las actividades del grupo de investigación de excelencia PROMETEO CIPROM/2021/072, financiado por Conselleria d'Innovació, Universitats, Ciència i Societat Digital de la Generalitat Valenciana. Queremos agradecer a colegas del grupo de investigación de Éticas Aplicadas y Democracia de la Universidad de Valencia y de Ética Práctica y Democracia de la Universitat Jaume I la discusión del texto así como sus recomendaciones. Así mismo a los/las dos revisores por sus aportaciones que han enriquecido el texto.

** Profesora titular de Filosofía Moral (ética) en el Departamento de Filosofía y Sociología de la Universitat Jaume I. Su labor investigadora se centra en el campo de la ética aplicada y los sistemas de gobernanza y gestión ética en las organizaciones. Ha publicado recientemente "Investigación e innovación en inteligencia artificial: responsabilidad y confianza" en la obra colectiva que coedita junto a Juan Carlos Siurana, *Inteligencia Artificial. Concepto, alcance, retos* (Tirant lo Blanch, 2023) y el capítulo "The ETHNA System and Support Tools" en el libro en abierto Ethics *and Responsible Research and Innovation in Practice* (Springer, 2023) del que es coeditora junto a Ramón Feenstra y Luis Camarinha-Matos. Contacto: esteban@uji.es

*** Profesora ayudante doctora en el Departamento de Ciencias de la Comunicación de la Universitat Jaume I. Entre sus líneas de investigación se encuentran tendencias de innovación en periodismo, inteligencia artificial, éticas de la comunicación y comunicación de la ciencia. Ha publicado recientemente: "Ética y uso periodístico de la inteligencia artificial. Los medios públicos y las plataformas de verificación como precursores de la rendición de cuentas en España" (*Estudios sobre el mensaje periodistico*. Num. 4. Vol. 28. pp. 959-970. 2022) y "Ámbitos de aplicación periodística de la Inteligencia Artificial. Mapa conceptual, funciones profesionales y tendencias en desarrollo en el contexto de la pandemia global de la Covid-19" (*Razón y palabra*. Num. 112. Vol. 25. pp. 432-449. 2022). Contacto: rosana.sanahuja@uji.es

tías de calidad y excelencia. El texto sostiene que es necesario acercarse críticamente al impacto actual, así como al potencial, que este presenta para los profesionales, las organizaciones y la sociedad, y propone hacerlo desde el método hermenéutico-crítico de las éticas aplicadas. El estudio se estructura en dos partes. Una primera que realiza una aproximación a la irrupción de la inteligencia artificial en la práctica del periodismo, explicitando las recomendaciones que desde diferentes instancias se ofrecen para orientar esta incorporación de la inteligencia artificial en la práctica comunicativa, concretamente en la periodística. Una segunda parte que identifica los principales riesgos éticos, recomendaciones y principios éticos para afrontarlos desde la autorregulación ética.
Palabras clave: Ética, Inteligencia artificial, Periodismo, Riesgos éticos, Autorregulación.

it is necessary to critically approach the current impact, as well as the potential, that this presents for professionals, organizations and society, and proposes to do so from the hermeneutic-critical method of applied ethics. The study is structured in two parts. The first part takes an approach to the irruption of artificial intelligence in the practice of journalism, explaining the recommendations offered by different bodies to guide the incorporation of artificial intelligence in the practice of communication, specifically in journalism. The second part identifies the main ethical risks, recommendations and ethical principles for dealing with them through ethical self-regulation.
Keywords: Ethics, Artificial intelligence, Journalism, Ethical risks, Self-regulation.

1. Una mirada crítica a la irrupción de la inteligencia artificial en el periodismo

La irrupción de la utilización generalizada de inteligencia artificial en el ámbito de la comunicación y en concreto en el periodismo muestra un resultado claroscuro que cabe analizar desde una perspectiva crítica. Su presencia y utilización creciente en la actividad periodística está revolucionando la velocidad y los métodos en las redacciones y comprometiendo las garantías de la calidad o excelencia del periodismo. Es necesario acercarse críticamente al impacto actual, así como al potencial, que esta realidad presenta para los profesionales, las organizaciones y la sociedad.

Este trabajo pretende realizar tal tarea utilizando el método hermenéutico-crítico propuesto desde las filas de la ética aplicada discursiva por Cortina (1996) y que ha sido empleado con éxito en otros ámbitos como el empresarial, el económico, el sanitario, el educativo, el desarrollo de los pueblos y el comunicativo, entre otros (García-Marzá et al., 2018; González-Esteban et al., 2019).

La hermenéutica-crítica busca explicitar la intersubjetividad ética que se encuentra en toda actividad humana. Se fundamenta en el principio ético discursivo que salvaguarda los conceptos de *persona* e *igualdad* a la hora de desentrañar qué es lo que consideramos como correcto o justo en las actividades humanas. Tomando este punto de partida crítico, Cortina argumenta que la tarea de la ética aplicada no le corresponde a los parlamentos sino a los afectados por la actividad. Atañe por tanto a esos interlocutores válidos que deben ser reconocidos por igual valía y en condiciones lo más cercanas a la simetría para que participen en aquellos diálogos y configuración de discursos en los que se tratan aquellos temas, o se discute sobre normas, que le afectan o le pueden afectar. Diálogos que tengan como objeto dilucidar qué es lo que se tiene por justo o correcto, en dicha actividad, porque humaniza (Cortina, 1996, 2007). En nuestro caso queremos centrar la mirada sobre la actividad

comunicativa en el periodismo y de qué modo su legitimidad y excelencia se ven, o pueden ver, alteradas por la utilización de la inteligencia artificial. El objetivo último es conocer críticamente la situación que está viviendo el periodismo en el contexto de la inteligencia artificial desde la perspectiva ético-discursiva.

El periodismo como actividad comunicativa cobra legitimidad socialmente por colaborar a formar una opinión crítica madura, la meta de su actividad. Y, cuenta con una serie de virtudes y valores asociados a esa meta que generan un hábito y un carácter concreto en la práctica periodística que requieren ser cultivados por la profesión y las organizaciones. Los códigos deontológicos y profesionales de la actividad periodísticas reflejan tanto la meta como los valores y virtudes que se consideran intersubjetivamente necesarios encarnar y exigir de un periodismo que atiende a la conciencia moral cívica que ha alcanzado la sociedad en la que se desarrolla (Bilbeny, 2012; Conill et al., 2004; Meyers, 2010). En líneas generales los valores y virtudes asociadas a la práctica periodística están ligados a la veracidad, la búsqueda de la objetividad (entendida como intersubjetividad), la responsabilidad, la libertad de opinión y crítica y al respeto a la intimidad, integridad y dignidad de las personas, entre otros.

Tanto la meta como algunos de los valores y virtudes mencionados se encuentran en la actualidad en entredicho ante el uso cuestionable que se está realizando de la inteligencia artificial en la práctica periodística. Por ese motivo, se hace necesario, a nuestro juicio, en primer lugar, comprender de qué modo se está utilizando y para qué la inteligencia artificial en el ámbito y la práctica del periodismo. En segundo lugar, siguiendo el método hermenéutico-crítico, es preciso explicitar los riesgos éticos a los que está sometida la actividad periodística y que pueden hacerla caer en el descrédito y la falta de legitimidad social. En tercer lugar, se deben conocer y analizar las recomendaciones que desde diferentes instancias se están proponiendo para orientar esta incorporación de la inteligencia artificial en la práctica comunicativa y concretamente en la periodística, de modo que se consideren las expectativas sociales y las exigencias éticas. Y, finalmente, cabe sacar a la luz y profundizar en los principios éticos que se están proponiendo para orientar éticamente su autorregulación. Principios éticos que son la expresión de las exigencias éticas que la sociedad esgrime para hacer valer y respetar el contrato moral (García-Marzá, 2004) que posee con respecto al bien social que provee el periodismo: la información contrastada y de calidad en sociedades democráticas, de modo que se colabore en formar una opinión pública madura.

2. Presencia creciente de la inteligencia artificial en el periodismo

Periodismo robot, periodismo algorítmico, periodismo computacional, periodismo automatizado y periodismo artificial son algunos de los nombres que se ha venido dando al uso de la inteligencia artificial en el ámbito periodístico. Estas y otras denominaciones se han utilizado para hacer referencia al proceso informativo apoyado de una u otra forma por esta tecnología (L.-M. Calvo-Rubio y Ufarte-Ruiz, 2021; Túñez-López et al., 2021; Ufarte Ruiz y Manfredi Sánchez, 2019). Aún a falta de un consenso conceptual, lo que resulta evidente es que la inteligencia artificial, incluida por Salaverría (2015) en un conjunto más amplio de altas tecnologías ligadas al periodismo *hi-tech*, es una realidad ya consolidada en las

redacciones españolas. A mediados de la segunda década del siglo XXI empezaron a surgir en España las consideradas como iniciativas pioneras en el uso de la inteligencia artificial como el bot Politibot creado en Telegram para cubrir las elecciones españolas de 2016, el bot puesto en marcha por *El País* en Facebook Messenger para informar sobre las elecciones presidenciales de Francia en 2017; el proyecto Medusa del grupo Vocento de generación automatizada de contenidos sobre la situación de playas y estaciones de esquí o el bot de *El Confidencial* AnaFut para la cobertura de los partidos de la Segunda División B (Ufarte Ruiz y Manfredi Sánchez, 2019). Apenas un lustro después, el uso de la inteligencia artificial en el periodismo ha dado un giro radical integrándose ya de forma habitual en la mayor parte de las redacciones. Un buen ejemplo de este crecimiento exponencial es la *startup* española Narrativa que en 2015 creaba el software Gabriel de procesamiento de lenguaje natural destinado, entre otros usos, a la redacción de piezas periodísticas. Narrativa es actualmente una empresa consolidada como referente a nivel internacional que utiliza herramientas de extracción y análisis de datos y de procesamiento y generación de lenguaje natural para la creación de contenido inteligente, automatización de informes y optimización de procesos en todo tipo de sectores. Entre sus últimas acciones en el campo del periodismo se encuentra el Proyecto de Seguimiento COVID-19[1], impulsado con la colaboración de Radio Televisión Española, que desde el principio de la pandemia ha generado textos en español, inglés e italiano a partir de datos actualizados facilitados por múltiples fuentes oficiales.

El uso de la inteligencia artificial en el ámbito de la comunicación se ha «convertido rápidamente en una parte fundamental de las operaciones modernas a todo nivel» (Newman, 2022). Según el informe de tendencias del Instituto Reuters para 2022, más de un 80% de los directivos y responsables de empresas de la comunicación considera que ya en este año la inteligencia artificial va a ser clave para la personalización y las recomendaciones de contenidos, así como para la automatización de los flujos de trabajo en las redacciones, por ejemplo, el etiquetado de contenidos y la transcripción de entrevistas. Alrededor de un 70% le atribuyen un papel clave para la búsqueda previa de información y un 40% para la redacción automática de artículos. Estos datos coinciden con las tendencias detectadas en el uso actual de la inteligencia artificial por parte de los medios de comunicación en Cataluña, donde un 76,2% de los medios analizados por Ventura (2021) afirma utilizar ya inteligencia artificial u otros sistemas algorítmicos en sus procesos informativos. De ellos, el 88,2% lo emplean en la selección de contenidos, la detección de tendencias o la elección del ángulo y alrededor de la mitad lo aplican a la recopilación de información y la creación y distribución de contenido.

Las razones de este auge se deben a la posibilidad de agilizar, simplificar y hacer más eficaces los procesos de producción que supone la inteligencia artificial para el periodismo (López-García y Vizoso, 2021). Así, la profesión periodística atribuye interesantes oportunidades al uso de estas tecnologías, como el potencial del procesamiento de datos para conocer a las audiencias y adaptar el producto, la eficiencia en la gestión de procesos internos o el apoyo en las búsquedas y la generación automatizada de contenidos (Ventura, 2021). Otras de las razones argumentadas para su uso son la búsqueda de mayor precisión, el aumento en la producción, la objetividad, la capacidad de agregar contenidos web, la personalización de informaciones, la identificación de eventos de interés periodístico para su posterior difusión

1 Ver en: https://covid19tracking.narrativa.com/

y la lucha contra la desinformación (L.-M. Calvo-Rubio y Ufarte-Ruiz, 2021). Ahora bien, cabe una mirada crítica ante las bondades atribuidas al uso de la inteligencia artificial en tales prácticas y contextos.

3. Riesgos éticos del uso de la inteligencia artificial en la práctica periodística

La inteligencia artificial afecta a tres esferas del periodismo: la organizacional, la profesional y la social; y lo hace además durante las tres etapas del proceso periodístico: recogida de datos, elaboración y difusión (Dörr y Hollnbuchner, 2017). La influencia en estas esferas y etapas es creciente y comporta claras oportunidades pero también importantes retos y riesgos. El informe pionero sobre los efectos de la inteligencia artificial en el periodismo lanzado por la plataforma internacional Journalism AI advertía en 2019 que el nuevo poder que supone el empleo de la inteligencia artificial en el periodismo conlleva también nuevas responsabilidad ya que esta tecnología tiene el potencial de influir de forma profunda en la forma de hacer periodismo y de consumir la información, generando cambios estructurales; de ahí la necesidad de que los medios de comunicación sean capaces de poner estas herramientas al servicio de sus valores y criterios editoriales (Beckett, 2019). Así como de explicarlos y ponerlos en relación con las expectativas éticas de sus grupos de interés o *stakeholders*, como apunta una ética empresarial de corte discursivo-crítico (González-Esteban, 2019).

A pesar de resultar evidente para muchas voces la necesidad de reflexionar sobre estas nuevas responsabilidades y riesgos, los estudios desarrollados hasta la fecha en el ámbito de la inteligencia artificial y el periodismo recogen de forma minoritaria las reflexiones éticas sobre las posibles consecuencias de estas tecnologías en el periodismo (Sanahuja-Sanahuja, 2022a,; L.-M. Calvo-Rubio y Ufarte-Ruiz, 2021). Desde 2015 se registra un crecimiento continuado de las publicaciones académicas sobre el uso de la inteligencia artificial en el periodismo (Sanahuja-Sanahuja, 2022b) pero es en los últimos tiempos cuando en el caso de España están proliferando diversos estudios sobre el tema centrados en su perspectiva ética y en los que se reivindica la necesidad de adoptar medidas para hacer frente a los riesgos de la inteligencia artificial en la práctica periodística. El Consell de la Informació de Catalunya ha publicado recientemente un informe con recomendaciones para dotar a la inteligencia artificial de los valores éticos del periodismo (Ventura, 2021), mientras estudios académicos destacan la necesidad de revisar el Código Deontológico de la Federación de Asociaciones de Periodistas de España (FAPE) para adaptarlo al periodismo automatizado, sobre todo en materia de autoría, elaboración, transparencia y jerarquización de las informaciones redactadas a través de inteligencia artificial (Ufarte Ruíz et al., 2021). En 2021 Túñez-López publicaba asimismo un estudio sobre el impacto de la inteligencia artificial en la comunicación alertando de que se trata de un «sistema diseñado para generar esferas de control social a través de productos o mensajes singularizados ajustados a necesidades detectadas por el procesamiento de datos masivos» (Túñez López, 2021, p. 8).

Por su parte, los expertos consultados por Ruíz Ufarte et al (2021), a través de un estudio Delphi con profesorado de Periodismo del ámbito de la ética y la deontología, concluye que los retos éticos en los que existe un mayor grado de consenso pasan por garantizar la intimidad y privacidad de las personas; potenciar la ética del periodista y del

uso de la tecnología; el contraste por parte de los periodistas de la información producida automáticamente; la formación de los profesionales de la información y la transparencia. La supervisión humana y la aplicación de un criterio editorial junto al tratamiento responsable de los datos de los usuarios son asimismo algunos de los principales retos detectados por Ventura (2021) a partir de entrevistas y grupos de discusión con expertos y responsables de medios de comunicación. Otros temas claves detectados en el estudio pasan por los riesgos de la personalización de contenidos, la supervisión y calidad de los datos para evitar sesgos o el riesgo de la independencia periodística ante la financiación de las grandes plataformas tecnológicas a la hora de desarrollar tanto la tecnología vinculada a la inteligencia artificial como la formación sobre la misma (Ventura, 2021). La revisión literaria realizada por Pérez-Seijo et al. (2020), recoge también algunos de los principales retos del periodismo que pasarían por los conflictos que comprometen la ética profesional del periodista y del periodismo, las cuestiones legales que dificultan la atribución de responsabilidades y de los derechos de autor, la percepción de los usuarios de la credibilidad y objetividad informativa cuando los contenidos son fruto de una mediación tecnológica, el rol del profesional de la información cuando una máquina sustituye su papel como mediador, la aproximación hacia un periodismo de corte más emocional que puede abrir la puerta al sesgo, y el elevado coste económico que la introducción de la más alta tecnología actual representa para las redacciones con menos recursos (Pérez-Seijo et al., 2020, p. 144).

A nivel tanto nacional como internacional diversas investigaciones han reflexionado a lo largo de los últimos años sobre el tema desde posturas críticas sobre los posibles efectos (Sanahuja-Sanahuja, 2022a), alertando de aspectos como la necesidad de transparencia en su uso (Diakopoulos, 2019a; Diakopoulos y Koliska, 2017; Trattner et al., 2021); la necesidad de una mayor regulación ante las posibilidades que abre la aplicación de la inteligencia artificial en el ámbito de la comunicación (Jina, 2019; Lie, 2021; Túñez López, 2021); la atribución de responsabilidades (Chen y Wen, 2021; Lewis et al., 2018); los riesgos de los recomendadores de noticias y su relación con los valores de los periodistas y de los medios (Salazar García, 2018); la forma en que la interrelación entre hombres y máquinas afectarán al periodismo y al ejercicio de su profesión (Lewis et al., 2018); la percepción de los humano por parte de la audiencia (Gonzales, 2017; Shin, 2021); la repercusión en la profesión periodística (Kim et al., 2020); la necesidad de formación de los nuevos periodistas ante las nuevas competencias que demanda (Salnikova, 2019; Ufarte Ruiz, Calvo Rubio, et al., 2020; Ufarte Ruiz, Fieiras-Ceide, et al., 2020); la necesidad de una educación ciudadana a todos los niveles para detectar el uso de automatismos (L. M. Calvo-Rubio y Ufarte-Ruiz, 2020; Túñez López, 2021); y los efectos de la tecnología en la difusión de desinformación pero también el potencial de su uso como herramienta para combatirla (Flores Vivar, 2019; Grmuša y Prelog, 2020; Manfredi Sánchez y Ufarte Ruiz, 2020; Túñez López, 2021).

4. La autorregulación como recomendación

La tecnología no es buena ni mala ni tampoco neutral afirma Kranzberg (1986) en sus leyes sobre la tecnología, en las que defiende asimismo cómo, aunque la tecnología puede ser un elemento primordial en muchos asuntos públicos, los factores no técnicos han de

tener prioridad en las decisiones de política tecnológica. Si coincidimos con Helberger et al. (2019) en que el periodismo desempeña un papel crucial en las democracias, ya que proporciona al público una fuente de información, una plataforma para la deliberación y una vigilancia crítica; resultará claramente necesario que sean factores no técnicos guiados por la ética y los valores del periodismo los que determinen el uso de la inteligencia artificial en las esferas y etapas del proceso informativo. La autorregulación a través de códigos deontológicos, libros de estilo, consejos de prensa o defensores de la audiencia es uno de los principales mecanismos de control ético de la profesión periodística (Ufarte Ruíz et al., 2021). De este modo, repensar a través de estos marcos el uso de la inteligencia artificial en los procesos periodísticos y el modo en el que los medios y los profesionales del periodismo se autorregulan, resultan aspectos clave. Es necesario plantear una hoja de ruta que permita buscar más puntos de convergencia en favor de un mejor periodismo para hacer frente a los riesgos profesionales y deontológicos de estas tecnologías (Murcia Verdú y Ufarte Ruiz, 2019; Parratt-Fernández et al., 2021) e incluso que fije unas «directrices claras sobre qué se puede y qué no automatizar» (Ufarte Ruíz et al., 2021, p. 679). Esto posibilitaría avanzar hacia un pacto comunicacional que preserve el bien interno del periodismo que no es otro que aportar información que favorezca la formación de una opinión pública madura en sociedades democráticas, de tal modo que el ciudadano tenga acceso a aquella información que tiene derecho a saber. En palabras de Cortina, «un periodismo comprometido con el objetivo de ayudar a promover una sociedad informada y abierta» necesario para construir y consolidar la democracia (Cortina, 2021, p. 17).

De cara a una autorregulación por parte de los medios a nivel nacional, por el momento, las recomendaciones planteadas por el Consell de la Informació de Catalunya para dotar a la inteligencia artificial de los valores éticos del periodismo parecen un primer paso en la línea de buscar una hoja de ruta a nivel práctico. En concreto, a partir del proceso de reflexión colectiva desarrollado, el informe establece ocho recomendaciones para los medios: (1) velar por la calidad de los datos y la gestión responsable de los mismos, manteniendo una vigilancia constante sobre su representatividad; (2) supervisar los procesos, asegurando su calidad técnica para minimizar los riesgos y evitar los errores; (3) transparencia y rendición de cuentas; (4) gestionar responsablemente los datos y la privacidad, recogiendo los datos personales estrictamente necesarios, anonimizándolos si no son relevantes y preservándolos de un mal uso por parte de terceros; (5) gestionar de forma responsable las personalizaciones y recomendaciones, evitando un uso de algoritmos que socave el pluralismo o perjudique a las personas vulnerables; (6) poner en valor el factor humano, recordando que es el profesional el que tiene el talante ético que no tiene la máquina; (7) impulsar la formación y la promoción de la interdisciplinariedad de los equipos con el fin de alcanzar una capacitación técnica y ética, y (8) promover la investigación encaminada a explorar la convergencia entre la eficacia técnica y los valore de un periodismo ético (Ventura, 2021).

Las recomendaciones planteadas recogen así algunos aspectos claves para hacer frente a los retos éticos de la inteligencia artificial y el periodismo en los que parece existir un claro consenso como es la necesidad de transparencia y explicabilidad (Diakopoulos y Koliska, 2017; Ufarte Ruíz et al., 2021) ya que los lectores tienen derecho a entender en términos comprensibles como es usada la inteligencia artificial y las decisiones que se toman a partir de las mismas (Hansen et al., 2017). Este punto no está exento a su vez de retos, incluida

la necesidad de determinar de qué automatizaciones es necesario informar al usuario ya que, como ejemplifica Ventura (2021, p. 35), igual no es necesario informar de que se ha utilizado inteligencia artificial para realizar una transcripción pero sí cuando el texto se ha generado de forma automática o se ha realizado una recomendación a partir de algoritmos. Otras dificultades se encuentran en la propia opacidad con la que pueden llegar a funcionar los algoritmos. Descampe et al. (2021) alertan además de que la transparencia por sí sola no es suficiente ya que hay que tener en cuenta la necesidad de reforzar desde la perspectiva ética los requisitos técnicos para hacer frente a problemas técnicos de aprendizaje automático, sesgos y manipulación interesada de datos. En esta misma línea (Diakopoulos, 2019b) reivindica la necesidad de la orientación ética del diseño tecnológico para la producción automatizada de noticias, resaltando la necesidad de métricas para medir la alineación de las implementaciones técnicas con los objetivos y valores de la organización.

El caso de la BBC puede ser un buen ejemplo de rendición de cuentas unido a una voluntad de establecer un diseño tecnológico responsable alineados con los valores del medio. La corporación británica de radiotelevisión presentó en 2017 a la Cámara de los Lores un escrito[2] en el que plasma su compromiso para liderar el uso responsable de todas las tecnologías de inteligencia artificial, para lo que establecía una serie de principios dirigidos a que los motores de aprendizaje automático de la cadena reflejen los valores de la organización. En cuanto a la audiencia, el texto establece que los datos recogidos se utilizan para mejorar sus experiencias con la BBC y recoge el derecho de los usuarios a saber qué hacen con sus datos por lo que se comprometen a explicar en un lenguaje sencillo qué datos recogen y cómo los utilizan, por ejemplo, en la personalización y las recomendaciones. Estos principios muestran además su apuesta por un desarrollo responsable de la tecnología con el fin de que sus algoritmos sirvan a sus audiencias de forma equitativa y justa y que la difusión de contenidos se ajuste a los valores editoriales de la BBC, tratando de ampliar, en lugar de reducir, los horizontes de la audiencia.

Poner las tecnologías emergentes al servicio de los valores que rigen el periodismo de calidad puede ser también una oportunidad para los medios de comunicación, de forma que un uso responsable y ético de la inteligencia artificial contribuya a reforzar los principios de verdad, justicia, libertad y responsabilidad que rigen la profesión periodística (Ventura 2021). En unos momentos en los que la inteligencia artificial se está consolidando a pasos de gigante en las redacciones, periodistas y medios de comunicación tendrían que jugar un papel clave para decidir el rumbo a tomar en esta implementación. Sin embargo, un estudio reciente de De Haan et al. (2022) alerta de la postura «sorprendentemente pasiva» que adoptan los periodistas hacia la introducción de la lógica algorítmica en su proceso de producción predominando la sensación de que es suficiente con su autonomía profesional para combatir la influencia de algoritmos y otros sesgos tecnológicos, lo que lleva a que aprender más sobre la inteligencia artificial y el proceso periodístico no parezca ser una prioridad ni un aspecto que se facilite a nivel de dirección. No obstante, para hacer frente de forma responsable a los retos que supone la inteligencia artificial, tanto medios como expertos destacan la necesidad de formación y de incorporar nuevos perfiles profesionales (Ufarte Ruíz et al., 2021;

2 BBC – Written evidence (AIC0204): http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/70493.pdf%20

Ventura, 2021). La inversión que supone la inteligencia artificial, tanto en tiempo -incluido el de formación- como en dinero, se percibe como un desafío que puede aumentar aún más la brecha ya existente entre grandes grupos mediáticos y medios más pequeños y locales, según advierte un estudio publicado por Associated Press (Rinehart y Kung, 2022). Contar con la apuesta clara de la dirección, recursos y formación se perfilan como aspectos claves para avanzar hacia una introducción responsable de la inteligencia artificial en las redacciones y promover una necesaria autorregulación por parte de los medios.

5. Principios éticos ante los retos que plantea el uso de inteligencia artificial en el periodismo

Los riesgos éticos identificados y las recomendaciones elaboradas desde diferentes instancias profesionales, académicas y de la sociedad civil permiten evidenciar expectativas y demandas éticas específicas y relativas al uso e incorporación de la inteligencia artificial en la práctica comunicativa. Podría hablarse, como se ha argumentado en el apartado anterior, de una interpretación y visión crítica del uso de la inteligencia artificial a la luz de la meta o bien interno de la práctica periodística, así como de los valores éticos reconocidos por la profesión y la sociedad.

Llevar a cabo esta interpretación y visión crítica en el día a día, no es tarea fácil. Así autores como Túñez López (2021) abordan la necesidad y la dificultad de la gobernanza de la inteligencia artificial en el ámbito de la comunicación, recogiendo las reflexiones de Robles Carrillo (2020) sobre cómo la gobernanza de la inteligencia artificial constituye, posiblemente, uno de los mayores desafíos que se ha planteado a la ciencia y a la técnica jurídicas ya que funcionalmente, la inteligencia artificial no es una categoría estanca, sino permeable, transversal, porque actúa sobre los elementos y realidades preexistentes afectando e interfiriendo en su desarrollo y funcionamiento. Por estas razones, la autora argumenta que no es posible proceder a la gestión y regulación de la inteligencia artificial exclusivamente desde postulados previos «creados para una sociedad en la que no existía un desarrollo científico y tecnológico con ese alcance, contenido y naturaleza» (Robles Carrillo, 2020, p. 25). La complementariedad con la autorregulación ética se hace pues necesaria, en entornos globales, interdependientes y donde el pluralismo moral está presente (González-Esteban, 2022).

Como ya se ha mencionado, en el último lustro se ha incrementado enormemente el interés y los esfuerzos desde el ámbito académico, político y organizativo, nacional e internacional, por desentrañar colectivamente de qué modo queremos como sociedades que sea utilizada la inteligencia artificial. Una mirada a estos avances puede ser de ayuda para identificar los principios éticos que se pueden utilizar en la autorregulación de la actividad periodística ante el uso de la inteligencia artificial.

Son numerosas las propuestas de guías, directrices o marcos para generar una inteligencia artificial o para hacer uso de la misma de modo que se considere socialmente aceptable y éticamente deseable (Hagendorff, 2020; Jobin et al., 2019). Iniciativas que proceden en ocasiones de organismos político-gubernamentales, como la Unión Europea (Comisión Europea, 2021) o las Naciones Unidas (UNESCO, 2021). Y, en otras ocasiones, del sector privado y/o de la sociedad civil, donde se suele invocar a la necesidad de

complementar la regulación de los estados mediante mecanismos jurídico-políticos con la autorregulación basada en principios éticos que permitan generar códigos éticos y de conducta, ser utilizados por comités éticos especializados en inteligencia artificial o para la identificación de buenas prácticas. Una comparativa de los principios que se proponen en estos documentos internacionales desarrollada por Jobin, Ienca y Veyana (2019) identifica 12 principios fundamentales: justicia y bienestar; no-maleficencia; responsabilidad; privacidad; beneficencia; libertad y autonomía; confianza; sostenibilidad; dignidad; solidaridad; transparencia y; explicabilidad y rendición de cuentas. Como apunta Cortina se trata de once principios tradicionalmente utilizados en diferentes contextos y prácticas humanas desde la ética aplicada, y uno original: el de explicabilidad y rendición de cuentas, que merecería una atención especial (Cortina, 2019, p. 388).

Este principio ético de explicabilidad y transparencia constituyen, junto al de rendición de cuentas y supervisión, los dos pilares argumentativos en la identificación de los riesgos éticos, así como en las recomendaciones del uso e integración de la inteligencia artificial en los contextos comunicativos. Por este motivo conviene una comprensión más profunda de los mismos, así como una interpretación adecuada de sus implicaciones.

El principio de explicabilidad y transparencia exige que los afectados puedan conocer qué algoritmos se han utilizado, de dónde proceden, quién los ha construido, con qué sesgos y con qué finalidades. Con este principio se sitúa la responsabilidad en las organizaciones y los profesionales que diseñan y utilizan la inteligencia artificial en la práctica periodística. La confiabilidad que se exige la inteligencia artificial es importante no confundirla con una cualidad antropomórfica de la tecnología (Ryan, 2020). Podremos confiar en la inteligencia artificial porque confiamos en la información que se nos proporciona sobre cómo y para qué se ha utilizado. Información que debemos recibir de modo inteligible, veraz y comprensible. Donde el ser humano como agente queda identificado. Así pues, es fundamental que quede claro quién ha elaborado, recopilado o distribuido la información, si es una persona o si es un sistema de inteligencia artificial. Sólo de ese modo podremos juzgar como afectados e interlocutores válidos si aceptamos o no ese uso y práctica periodística. Con este principio se exige que la autonomía y la dignidad de las personas sea respetada en todo momento reconociéndolas como interlocutores válidos que pueden ejercer su libertad.

El principio de rendición de cuentas y supervisión orienta las actuaciones para que en todo momento sea posible rastrear al agente que ha desarrollado y aplicado la inteligencia artificial, y por tanto, al sujeto al que cabe pedir responsabilidad, aunque ese sea un actor múltiple. Ellos son los responsables de las acciones que se generan o de los efectos —positivos o negativos— que produce el uso de la inteligencia artificial. En el caso del periodismo, los sujetos de responsabilidad deben quedar identificados en el medio de comunicación, de forma que no se diluya. Por otra parte, para que la rendición de cuentas sea adecuada debe existir un sistema de supervisión basado en una continua evaluación de los riesgos éticos, presentes o potenciales, junto con un sistema para garantizar que los diferentes *stakeholders* puedan informar de sus preocupaciones. De nuevo son los valores de la autonomía, dignidad y libertad los que se pretenden exigir al impulsar este principio ético en la práctica periodística en el contexto de la inteligencia artificial.

Como se ha venido señalando el uso de la inteligencia artificial en el contexto de la práctica comunicativa y periodística puede ser beneficioso siempre y cuando el progreso técnico que ofrece no menoscabe el progreso ético que es capaz de ofrecernos la actividad periodística a través de su bien interno: una información que nos permita ir formando en sociedades moralmente pluralistas con una opinión pública madura, desde la salvaguarda y exigencia de la autonomía, la libertad y la dignidad que son propiamente humanas. Para ello, es necesario seguir profundizando en cómo desplegar los principios de explicabilidad y rendición de cuentas en el terreno periodístico para avanzar en una autorregulación capaz de dar respuesta de las exigencias éticas que se erigen por parte de los afectados.

Referencias

Beckett, C. (2019). *New powers, new responsibilities. A global survey of journalism and artificial intelligence | | Polis*. https://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities/

Bilbeny, N. (2012). *Ética del periodismo : la defensa del interés público por medio de una información libre, veraz y justa* [Book]. Publicacions i Edicions de la Universitat de Barcelona.

Calvo-Rubio, L.-M., & Ufarte-Ruiz, M.-J. (2021). Artificial intelligence and journalism: Systematic review of scientific production in Web of Science and Scopus (2008-2019). *Communication & Society*, *34*(2), 159–176. https://doi.org/10.15581/003.34.2.159-176

Calvo-Rubio, L. M., & Ufarte-Ruiz, M. J. (2020). Perception of teachers, students, innovation managers and journalists about the use of artificial intelligence in journalism. *Profesional de La Informacion*, *29*(1). https://doi.org/10.3145/epi.2020.ene.09

Chen, Y. N. K., & Wen, C. H. R. (2021). Impacts of Attitudes Toward Government and Corporations on Public Trust in Artificial Intelligence. *Communication Studies*, *72*(1), 115–131. https://doi.org/10.1080/10510974.2020.1807380

Comisión Europea, U. . (2021). Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de IA (Ley de IA) y se modifican determinados actos legislativos de la Unión. In *Diario Oficial de las Comunidades Europeas: Vol. COM 206*.

Conill, J., Gozálvez Pérez, V. E., & Camps, V. (2004). *Ética de los medios : una apuesta por la ciudadanía audiovisual*. Gedisa.

Cortina, A. (1996). El estatuto de la ética aplicada. Hermenéutica crítica de las actividades humanas. *Isegoría: Revista de Filosofía Moral y Politica*, *13*, 119–134. https://doi.org/https://doi.org/10.3989/isegoria.1996.i13

Cortina, A. (2007). Ethica cordis. *Isegoría: Revista de Filosofía Moral y Política*, *37*, 113–126. https://doi.org/10.3989/isegoria.2007.i37.112

Cortina, A. (2019). Ética de la inteligencia artificial. In *Anales de la Real Academia de Ciencias Morales y Políticas* (pp. 379–394). Real Academia de Ciencias Morales y Políticas.

Cortina, A. (2021). Periodismo ético en tiempos de polarización. *Cuadernos de Periodistas*, *43*, 9–18. https://www.cuadernosdeperiodistas.com/periodismo-etico-en-tiempos-de-polarizacion/

de Haan, Y., van den Berg, E., Goutier, N., Kruikemeier, S., & Lecheler, S. (2022). Invisible Friend or Foe?How Journalists Use and Perceive Algorithmic-Driven Tools in Their Research Process. *Https://Doi.Org/10.1080/21670811.2022.2027798*. https://doi.org/10.1080/21670811.2022.2027798

Descampe, A., Massart, C., Poelman, S., Standaert, F. X., & Standaert, O. (2021). Automated news recommendation in front of adversarial examples and the technical limits of transparency in algorithmic accountability. *AI and Society*. https://doi.org/10.1007/s00146-021-01159-3

Diakopoulos, N. (2019a). *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.

Diakopoulos, N. (2019b). Towards a Design Orientation on Algorithms and Automation in News Production. *Digital Journalism*, *7*(8), 1180–1184. https://doi.org/10.1080/21670811.2019.1682938

Diakopoulos, N., & Koliska, M. (2017). Algorithmic Transparency in the News Media. *Digital Journalism*, *5*(7), 809–828. https://doi.org/10.1080/21670811.2016.1208053

Dörr, K. N., & Hollnbuchner, K. (2017). Ethical Challenges of Algorithmic Journalism. *Digital Journalism*, *5*(4), 404–419. https://doi.org/10.1080/21670811.2016.1167612

Flores Vivar, J. M. (2019). Artificial intelligence and journalism: diluting the impact of disinformation and fake news through bots. *Doxa Comunicación. Revista Interdisciplinar de Estudios de Comunicación y Ciencias Sociales*, *29*, 197–212. https://doi.org/10.31921/doxacom.n29a10

García-Marzá, D., Lozano Aguilar, J. F., Martínez Navarro, E., & Siurana Aparici, J. C. (Eds.). (2018). *Ética y política. Homenaje a Adela Cortina*. Tecnos.

Gonzales, H. M. S. (2017). Bots as a news service and its emotional connection with audiences. The case of Politibot | Publons. *DOXA COMUNICACION*, *25*, 63–84. https://publons.com/publon/46955778/

González-Esteban, E. (2019). El reconocimiento de los stakeholders desde una ética empresarial cordis. In E. González-Esteban, J. C. Siurana Aparisi, J. L. López-González, y M. García-Granero (Eds.), Ética y Democracia. Desde la razón cordial (pp. 59–66). Comares.

González-Esteban, E., y Calvo, P. (2022). Ethically governing artificial intelligence in the field of scientific research and innovation. Heliyon, 8(2), 1–9. https://doi.org/10.1016/j.heliyon.2022.e08946

González Esteban, E., Siurana, J. C., López González, J. L., y García-Granero Gascó, M. (Eds.). (2019). Ética y democracia: desde la razón cordial [Book]. Editorial Comares.

Grmuša, T., & Prelog, L. (2020). The Role of New Technologies in Combatting Fake News –Experiences and Challenges of Croatian Media Organisations. *Medijske Studije*, *11*(22), 62–80. https://doi.org/10.20901/MS.11.22.4

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hansen, M. , Roca-Sales, M. , Keegan, J. , & King, G. . (2017). *Artificial Intelligence: Practice and Implications for Journalism*. https://doi.org/10.7916/D8X92PRD

Helberger, N., Eskens, S., Van Drunen, M., Bastian, M., & Moeller, J. (2019). *Implications of AI-Driven Tools in the Media for Freedom of Expression*. Helberger, Natali Eskens,

Sarah Van Drunen, Max Bastian, Mariella Moeller, Judith. https://rm.coe.int/coe-ai-report-final/168094ce8f#:~:text=While the introduction of AI,opportunities of access to information.

Jina, P. (2019). Legal issues related to artificially intelligent journalism. *Science, Technology and Law*, *10*(2), 119–154.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kranzberg, M. (1986). Technology and History: "Kranzberg's Laws." *Technology and Culture*, *27*(3), 544. https://doi.org/10.2307/3105385

Lewis, S. C., Sanders, A. K., & Carmody, C. (2018). Libel by Algorithm? Automated Journalism and the Threat of Legal Liability: *Journalism & Mass Communication Quarterly*, *96*(1), 60–81. https://doi.org/10.1177/1077699018755983

Lie, J. W. (2021). How has the Entertainment News Production Practices Changed Since the Portal Site's Introduction of AI News Curation?: An Exploratory Study. *Korean Journal of Broadcasting & Telecommunications Research*, *113*, 93–121.

López-García, X., & Vizoso, Á. (2021). Periodismo de alta tecnología: signo de los tiempos digitales del tercer milenio. *Profesional de La Información*, *30*(3). https://doi.org/10.3145/EPI.2021.MAY.01

Manfredi Sánchez, J. L., & Ufarte Ruiz, M. J. (2020). Artificial intelligence and journalism: A tool to fight disinformation. *Revista CIDOB d'Afers Internacionals*, *124*, 49–72. https://doi.org/10.24241/RCAI.2020.124.1.49

Meyers, C. (2010). *Journalism ethics : a philosophical approach* (C. Meyers (Ed.)) [Book]. Oxford University Press.

Murcia Verdú, F. J., & Ufarte Ruiz, M. J. (2019). Risk map of hi-tech journalism. *Hipertext.Net: Revista Académica Sobre Documentación Digital y Comunicación Interactiva*, *0*(18), 47–55. https://doi.org/10.31009/hipertext.net.2019.i18.05

Newman, N. (2022). *Journalism, media, and technology trends and predictions 2022*. *Reuters Institute for the Study of Journalism*. https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2022

Parratt-Fernández, S., Mayoral-Sánchez, J., & Mera-Fernández, M. (2021). The application of artificial intelligence to journalism: an analysis of academic production. *Ediciones Profesionales de La Información SL*, *30*(3). https://doi.org/10.3145/EPI.2021.MAY.17

Pérez-Seijo, S., Gutiérrez-Caneda, B., & López-García, X. (2020). Periodismo digital y alta tecnología: de la consolidación a los renovados desafíos. *INDEX COMUNICACION*, *10*(3), 129–151. https://doi.org/10.33732/ixc/10/03period

Rinehart, A., & Kung, E. (2022). *Artificial Intelligence in Local News. A survey of US newsrooms' AI readiness*.

Robles Carrillo, M. (2020). La gobernanza de la inteligencia artificial: contexto y parámetros generales. *Revista Electrónica de Estudios Internacionales*, *2020*(39). https://doi.org/10.17103/REEI.39.07

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, *26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Salaverría, R. (2015). Periodismo en 2014: balance y tendencias. *Cuadernos de Periodistas*, *29*, 9–22. https://www.cuadernosdeperiodistas.com/periodismo-en-2014-balance-y-tendencias/

Salazar García, I. A. (2018). Los robots y la Inteligencia Artificial. Nuevos retos del periodismo. *Doxa Comunicación. Revista Interdisciplinar de Estudios de Comunicación y Ciencias Sociales*, *27*, 295–315. https://doi.org/10.31921/doxacom.n27a15

Salnikova, L. (2019). Robots Versus Journalists: Does Journalism Have a Future? *Theoretical and Practical Issues of Journalism*, *8*(4), 668–678. https://doi.org/10.17150/2308-6203.2019.8(4).668-678

Sanahuja Sanahuja, R, y López Rabadán, P. (2021). Ámbitos de aplicación periodística de la Inteligencia Artificial. Mapa conceptual, funciones profesionales y tendencias en desarrollo en el contexto de la pandemia global de la Covid-19. Razón y Palabra, 25(112). https://doi.org/10.26807/RP.V25I112.1827

Sanahuja Sanahuja, R., y López Rabadán, P. (2022). Aspectos éticos del uso de la inteligencia artificial en el periodismo: posicionamientos y preocupaciones desde el ámbito académico. En *Nuevas tendencias en la comunicación y en la investigación: Su reflejo profesional y académico*. Ed. Editorial Gedisa. Col·lección Biblioteca de la Educación. Herramientas Universitarias. ISBN 9788418914546.

Shin, D. (2021). The perception of humanness in conversational journalism: An algorithmic information-processing perspective. *New Media and Society*. https://doi.org/10.1177/1461444821993801

Trattner, C., Jannach, D., Motta, E., Meijer, I. C., Diakopoulos, N., Mehdi Elahi, ·, Opdahl, A. L., Tessem, · Bjørnar, Borch, N., Fjeld, M., Øvrelid, L., Koenraad, ·, Smedt, D., & Moe, · Hallvard. (2021). Responsible media technology and AI: challenges and research directions. *AI and Ethics 2021*, *1*, 1–10. https://doi.org/10.1007/S43681-021-00126-4

Túñez-López, J. M., Fieiras Ceide, C., & Vaz-Álvarez, M. (2021). Impacto de la Inteligencia Artificial en el Periodismo: transformaciones en la empresa, los productos, los contenidos y el perfil profesional . *Communication & Society*, *34 (I)*, 177–193. https://dadun.unav.edu/bitstream/10171/59967/2/12. Túñez et al. ESP VF.pdf

Túñez López, J. M. (2021). Tendencias e impacto de la inteligencia artificial en comunicación: cobotización, gig economy, co-creación y gobernanza. *Fonseca, Journal of Communication*, *22*, 5–22. https://dialnet.unirioja.es/servlet/articulo?codigo=7955730&info=resumen&idioma=ENG

Ufarte Ruiz, M. J., Calvo Rubio, L. M., & Murcia Verdú, F. J. (2020). Las tecnologías hi-tech en los grados en Periodismo. Planes de estudios, formación de los periodistas y propuestas de inserción curricular. *AdComunica. Revista Científica de Estrategias, Tendencias e Innovación En Comunicación*, *20*, 43–66. https://doi.org/10.6035/2174-0992.2020.20.3

Ufarte Ruíz, M. J., Calvo Rubio, L. M., & Murcia Verdú, F. J. (2021). Los desafíos éticos del periodismo en la era de la inteligencia artificial. *Estudios Sobre El Mensaje Periodístico*, *27*(2), 673–684. https://doi.org/10.5209/esmp.69708

Ufarte Ruiz, M. J., Fieiras-Ceide, C., & Túñez-López, M. (2020). The teaching-learning of automated journalism in public institutions: Studies, feasibility proposals and future impact of artificial intelligence. *Analisi*, *62*, 131–146. https://doi.org/10.5565/rev/analisi.3289

Ufarte Ruiz, M. J., & Manfredi Sánchez, J. L. (2019). Algorithms and bots applied to journalism. The case of Narrativa Inteligencia Artificial: structure, production and informa-

tive quality. *Doxa Comunicación. Revista Interdisciplinar de Estudios de Comunicación y Ciencias Sociales*, *29*, 213–233. https://doi.org/10.31921/doxacom.n29a11

UNESCO. (2021). *Proyecto de texto de la recomendación sobre la ética de la inteligencia artificial* (Issue September).

Ventura, P. (2021). *Algoritmos en las redacciones: Retos y recomendaciones para dotar a la inteligencia artificial de los valores éticos del periodismo*. Consell de la Informació de Catalunya.

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 147-161
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.551741

Sobre los diferentes ritmos del derecho y la Inteligencia Artificial. La desincronización como patología social

On the different rhythms of law and Artificial Intelligence De-synchronization as a social pathology

CÉSAR ORTEGA-ESQUEMBRE[*]

Resumen. El objetivo de este trabajo es estudiar el desajuste temporal que se produce, dentro de la Unión Europea (UE), entre las innovaciones tecnológicas en materia de Inteligencia Artificial y sus regulaciones jurídicas[1]. Para ello se parte de una tesis formulada por Hartmut Rosa, sociólogo alemán cercano a la Teoría Crítica de la sociedad, de acuerdo con la cual las estructuras temporales de la política no resultan hoy ya compatibles con el ritmo de cambio de algunas esferas sociales. Esto produce una nueva forma de patología social, que Rosa denomina "riesgos de la desincronización", y cuyo efecto más preocupante hay que buscar en el desplazamiento de los procesos de toma de decisiones desde el ámbito de la política hacia otros ámbitos de la sociedad más rápidos, fundamentalmente el mercado. Para analizar esta problemática se reconstruye, en primer lugar, la tesis de la desincronización como patología social derivada de la aceleración. Tras

Abstract. The aim of this paper is to study the temporal de-synchronization that occurs, within the European Union (EU), between technological innovations in Artificial Intelligence and its legal regulations. To do that, I base on a thesis formulated by Hartmut Rosa, a German sociologist close to the Critical Theory of society, according to which the temporal structures of politics are no longer compatible with the pace of change in some others social spheres. This produces a new form of social pathology, which Rosa calls "risks of de-synchronisation", and whose most worrying effect must be sought in the displacement of decision-making processes from the political sphere to faster areas of society, fundamentally the market. To analyze this problem, I first reconstruct the thesis of desynchronization as a social pathology derived from acceleration. Then, I map the AI legislation process at EU level. Finally, I defend the thesis

Recibido: 19/12/2022. Aceptado: 04/06/2023.

* Profesor Ayudante Doctor en el Departamento de Filosofía de la Universidad de Valencia. Correo electrónico: cesar.ortega@uv.es Líneas de investigación: teoría crítica de la sociedad, filosofía social, filosofía política. Publicaciones recientes: "La prehistoria filosófica de la Teoría Crítica como crítica de la racionalización sociocultural. ¿Patologías sociales o patologías culturales?". *Anales del Seminario de Historia de la Filosofía*. 39 -1, 2022, pp. 157 – 168; *Habermas ante el siglo XXI*. Madrid, Tecnos, 2021.

1 Este estudio se inserta en el Proyecto Coordinado de Investigación Científica y Desarrollo "Ética discursiva y Democracia ante los retos de la Inteligencia Artificial" PID2019-109078RB-C21 y PID2019-109078RB-C22 financiado por MCIN/ AEI /10.13039/501100011033. Asimismo, quisiera agradecer a los profesores Jesús Conill, Agustín Domingo, Juan Carlos Siurana, Pedro Jesús Teruel y José Luís López los estimulantes comentarios que formularon a la primera versión de este trabajo, así como a los dos revisores anónimos por sus sugerencias de mejora.

ello, se hace un mapeo del proceso de legislación en materia de IA a nivel de la UE. Por último, se defiende la tesis de que este proceso legislativo, ciertamente muy admirable, ha llegado sistemáticamente tarde.
Palabras clave: desincronización, Inteligencia Artificial, Unión Europea, aceleración, derecho

that this legislative process, which is certainly very admirable, has consistently come too late.
Keywords: de-synchronisation, Artificial Intelligence, European Union, acceleration, law

Introducción

Es conocida la historia de los denominados "ludditas". Entre 1811 y 1812, algunos condados centrales de Gran Bretaña, entre ellos Yorkshire, Lancashire o Nottinghamshire, vieron nacer un curioso movimiento que tomó su nombre de Ned Ludd, un personaje legendario cuya identidad no es del todo cierta. Los integrantes de este movimiento, en su mayoría trabajadores artesanales del sector textil, reaccionaron de manera organizada y más o menos violenta contra el proceso de industrialización, cuyo símbolo más diabólico veían encarnado en las máquinas textiles de las incipientes factorías. Muy al contrario de lo que solamente medio siglo después Marx establecería como condición *sine qua non* para la emancipación del ser humano, estas "víctimas del progreso" no veían en el desarrollo de las fuerzas productivas precisamente la avanzadilla de la evolución social (Marx, 1970), sino más bien el camino directo hacia la decadencia de sus tradicionales, y muy queridas, formas de vida. Por eso la acción principal de su lucha fue la destrucción sistemática de telares automáticos.

Que los procesos de racionalización social resultaron finalmente un estadio necesario en el camino hacia la emancipación es algo tan evidente, que hoy se utiliza el término "luddita" en un sentido puramente despectivo[2]. A mi modo de ver, esta utilización del término no resulta, sin embargo, enteramente justa. La intuición que aquellos primitivos críticos del progreso solamente pudieron articular en forma muy tosca no deja de contener el germen de una idea de la que aún hoy podemos aprender algo. Pues lo cierto es que los ludditas no eran sujetos esencialmente tecnófobos, sino más bien ciudadanos con la conciencia de clase suficiente como para saber que únicamente podrían aceptar la introducción de las nuevas técnicas de producción cuando éstas estuvieran al servicio de la satisfacción general de las necesidades (Jones, 2006). Cuestionarse al servicio de qué fines se encuentran los avances técnicos, así como los posibles peligros que pueda entrañar una aplicación precipitada, no es desde luego una extravagancia propia de artesanos enloquecidos, sino algo que cada generación debería practicar de manera cuidadosa.

La dialéctica entre los progresos técnicos y los movimientos de reacción ha resultado una constante a lo largo de las cuatro hornadas de la revolución industrial —mecanización; producción en masa y electricidad; informática y automatización; y espacios ciberfísicos e Inteligencia Artificial (IA)—. En esta dialéctica, los movimientos de reacción no siempre

2 Véase, por ejemplo, el uso que hace del término C. P. Snow en su famosa conferencia "Las dos culturas" (Snow, 2000). Mucho más evidente es el caso de la *Information Technology & Innovation Foundation*, un lobby tecnológico estadounidense que ha adoptado la curiosa tradición de repartir anualmente los denominados Premios Ludditas, cuyos galardonados son supuestos enemigos del progreso socio- económico.

han ocupado el espectro político que denominamos "conservador", sino que son y han sido muchos los críticos del progreso que han motivado sus posturas en convicciones estrictamente progresistas. Basta echar un vistazo, por ejemplo, a los "críticos del crecimiento" que se encuentran normalmente a la izquierda de los partidos socialdemócratas en los parlamentos europeos. Sea como fuere, es evidente que la cuarta revolución industrial, también denominada Industria 4.0, ha reeditado esta dialéctica. Y acaso de una forma más extrema que nunca, pues parece existir consenso sobre el hecho de que algunos de los elementos que componen esta nueva hornada, especialmente los que tienen que ver con la IA, están transformando nuestra forma de vida de una manera nunca vista hasta el momento.

Preguntarse por las posibles consecuencias adversas de la IA no constituye un acto de reacción antimoderna, sino un ejemplo de ese venerable ejercicio de discernimiento al que solemos dar el nombre de "crítica". Ciertamente, nuestro *Zeitgeist* o *espíritu de la época* parece ser tan entusiasta con respecto a los avances que traerán los sistemas de IA —muchos de los cuales, sin duda, constituyen ya un progreso incontrovertible—, que hoy uno apenas puede plantearse la pregunta sobre la *conveniencia* de esos avances sin ser tachado inmediatamente de reaccionario. Que la respuesta a esta pregunta no es en todos los casos afirmativa es algo que puede constatarse observando un hecho muy sencillo: todo el mundo reconoce que la introducción de sistemas de IA en nuestras vidas requiere el establecimiento de ciertas normas —de tipo ético y/o jurídico— que lo limite y regule.

Ahora bien, los ordenamientos normativos no siempre transcurren a la misma velocidad que los progresos tecnológicos, de suerte que suelen darse casos en los que aparecen nuevas tecnologías para cuya regulación no tenemos a disposición normas operativas[3]. El desarrollo de la IA, muy particularmente, se encuentra acelerado hasta un punto tal, que los procesos legislativos de creación de derecho están condenados a irle siempre a la zaga. El objetivo de este trabajo es estudiar el desajuste temporal que se produce, dentro de la Unión Europea (UE), entre las innovaciones tecnológicas en materia de IA y sus regulaciones jurídicas. Para ello voy a partir de una tesis formulada por Hartmut Rosa, sociólogo alemán cercano a la llamada Teoría Crítica, de acuerdo con la cual las estructuras temporales de la política, es decir, el tiempo requerido para la toma de decisiones políticas traducibles al lenguaje del derecho, no resulta ya compatible con el ritmo de cambio de algunas esferas sociales. Esto produce una curiosa forma de patología social, que Rosa denomina "riesgos de la desincronización", y cuyo efecto más preocupante hay que buscar en el desplazamiento de los procesos de toma de decisiones desde el ámbito de la política hacia otros ámbitos de la sociedad más rápidos. Para analizar esta problemática daré tres pasos. En primer lugar, reconstruiré la tesis de la desincronización como patología social derivada de la aceleración. Tras ello, haré un mapeo del proceso de legislación en materia de IA a nivel de la UE. Por último, defenderé la tesis de que este proceso legislativo, ciertamente muy admirable y sin duda en consonancia con los valores y principios supremos de la Unión, ha llegado tarde.

3 Tal y como me han hecho ver con mucho acierto Jesús Conill y Juan Carlos Siurana, aunque las regulaciones jurídicas suelen aparecer, como es natural, después de las innovaciones tecnológicas que reclaman dicha regulación, esto no siempre ocurre así. Por ejemplo, aunque hoy todavía no parece existir la tecnología necesaria para la clonación de seres humanos, lo cierto es que existe desde hace tiempo una regulación jurídica que prohíbe su uso.

1. La desincronización como patología social

En su estudio sobre la aceleración social, Hartmut Rosa parte de la idea de que nuestra forma de estar en el mundo depende de las "estructuras temporales" de la sociedad en que vivimos. Sobre la base de esta idea, su tesis central es que la modernidad occidental está sometida a un proceso creciente de aceleración social, proceso que en la modernidad tardía –la que arranca aproximadamente en los años setenta del siglo XX– supera un umbral tal que comienza a ocasionar efectos patológicos (Rosa, 2005; Ortega-Esquembre, 2021). En términos abstractos, la aceleración puede ser entendida como un «incremento en cantidad por unidad de tiempo (o, lógicamente equivalente, una reducción de la cantidad de tiempo para cantidades fijas)» (Rosa, 2005). Esta definición no resulta sin embargo operativa para apresar las múltiples formas de aceleración social empíricamente observables. Para resolver este déficit, Rosa propone distinguir tres dimensiones, que permanecen por lo demás conectadas entre sí: la aceleración técnica, la aceleración del cambio social y la aceleración del ritmo de vida.

Los ejemplos paradigmáticos de la aceleración técnica se encuentran en las transformaciones de los medios de transporte, en las comunicaciones y en la producción de bienes. Con la aceleración de los medios de transporte —primero con el ferrocarril y más tarde con el surgimiento del avión y el automóvil—, la conciencia del espacio sufrió una importante transformación. El espacio se convierte, por así decirlo, en una mera función del tiempo: tardamos tantas horas en llegar de un país a otro, en atravesar un país de norte a sur, en recorrer una ciudad. Este proceso ha sufrido a su vez un impulso de consecuencias incalculables con la revolución digital y la transmisión electrónica de información. Por otro lado, la aceleración masiva de la producción ha hecho posible a juicio de Rosa satisfacer el imperativo de la sociedad capitalista, es decir, la conversión de los objetos en mercancías que se vuelven obsoletas cada vez más rápidamente.

En segundo lugar, la aceleración del cambio social tiene que ver con el mayor ritmo de cambio en prácticas y orientaciones de acción. Para definir esta forma de aceleración, Rosa se sirve del concepto, acuñado por Hermann Lübbe, de "contracción del presente", siendo el presente un periodo temporal en el que las experiencias y expectativas permanecen estables (Lübbe, 1988). La contracción del presente se aprecia de forma especialmente clara desde el punto de vista de los ritmos generacionales: si en las sociedades de la modernidad temprana los cambios en las prácticas y orientaciones de acción —por ejemplo, los cambios de profesión— ocurrían solo a lo largo de varias generaciones, en la modernidad clásica estos cambios pasan a sincronizarse con la secuencia de generaciones. En la modernidad tardía o postmodernidad, en tercer lugar, nos encontramos con un ritmo de cambio intrageneracional. Tres son las consecuencias más preocupantes de esta contracción del presente: en primer lugar, el rápido deterioro del acervo de saber cultural; en segundo lugar, la creciente brecha intergeneracional que se abre con la creación de mundos de la vida generacionales totalmente extraños entre sí; en tercer lugar, la creación de una pendiente resbaladiza de cambio social a la que los sujetos no pueden sustraerse más que al precio de quedar fuera de la carrera —y el consecuente incremento de trastornos de ansiedad, estrés o depresión—.

En tercer lugar, la aceleración del ritmo de vida es entendida como un incremento de episodios de acción por unidad de tiempo. Objetivamente, la aceleración del ritmo de vida implica una condensación de episodios de acción —por ejemplo, el acortamiento del tiempo dedicado a comer, la reducción de las horas de sueño o la implementación de multitareas—. Subjetivamente, este proceso se expresa en la creciente experiencia de terror ante la posibilidad de perder el tiempo y la sensación de no tener nunca tiempo suficiente para emprender lo que se considera "realmente importante".

Rosa se esfuerza por analizar algunas de las consecuencias más preocupantes de estos procesos de aceleración social. La última ola de aceleración, que tuvo lugar a finales de los años setenta, conduce a una transformación de las formas individuales y colectivas de identidad. Los elementos que caracterizan los diagnósticos de las identidades postmodernas tienen que ver con esto: la disolución de estructuras del sujeto antes estables en favor de una identidad abierta, experimental y constantemente transitoria. Este estado situacional de la identidad se traslada, como segunda consecuencia patológica de los procesos de aceleración, al ámbito de la política. Y éste es justamente el aspecto de mayor relevancia para los propósitos de nuestro trabajo. Si en la modernidad clásica los tiempos requeridos para la institucionalización de la formación de la voluntad política, la toma de decisiones democrática y su implementación eran compatibles con el ritmo de los desarrollos sociales, de suerte que el sistema político disponía de tiempo suficiente para tomar decisiones sobre cómo organizar estos desarrollos, con la última ola de aceleración esto cambia, y la posibilidad del auto-control social queda puesta en duda. Surge así una forma de patología social nunca vista hasta el momento, a la que Rosa da el nombre de "riesgos de la desincronización". Efectivamente, es claro que la temporalidad intrínseca a la política deliberativa solo puede ser acelerada hasta cierto punto —uno no puede, por así decirlo, acelerar el propio acto de argumentación racional en la esfera pública—. Sin embargo, los desarrollos en otros terrenos, como la economía o las nuevas tecnologías, que requieren ser legislados por dicha política, no están sujetos a estas limitaciones temporales. En ellos el proceso de aceleración opera de una forma prácticamente incontrolable.

¿En qué sentido, sin embargo, podemos decir que esta desincronización constituye una patología social? Aunque, ciertamente, existe un gran debate sobre el significado preciso del término "patología social" (Honneth, 2011; Neuhouser, 2022), a mi modo de ver podemos hablar de ellas en sentido estricto únicamente cuando se trata de prácticas o dinámicas sociales que reúnen, al menos, los siguientes rasgos: ser inducidas por el propio sistema social, y no por agentes externos o por elementos internos individuales; poseer una dinámica claramente identificable, que les ofrezca un carácter permanente en lugar de esporádico; y tener como consecuencia no solamente una distribución injusta de recursos, sino una paralización de la posibilidad de desarrollar formas de vida autorrealizadas y autodeterminadas. Parece claro que la desincronización entre las esferas de la política y el desarrollo tecnológico reúne estas características. En primer lugar, el proceso de desincronización es la consecuencia de una dinámica interna al propio sistema social, a saber, la permanente tendencia a la aceleración. En segundo lugar, este proceso no constituye un fenómeno esporádico, sino un rasgo permanente de dicha dinámica sistémica. Por último, la desincronización no tiene como resultado, o al menos no prioritariamente, una distribución material asimétrica, sino un impedimento sistemático para la consecución de vidas *logradas*. Al afirmar que la

desincronización constituye una patología social, no quiero decir que cualquier desajuste temporal entre dos esferas sociales diferentes sea necesariamente patológico. Naturalmente, existen esferas más rápidas que otras, y es natural que la creación de derecho acontezca después de la aparición de fenómenos nuevos que reclaman su regulación. Lo patológico, esto es lo que quiero defender, es que el *ritmo* de aceleración de una de esas esferas, en este caso los avances en materia de IA, sea hasta tal punto *mayor* que el ritmo de los procesos de creación de derecho, que sus "regulaciones" sean asumidas finalmente por una esfera no siempre sometida al control democrático, a saber, el mercado[4]. Sea como fuere, Aunque Rosa ha analizado algunos de los corolarios patológicos de este proceso en obras posteriores (Rosa, 2009; 2016a), a nosotros nos interesa conectar ahora el diagnóstico sobre la desincronización con la rápida implementación de los sistemas de IA y los procesos de regulación jurídica que se dan en la UE.

2. La legislación europea sobre Inteligencia Artificial

Antes de analizar la legislación vigente en la UE en materia de IA, tal vez convendría ofrecer una breve caracterización de esta tecnología, así como de sus beneficios y riesgos presentes y potenciales. En el denominado "Libro Blanco sobre la Inteligencia Artificial", elaborado por la Comisión Europea en el año 2020, se ofrece la siguiente definición:

> Los sistemas de IA son programas informáticos (y posiblemente también equipos informáticos) diseñados por seres humanos que, dado un objetivo complejo, actúan en la dimensión física o digital mediante la percepción de su entorno mediante la adquisición de datos, la interpretación de los datos estructurados o no estructurados, el razonamiento sobre el conocimiento o el tratamiento de la información fruto de estos datos y la decisión de las mejores acciones que se llevarán a cabo para alcanzar el objetivo fijado (COM (2020) 65 final (19-02-2020): 20).

La IA permite que operaciones o decisiones tradicionalmente ejecutadas por seres humanos sean emprendidas por algoritmos, que se "nutren" de datos recogidos de su entorno a fin de obtener *outputs* en principio más racionales. Los ejemplos son muy numerosos, y van desde los vehículos sin conductor hasta la recomendación de canciones en plataformas de música online, pasando por la emisión de diagnósticos médicos, la toma de decisiones en los procesos de selección de candidatos para un puesto de trabajo o las técnicas de reconocimiento facial en tiempo real. Las ventajas actuales y potenciales de la utilización de esta nueva tecnología son evidentes, y no es descabellado afirmar que se trata del instrumento más importante del que dispondremos a corto plazo para afrontar retos tan apremiantes como el diagnóstico temprano de enfermedades crónicas, el cambio climático o la seguridad ciudadana.

Aunque negar estas virtualidades sería propio, naturalmente, de un neo-luddismo más bien romántico y reaccionario, lo cierto es que el hecho de que nuestra interacción con

4 Jesús Conill me ha llamado la atención sobre este problemático asunto, que sin esta aclaración adicional conduciría a la sorprendente conclusión de calificar como patológico un fenómeno *natural y necesario*.

el medio y con el resto de seres humanos quede crecientemente mediada por algoritmos presenta riesgos y problemas de magnitud comparable a sus ventajas (Mittelstadt et al, 2016). En primer lugar, es evidente que los algoritmos son diseñados por personas, y aunque el supuesto punto fuerte de estos sistemas es que están libres de los sesgos típicos del ser humano[5], en principio no hay razón para creer que los programadores no trasmitirán, consciente o inconscientemente, sus propios prejuicios al sistema diseñado. En segundo lugar, incluso cuando el diseño y los parámetros sean transparentes y aceptados, ello no garantiza la consecución de decisiones éticamente aceptables, como se puede ver, por ejemplo, en algoritmos que discriminan inadvertidamente a ciertos grupos sociales marginados (Mittelstadt et al, 2016). Los sistemas algorítmicos no son, en una palabra, éticamente neutrales, lo cual exige, como dice Adela Cortina, reflexionar sobre «cómo orientar el uso humano de estos sistemas de forma ética» (Cortina, 2019).

Según Tsamados *et al*, son seis fundamentalmente los problemas éticos planteados por los algoritmos. En primer lugar, el problema de la "evidencia no concluyente", es decir, el hecho de que los algoritmos produzcan *outputs* que no se basan en conexiones causales, sino en meras correlaciones estadísticas identificadas entre los datos disponibles. Este hecho puede conducir a acciones injustificadas. En segundo lugar, el problema de la "evidencia impenetrable", que tiene que ver con la opacidad que caracteriza a los algoritmos a la hora de tomar decisiones, lo cual puede traducirse en la inexistencia de una "rendición de cuentas". En tercer lugar, el problema de la evidencia errónea, que puede conducir al surgimiento de sesgos. En cuarto lugar, el problema de los "resultados injustos", que es la consecuencia de que los algoritmos realicen su "minería de datos" sin tener en cuenta criterios de tipo ético como la no discriminación por sexos o estratos sociales. En quinto lugar, el problema de los "efectos transformadores", es decir, el riesgo de que los sistemas algorítmicos obstaculicen la autonomía humana, por ejemplo prefigurando las decisiones humanas mediante el envío de información política dirigida y segmentada en función de los datos generados por cada persona. En sexto y último lugar, el problema de la trazabilidad. La falta de transparencia y explicabilidad de los algoritmos dificulta el trazado de la responsabilidad moral/jurídica que se deriva de las decisiones adoptadas (Tsamados et al, 2021).

Ciertamente, la literatura sobre los problemas éticos derivados de la IA es muy prolija[6]. Pero a nosotros no nos interesa analizar en detalle estos problemas, sino más bien estudiar la forma en que la "conciencia de riesgo" por ellos suscitada ha empujado a la UE a establecer un marco jurídico común sobre la materia. La preocupación de la UE por esta cuestión se hizo patente, por ejemplo, en la constitución del "Comité especial sobre Inteligencia Artificial en la Era digital" puesto en marcha por el Parlamento Europeo en septiembre de 2020. Este comité se ha venido reuniendo periódicamente desde entonces, y en abril de 2022 emitió el llamado "European Parliament final Report on Artificial Intelligence in a digital age" (2020/2266(INI)).

5 Éste es el argumento principal de los defensores de lo que se ha dado en llamar "democracia algorítmica". Para un estudio crítico de esta propuesta véase Calvo, 2019.

6 Una excelente aproximación a esta problemática, que además disecciona analíticamente los problemas éticos derivados de cada uno de los tres tipos diferentes de IA –superior, general y especial– se encuentra en Cortina, 2019. Agustín Domingo, por su parte, ha analizado los posibles riesgos derivados de una aplicación de los sistemas de IA a las tareas de cuidado. Véase Domingo, 2020.

En abril de 2021, la Comisión Europea publicó un documento que establece las reglas que deben regir la producción, comercialización y uso de sistemas de IA dentro de la UE. Este documento constituye el primer marco legal unificado sobre IA que se lleva a cabo a nivel planetario, y busca hacer de Europa el centro mundial de una IA "digna de confianza" (*trustworthy*) (COM (2021) 206 final). Pero este marco constituye el *resultado último* de un proceso de regulación que se inició bastante tiempo antes. Para los objetivos de este trabajo, resulta especialmente importante rastrear los orígenes de este proceso.

Si no me equivoco, el primer documento de la Comisión que abordó sistemáticamente la necesidad de una regulación jurídica de la IA —dejando de lado documentos que, de forma tangencial, trataron algunos de los aspectos vinculados a esta nueva tecnología, como las regulaciones en materia de automatización de la fuerza de trabajo y digitalización (COM (2016) 180 final)— es la denominada "Estrategia Europea para la IA", publicada en abril de 2018 (COM (2018) 237 final). En este texto ya se llama la atención sobre la necesidad de trabajar en un marco europeo sólido que permita «aprovechar al máximo las oportunidades que brinda la IA y abordar los nuevos retos que conlleva». Resulta muy significativo que ya en la primera página se advierta que dicho marco, que naturalmente no debe obstaculizar, sino favorecer una innovación tecnológica que permita a Europa jugar un rol relevante en el panorama internacional, debe estar basado en los valores y derechos fundamentales de la Unión. La regulación en materia de IA debe ser coherente con los valores enunciados en el Artículo 2 del Tratado de la Unión Europea, es decir, con los valores del respeto hacia la dignidad humana, la libertad, la igualdad, el pluralismo, la tolerancia o la solidaridad. Asimismo, dicha regulación debe garantizar la correcta protección de los derechos reconocidos en la Carta de los Derechos Fundamentales de la Unión Europea.

Aunque este documento supone sin duda una declaración de intenciones, lo cierto es que no se trata más que de una estrategia sin especificaciones concretas, y, obviamente, también sin carácter vinculante. La Estrategia Europea para la IA describe el camino a seguir para lograr «poner al servicio del progreso humano el potencial de la IA» (COM (2018) 237 final: 22). Pocos meses después de la publicación de esta Estrategia, la Comisión presentó un plan coordinado con los Estados miembro para armonizar estrategias (COM (2018) 795 final), a lo que siguió la redacción, por parte del Grupo de Expertos de Alto Nivel en IA de la UE, de unas directrices no vinculantes para una IA confiable, seguido de una comunicación de la Comisión en la que se acogían favorablemente dichas directrices (COM (2019) 168 final). Tras la publicación de todos estos documentos, en febrero del año 2020 apareció el denominado "Libro Blanco sobre la IA", sin duda uno de los textos más importantes para comprender el enfoque ético-jurídico de la UE en relación con la IA (COM (2020) 65 final).

Dos son los conceptos fundamentales en torno a los que se articula el Libro Blanco: "excelencia" y "confianza". Su objetivo es formular una serie de alternativas políticas para alcanzar un modelo de IA que, promoviendo el desarrollo y adopción de esta nueva tecnología por parte de ciudadanos, empresas y administraciones públicas, así como la inversión pública y privada, minimice sus riesgos potenciales y sea consistente con el orden normativo europeo. En este sentido, se afirma que uno de los obstáculos principales para la implementación de la IA es la falta de un marco regulador vinculante para todos los Estados miembro. Solo una regulación jurídica a escala de la Unión, tal es la convicción, puede prevenir los riesgos considerados más inminentes, entre los que se menciona la fragmentación

del mercado interior europeo, la vigilancia masiva por parte de las autoridades estatales, las cuestiones vinculadas con la responsabilidad civil[7], la inseguridad, la desprotección de los datos personales, la falta de privacidad y la discriminación. Aunque, obviamente, muchas de las normas ya existentes en la UE, por ejemplo el Reglamento General de Protección de Datos, pueden resultar operativas para combatir algunos de estos riesgos, la tesis de la Comisión es que resulta imprescindible la creación de una legislación específica sobre IA.

Tras la publicación del Libro Blanco, la Comisión abrió un proceso de consulta pública a todos los *stakeholders*, incluidas empresas, administraciones públicas, grupos de investigación y ciudadanos particulares. Esta consulta estuvo abierta entre el 19 de febrero y el 14 de junio de 2020, y recibió un total de 1215 contribuciones. Teniendo en cuenta las respuestas ofrecidas, así como algunas resoluciones adoptadas entretanto por el Parlamento Europeo[8], el 21 de abril de 2021 la Comisión publicó su "Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts" (COM (2021) 206 final). Este documento constituye la propuesta para un marco jurídico común y vinculante para todos los Estados miembro de la UE en materia de IA.

Este marco establece normas de obligado cumplimiento para el uso y la comercialización de sistemas de IA, normas que se aplican a proveedores que comercializan estos sistemas dentro la UE, independientemente de si están establecidos en la Unión o en un país tercero, así como a los usuarios de sistemas de IA localizados dentro de nuestras fronteras. La regulación adopta un enfoque basado en el riesgo (*risk-based approach*), y el punto de partida es que la intervención debe limitarse a los casos que presentan riesgos claros. En este sentido, se diferencian tres niveles de riesgo: aquellos sistemas de IA que crean un "riesgo inaceptable", aquellos que crean un "riesgo alto", y aquellos que crean un "riesgo bajo" o "mínimo". Mientras que los primeros quedan estrictamente prohibidos, en la medida en que contravienen los valores de la Unión, los segundos quedan sometidos a una serie de restricciones específicas en relación con ciertos usos. Para los sistemas que entrañan un riesgo bajo o mínimo, en tercer lugar, se recomienda diseñar y adoptar códigos de conducta ética.

Con respecto al primer tipo, se establece que quedan prohibidos los siguientes sistemas de IA: aquellos que «implementan técnicas subliminales más allá de la conciencia de una persona con el fin de distorsionar materialmente su comportamiento de una forma que causa o es probable que cause daños físicos o psicológicos»; aquellos que «explotan cualquiera de las vulnerabilidades de un grupo específico de personas debido a su edad, minusvalía física o mental, con el fin de distorsionar materialmente el comportamiento de una persona perteneciente a ese grupo»; aquellos sistemas «establecidos por una autoridad pública para la evaluación o clasificación de la fiabilidad de una persona natural durante cierto periodo

7 Este problema afecta sobre todo a aquellos sistemas capaces de aprendizaje. Efectivamente, la imputación de responsabilidad resulta muy sencilla cuando los resultados pueden preverse en la fase de diseño, pero no cuando el propio sistema algorítmico es capaz de aprender y tomar decisiones imposibles de anticipar por el programador.

8 De especial interés entre ellas resultan las siguientes: European Parlamient resolution of 20 October 2020 on a framework of ethical aspects of artificial intelligence, robotics and related technologies 2020/2012(INL); European Parliament Draft Report, Artificial Intelligence in criminal law and its use by the police and judicial authorities in criminal matters 2020/2016(INI); European Parliament Draft Report, Artificial intelligence in education, culture, and the audiovisual sector 2020/2017(INI).

de tiempo basándose en su comportamiento social o sus características personales conocidas o predichas» ; y aquellos «sistemas de identificación biométrica remota en tiempo real en espacios de acceso público para el propósito de la aplicación de la ley».

Con respecto al segundo tipo, la Comisión ha publicado un anexo donde aparecen detallados los sistemas considerados de alto riesgo. Entre ellos se incluyen los siguientes: sistemas de identificación biométrica y categorización de personas naturales; sistemas destinados a determinar la idoneidad de candidatos para el disfrute de ayudas y servicios sociales; o sistemas utilizados por autoridades judiciales para detectar el estado emocional de una persona natural. Estos sistemas están permitidos siempre y cuando cumplan con una serie de requisitos obligatorios, para lo cual se establece un sistema de gestión de riesgo y de evaluación del seguimiento de las normas.

Asentado en este enfoque, el reglamento detalla un sistema de gobernanza a nivel de la Unión y a nivel nacional. Siguiendo el espíritu del principio de subsidiariedad, se crea un Equipo Europeo de Inteligencia Artificial, cuyo objetivo es «proveer consejo y asistencia a la Comisión con el fin de contribuir a la cooperación efectiva entre las autoridades supervisoras nacionales y la Comisión». En todo caso, las normas propuestas serán aplicadas a través de un sistema de gobernanza nacional, construyéndose dicho sistema sobre la base de estructuras y autoridades ya existentes. Corresponde pues a los Estados miembro establecer las reglas específicas sobre las sanciones aplicables a la infracción de las normas recogidas en este documento, así como asegurar su correcto cumplimiento. No obstante, la Comisión establece una serie de multas administrativas a las que necesariamente deben estar sujetas determinadas infracciones.

Una vez analizado brevemente el contenido del marco jurídico común europeo sobre IA, así como el histórico de documentos europeos que han conducido a dicho marco, ahora estamos en condiciones de estudiar, utilizando para ello la categoría de "desincronización" desarrollada en la primera sección, la problemática relación entre los ritmos temporales de este proceso de legislación y los ritmos de la innovación tecnológica en materia de IA.

3. Comparativa entre dos ritmos temporales

Hemos visto que, si bien la IA está presente en documentos y planes presupuestarios europeos desde hace algunos años, el primer documento que aborda de forma sistemática la necesidad de una regulación jurídica de esta tecnología está fechado en abril de 2018. Aunque, ciertamente, la UE es pionera en esta regulación, llama la atención que la legislación en materia de IA haya llegado tan tarde, sobre todo si se tiene en cuenta que, en sentido estricto, *no ha habido legislación* hasta abril de 2021. ¿Cuándo apareció, así pues, la IA?

Parece existir consenso académico en torno a la idea de que el primer gran ejemplo de sistema de IA es el llamado "Logic Theorist", un programa diseñado en 1956 por Allen Newell, John Clifford Shaw y Herbert Simon que era capaz de hallar demostraciones de teoremas de lógica pura. Aunque éste fue el primer hito significativo de la IA, el nacimiento de la disciplina suele ubicarse en la celebración, también en 1956, del famoso congreso organizado por John McCarthy bajo el título "The Dartmouth Summer Research Project on Artificial Intelligence". Por lo demás, la investigación en torno a la pregunta de si las

máquinas son capaces de pensar se retrotrae todavía más, al menos hasta el célebre artículo de Alan Turing publicado en 1950 con el título "¿Puede pensar una máquina?" (Copeland, 1966: cap. 1). Durante los años cincuenta se desarrollan también sistemas capaces de participar en diversos juegos, como el ajedrez o las damas, con una habilidad muchas veces superior a la humana (Caparrini, 2018).

Ante la presencia de esta impresionante explosión de innovaciones en IA acontecida durante los años cincuenta y sesenta, uno podría preguntarse cómo es posible que hayamos tenido que esperar más de medio siglo para contar con una regulación jurídica sólida en Europa. Esta pregunta resultaría, sin embargo, injusta, pues el proceso de desarrollo de la IA no se ha mantenido en los niveles de aceleración vistos en sus orígenes. A comienzos de los años 70, los fondos para la investigación en IA, así como el interés de los departamentos universitarios, habían sufrido una fuerte caída. Este proceso de desaceleración, que duró hasta 1980, se conoce con el nombre de "the first AI Winter". A comienzos de los años ochenta, el interés por la IA vuelve a resurgir de la mano de la creación de los llamados "sistemas expertos". Tras un nuevo periodo de desaceleración entre 1987 y 1993, conocido como "the second AI Winter", la IA se adentra en una especie de "nueva primavera" que dura hasta nuestros días, y cuyos impulsos fundamentales hay que buscar en el surgimiento de los llamados "agentes inteligentes" capaces de comunicarse con el ser humano en un lenguaje natural, el desarrollo de las "redes neuronales artificiales", el "Machine Learning" y, ya durante la segunda década del siglo XXI, las técnicas de entrenamiento de algoritmos y la "minería de datos" basada en las nuevas tecnologías de la comunicación (Foote, 2016).

Aunque la existencia de estadios de ralentización de la IA explica que no se desarrollaran marcos jurídicos durante las últimas décadas del siglo XX, resulta llamativo que la "nueva primavera" haya operado a la intemperie de cualquier paraguas normativo. Siguiendo la tesis de Rosa, creo que la desincronización entre el proceso de creación de derecho y el proceso de innovación tecnológica en materia de IA no es una mera casualidad, sino algo que *tenía que ocurrir* dada la diferencia de los ritmos temporales de cada esfera. Esta tesis, sin embargo, no debería conducir a la asunción de una especie de filosofía de la historia de carácter negativo o catastrofista. Que la desincronización sea más o menos inevitable –como hemos visto más arriba, normalmente no ocurre que los fenómenos necesitados de regulación jurídica surjan al mismo tiempo que dicha regulación– no significa que la sociedad esté conducida necesariamente a una patología social, pues el rasgo patológico solamente se da cuando el grado de diferencia en los ritmos temporales es tan alto, que nuevas esferas más rápidas (mercado) aparecen para sustituir a otras menos rápidas (legislación) en las tareas de "regulación".

La diferencia de los ritmos temporales en las esferas de la legislación y las innovaciones tecnológicas es algo que podemos observar sin muchas dificultades. El progreso tecnológico posee unas dinámicas temporales concretas, que dependen de factores tan diversos como el éxito en la formación educativa de talentos, las inversiones de los sectores público y privado o el cambio en las exigencias de la industria. Por su parte, el proceso de creación de derecho posee también su propia dinámica temporal. Antes de que un parlamento pueda adentrarse en el proceso de legislación sobre un determinado asunto han de ocurrir muchas cosas. En primer lugar, hace falta que exista en la sociedad una conciencia suficiente sobre la relevancia de dicho asunto; es decir, hace falta que la opinión pública *tematice* ese asunto

como algo que *debería* ser regulado jurídicamente. En segundo lugar, hace falta desde luego que los legisladores dispongan de los conocimientos técnicos suficientes como para tomar decisiones suficientemente fundadas, lo cual exige, por ejemplo, la creación de comisiones de expertos. Por último, es preciso ponerse de acuerdo sobre el contenido y el alcance de dicha legislación[9].

Estos tres procesos no son algo que pueda acelerarse sin más. Uno no puede acortar los tiempos necesarios para que la opinión pública adquiera la conciencia de que un determinado fenómeno es lo suficientemente importante como para ser regulado jurídicamente. De la misma manera, tampoco puede acelerar el proceso a través del cual los representantes de la soberanía, que se supone han de haber dispuesto del tiempo suficiente para informarse del fenómeno a tratar, se ponen de acuerdo sobre el contenido de una ley. El intercambio de argumentos, la elevación de pretensiones de validez que han de ser confrontadas con una refutación racional por parte de todos los afectados —o por sus representantes democráticamente elegidos— no es algo que pueda hacerse a una velocidad ilimitadamente creciente.

Cuando la velocidad de las transformaciones tecnológicas es tan apabullantemente superior a la velocidad de los procesos de creación de derecho, la desincronización ocasionada puede tener como consecuencia una progresiva migración de las tareas supuestamente propias de la segunda esfera hacia terceras esferas más rápidas. Cuando ello ocurre, entonces ya no son las leyes creadas por un parlamento soberano, sino las leyes de la oferta y la demanda, las leyes de la revalorización del capital de las grandes multinacionales tecnológicas, las que "regulan", por así decirlo, los avances.

Ciertamente, podríamos pensar que una posible solución para este problema residiría en que esta migración no fuera hacia la esfera del mercado, sino hacia una cuarta esfera que, tal vez, podría asumir tareas regulatorias a una velocidad mayor que la de la esfera legislativa. Esta cuarta esfera podría ser la de una sociedad civil éticamente articulada, en la que podríamos incluir también la propia esfera mercantil (Conill, 2013; García-Marzá, 2013)[10]. La auto-obligación que nos ofrecen las consideraciones morales es absolutamente central también en el terreno de la IA, y existen razones para pensar que este proceso podría resultar más ágil que el de la regulación jurídica. Ahora bien, aunque la ética ofrece herramientas fundamentales para abordar el problema de las consecuencias indeseables derivadas de la IA, sobre todo porque la propia regulación jurídica está inspirada por valores y principios de naturaleza moral, creo que sería un error considerar que esta esfera puede sustituir sin más a la esfera legislativa en este terreno. En primer lugar, es evidente, como ha mostrado Jürgen Habermas (2010), que la institucionalización de normas jurídicas es esencial a la hora de compensar los déficits cognitivos, organizacionales y motivacionales de la moral postconvencional. En segundo lugar, creo que la "mayor velocidad" de la ética sobre el derecho no es algo que podamos dar por sentado sin más. Desde luego, la regulación ética "ahorra tiempo" en lo que hace al acto de aprobación de normas. No obstante, comparte con la legislación el resto de condiciones que hacen de ésta un proceso más lento que el de las

9 Me baso aquí en el modelo de "democracia de doble vía" desarrollado por Jürgen Habermas. Según este modelo, la legislación acontecida en el interior del Parlamento es solamente el último estadio de una secuencia que se inicia ya con los procesos de deliberación acontecidos en la sociedad civil. Véase Habermas, 2010.

10 Agradezco muy especialmente a José Luís López la estimulante discusión a propósito de las virtualidades y límites de esta posible salida al problema de la desincronización. Para su posición, véase López, 2022.

innovaciones tecnológicas: también en este caso hace falta que la opinión pública *tematice* el asunto como algo que *debería* ser regulado, y también es preciso que las personas que lo tematizan dispongan de los conocimientos técnicos suficientes como para tomar decisiones fundadas. La deliberación en el terreno moral, igual que la deliberación en el terreno jurídico, no es algo que pueda acelerare de forma ilimitada.

Conclusiones

La UE ha desarrollado un marco jurídico sobre IA sin duda garantista y comprometido con los valores, principios y derechos fundamentales sobre los que ella misma se asienta. Aunque este admirable proyecto hace de Europa el centro de una IA confiable y éticamente articulada, lo cual sin duda supone una "ventaja competitiva" (!) con respecto a otras regiones del planeta, lo cierto es que ha llegado más tarde de lo que hubiera sido deseable, y amenaza con volverse obsoleto a la misma velocidad a la que avanzan las vertiginosas innovaciones de la IA. Ciertamente, es preferible contar con una regulación ético-jurídica *lenta* que no contar con ninguna regulación *en absoluto*. Pero reconocer la necesidad de regulación no significa adoptar una postura reaccionaria sobre el objeto a regular. Negar las virtualidades de la IA resulta hoy una postura insostenible. No obstante, es preciso reconocer al mismo tiempo la existencia de peligros presentes y potenciales, que de ninguna manera pueden quedar sustraídos a una orientación normativa. Tienen razón, a mi modo de ver, aquellos que tachan de "neo-luddismo" las posturas que se oponen por principio al avance de la IA, permaneciendo insensibles a las muchas mejoras que su implementación puede traer a nuestras vidas. No la tienen, sin embargo, aquellos que aplican esta categoría despectiva a cualquier postura que considera necesaria la reflexión crítica sobre la idoneidad, y no solamente sobre la viabilidad, de algunos de estos avances tecnológicos. Muchos de ellos presentan amenazas tan evidentes contra nuestros principios fundamentales, que solamente el derecho positivo puede ofrecer los instrumentos adecuados para su contención. Otros, sin embargo, no presentan riesgos manifiestos, aunque tampoco ventajas que justifiquen su implementación.

Pero nuestro *Zeitgeist*, en ocasiones tan naif en lo que hace a las bondades de las nuevas tecnologías, conduce hoy a una generalizada sensación de "tensión de lo posible", de acuerdo con la cual todo lo que técnicamente *puede* ser realizado, *ha de serlo* de hecho. No hace falta que un proyecto sea monstruosamente peligroso para que nos cuestionemos si es deseable llevarlo a la realidad. Basta con que sea, por ejemplo, innecesario o ridículo. En los foros de discusión sobre IA, es más o menos habitual escuchar a sus entusiastas defensores esgrimir ejemplos como el que sigue para probar sus virtudes. Existen hoy, se dice, algoritmos lo suficientemente inteligentes como para completar, por ejemplo, la "Sinfonía inacabada" de Schubert. O como para producir composiciones poéticas tan absolutamente parecidas a las de un Rilke o un Hölderlin, que ni siquiera un experto sería capaz de distinguirlas. La primera reacción del asistente a estas exuberantes apologías es, sin duda, de asombro y admiración. Uno se siente, por de pronto, abrumado ante la gigantesca capacidad de estos algoritmos. Pero pasados unos minutos, los primeros asombros del impresionado asistente van dando lugar a una sensación diferente. Uno puede ver que éstas no son, desde

luego, implementaciones "monstruosamente peligrosas". Pero se pregunta, con un talante cada vez menos asombrado y cada vez más suspicaz, y hasta diríase que cada vez *más triste*, en qué sentido un mundo sin Rilkes ni Hölderlins, pero con algoritmos capaces de producir cosas muy parecidas a las que ellos, desde su absoluta singularidad, desde su biografía absolutamente única, lograron producir, es un mundo *mejor*.

Referencias

Calvo, P. (2019). "Democracia algorítmica: consideraciones éticas sobre la *dataficación* de la esfera pública". *Revista del CLAD Reforma y Democracia*, 74, 5-30

Caparrini, F. S. (2018). "Breve historia de la Inteligencia Artificial". *Revista de Occidente*, 446/7, 19-33

Carta de los Derechos Fundamentales de la Unión Europea. Diario Oficial de la Unión Europea (07.06.2016)

Copeland, J. (1996). *Inteligencia Artificial*. Madrid: Alianza

Conill, J. (2013). *Horizontes de economía ética*: *Aristóteles, Adam Smith, Amartya Sen*. Madrid, Tecnos

Cortina, A. (2019). "Ética de la Inteligencia Artificial". *Anales de la Real Academia de Ciencias Morales y Políticas*, 96

Digitalización de la industria europea Aprovechar todas las ventajas de un mercado único digital. COM (2016) 180 final (19.04.2016)

Domingo, A. (2020). "Cuidado generativo y ciudadanía digital: confianza, pandemia y proximidad". *Corintios XIII*, 176, 101-125

Estrategia Europea para la IA. COM (2018) 237 final (25.04.2018)

European Parlamient resolution of 20 October 2020 on a framework of ethical aspects of artificial intelligence, robotics and related technologies 2020/2012(INL)

European Parliament Draft Report, Artificial Intelligence in criminal law and its use by the police and judicial authorities in criminal matters 2020/2016(INI)

European Parliament Draft Report, Artificial intelligence in education, culture, and the audiovisual sector 2020/2017(INI)

European Parliament final Report on Artificial Intelligence in a digital age 2020/2266(INI)

Foote, K. D. (2016). "A Brief History of Artificial Intelligence". *Dataversity*. 05.04.2016

García Marzá, D. (2013). "Democracia de doble vía: el no-lugar de la empresa en la sociedad civil". *Revista del CLAD Reforma y Democracia*, 57, 67-92

Generar confianza en la Inteligencia Artificial centrada en el ser humano. COM (2019) 168 final (08.04.2019)

Habermas, J. (2010). *Facticidad y validez*. Madrid: Trotta

Honneth, A. (2011). "Patologías de lo social: tradición y actualidad de la filosofía social". En Honneth, A. *La sociedad del desprecio*. Madrid: Trotta, 75-126

Jones, S. E. (2006). *Against Technology: From the Luddites to Neo-Luddism*. Routledge

Libro Blanco sobre la Inteligencia Artificial. Un enfoque europeo orientado a la excelencia y la confianza. COM (2020) 65 final (19-02-2020)

López, J. L. (2022). *La ética ante la cinética del turismo. Aportaciones desde la teoría crítica de la resonancia de Hartmut Rosa* Tesis doctoral. (Universitat Jaume I, 2022)

Lübbe, H. (1988). "Gegenwartsschrumpfung". En Backhaus, K. y Bonus, H. (Eds.). *Die Beschleunigungsfalle oder der Triumph der Schildkröte*. Stuttgart, Schäffer/ Pöschel, 129-164

Marx, K. (1970). *Contribución a la crítica de la economía política*. Madrid, Alberto Corazón Editor, Prefacio

Mittelstadt, B. D.; Alli, P.; Taddeo, M.; Wachter, S. y Floridi, L. (2016). "The ethics of algorithms. Mapping the debate". *Big Data & Society*, 1-21

Neuhouser, F. (2022). *Diagnosing Social Pathology: Rousseau, Hegel, Marx, and Durkheim*. Cambridge University Press

Ortega-Esquembre, C. (2021). *Habermas ante el siglo XXI*. Madrid: Tecnos

Plan Coordinado sobre la IA. COM (2018) 795 final (07.12.2018)

Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence act) and amending certain Union legislative acts. COM (2021) 206 final (21.04.2021)

Rosa, H. (2005). *Beschleunigung*. Suhrkamp: Frankfurt a. M.

Rosa, H. (2016a). "Kritik der Zeitverhältnisse; Beschleunigung und Entfremdung als Schlüsselbegriffe der Sozialkritik". En Jaeggi, R. & Wesche, T. (Eds.). *Was ist Kritik*. Frankfurt a.M., Suhrkamp, 2009, 23-54 Rosa, H. *Alienación y aceleració*n. Buenos Aires: Katz

Rosa, H. (2016). *Resonanz. Eine Soziologie der Weltbeziehung*. Suhrkamp

Snow, C. P. (2000). *Las dos culturas*. Buenos Aires: Ediciones Nueva Visión

Tsamados, A.; Aggarwal, N.; Cowls, J.; Morley, J.; Roberts, H.; Taddeo, M.; y Floridi, L. (2021). "The ethics of algorithms: key problems and solutions". *AI and Society*

Versión consolidada del Tratado de la Unión Europea. Diario Oficial de la Unión Europea (26.10.2012)

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 163-174
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.560931

El estudio de la polarización política como terapia académica*

The study of political polarization as academic therapy

*PEDRO JESÚS PÉREZ ZAFRILLA***

Resumen. Los estudios sobre polarización política revelan que la polarización es mayor entre las personas con mayor nivel de estudios, mientras que las personas con menos estudios son más tolerantes. Estos resultados perturbadores me llevan a realizar en este trabajo una terapia académica sobre dos ideas centrales en la reflexión académica: la idea de la educación como forjadora de una ciudadanía crítica y tolerante, así como la idea del sujeto racional y autoconsciente. Finalmente, aplicaré la terapia académica también sobre las metodologías de trabajo en la Universidad. Esta terapia académica contribuirá a fomentar la tolerancia en la academia.
Palabras clave: Polarización política, educación, tolerancia, estatus, terapia.

Abstract. Studies on political polarization reveal that polarization is greater among people with a higher level of education, while people with less education are more tolerant. These disturbing results lead me to undertake an academic therapy in this work on two central ideas in academic reflection: the idea of education as a forger of a critical and tolerant citizenship, as well as the idea of the rational and self-aware subject. Finally, I will also apply academic therapy to work methodologies at the University. This academic therapy will help foster tolerance in the Academia.
Keywords: Political polarization, education, tolerance, status, therapy.

Recibido: 15/03/2023. Aceptado: 04/06/2023.

* Este estudio se inserta en el Proyecto de Investigación Científica y Desarrollo "Ética cordial y Democracia ante los retos de la Inteligencia Artificial" PID2019-109078RB-C22 financiado por MCIN/ AEI /10.13039/501100011033 y en las actividades del grupo de investigación de excelencia PROMETEO CIPROM/2021/072, financiado por la Conselleria d'Innovació, Universitats, Ciència i Societat Digital de la Generalitat Valenciana.

** Profesor Titular de Filosofía Moral en el Departamento de Filosofía de la Universidad de Valencia. Correo electrónico: p.jesus.perez@uv.es. Mi trabajo actual se centra en la neuropolítica, el tribalismo político, la polarización política y la distorsión del debate público en el mundo digital. Publicaciones recientes: "La paradoja aristotélica: cómo los discursos expresivos animalizan el debate público", *Isegoría*, 67:e03, https://doi.org/10.3989/isegoria.2022.67.03; "El tribalismo digital, entre la furia y la farsa: pinchemos la burbuja de la polarización artificial en internet", *Opinião Pública*, 28 (1), 2022, 33-61.

Introducción

La polarización política es un fenómeno de gran actualidad. Ahora bien, profundizar en el estudio de la polarización política tiene una vertiente que resulta perturbadora: la polarización política guarda una relación directamente proporcional con el nivel educativo de los sujetos (Abramowitz 2010, 23). Es decir, los sujetos con un mayor nivel de estudios mantienen posiciones más polarizadas y son menos tolerantes con el diferente. En cambio, los sujetos con menos estudios son más moderados y están menos ideologizados.

Este es un dato inquietante que merece ser analizado. Por un lado, esta evidencia nos obliga a reconsiderar algunas de las convicciones más firmes asentadas en la academia, como es la idea de que la formación académica cultiva en la persona el pensamiento crítico y la tolerancia. Pero también, resulta necesario averiguar las causas de esa conexión entre formación académica y polarización. Para comprender qué lleva a las personas más instruidas a tener actitudes más polarizadas y cómo este hecho socava algunas de las ideas más extendidas en la academia, considero que el método más adecuado para proceder es acometer una terapia académica.

La terapia académica (no confundir con la ya asentada terapia filosófica) procede por analogía en el ámbito académico a las formas de terapia psicoanalítica y de terapia cognitiva en el nivel clínico. De la terapia freudiana (Freud 1989) la terapia académica adopta la estrategia de indagación en los elementos subyacentes que son ignorados (o, al menos, no considerados en su verdadera dimensión) y que emergen inadvertidamente en la vida cotidiana. Por otro lado, de la terapia cognitiva (Ellis 1980) adopta el método de cuestionamiento de ciertas ideas centrales asumidas por los sujetos para, a través de un análisis crítico, proceder a su revisión.

Empleo el método de la terapia académica por tres motivos: El primero es que esta indagación permitirá a la institución universitaria ser consciente de ciertas ideas erróneas en las que asienta su labor tradicionalmente. Segundo, porque la terapia académica arroja luz sobre un impulso latente en la conducta humana que ha sido opacado en la tradición filosófica: el afán de estatus. Finalmente, esa terapia académica ayudará a aceptar los hechos perturbadores que la propia reflexión académica, de forma inesperada, hace aflorar, como es el caso de esta conexión entre polarización política y nivel de estudios.

En primer lugar, explicaré, a modo de introducción, en qué consiste la polarización política. A continuación, emprenderé una terapia académica relativa a dos ideas fundamentales que la relación entre polarización política y nivel de estudios pone en cuestión: la idea de la educación como forjadora de una ciudadanía crítica y tolerante y la idea de sujeto racional y autoconsciente. A partir de los elementos aflorados mediante la terapia académica se articulará una hipótesis explicativa de la mayor polarización de las personas más instruidas. Para concluir, esta terapia académica se aplicará a las metodologías de trabajo en la Universidad.

1. Polarización política

En ciencia política la polarización política se entiende como la división de la sociedad en dos grupos ideológicos contrapuestos. Así, una sociedad está más polarizada cuando una

mayoría de sujetos se identifica más fuertemente con las posiciones de uno de los grupos frente a las posiciones del otro grupo. En cambio, la polarización es menor cuando no son mayoría los sujetos que se identifican fuertemente con las posiciones de uno de los bloques frente al adversario (Pérez Zafrilla 2020).

De esta forma, la polarización parte de una adhesión acrítica a un partido. Los sujetos compran el paquete ideológico de los partidos, renunciando a formarse un juicio propio sobre los diferentes asuntos. Pero una sociedad en la que los sujetos hacen primar su identidad política sobre su propia autonomía en el juicio es una sociedad en la que el diálogo con el diferente se hace imposible, al verlo como un enemigo. Esto deja el terreno abonado para que proliferen las burbujas informativas, lo que pone en riesgo la propia democracia (Herreras 2021, 124).

Por ese motivo, la polarización se convierte hoy en un tema de máxima actualidad al que resulta necesario hacer frente. Uno de los ámbitos desde el que podría afrontarse la polarización es el educativo. Sin embargo, como mostraré a continuación, el ámbito académico no parece ser un espacio en el que esta labor se esté acometiendo con éxito.

2. Tópicos para una terapia académica

2.1. La formación académica como cultivadora de tolerancia

El estudio de la polarización política ha evidenciado un primer fenómeno inquietante: resulta que los sujetos con estudios universitarios se identifican más con las posiciones de un partido y manifiestan un mayor rechazo hacia los oponentes ideológicos. En cambio, los sujetos con estudios básicos o medios son más tolerantes con el diferente y manifiestan una menor identificación ideológica con las posiciones de un partido (Pew Research Center 2016, 7).

Esta relación entre polarización política y nivel de estudios representa, sin duda, un hecho perturbador que socava las convicciones asumidas en la academia sobre la educación como un proceso que fomenta en las personas la autonomía y la civilidad. Efectivamente, a lo largo de la historia, desde la *paideia* de los griegos hasta las reformas pedagógicas actuales, pasando por la Ilustración y la propia idea alemana de *Bildung*, una tesis asumida es que la educación representa un proceso necesario para que el niño adquiera no solo los conocimientos sino también las virtudes necesarias para desarrollar una vida autónoma. La educación se ha considerado siempre como un proceso que proporciona al alumnado una mayor capacidad cognitiva y crítica, así como unas actitudes dirigidas a la apertura de mente, la humildad epistémica y el respeto al diferente. Así también, la carencia de educación se asocia a la ignorancia, la sumisión a autoridades heterónomas, la cerrazón y la intolerancia. A este respecto, la excepción la representa Schopenhauer. Él defiende que el intelecto constituye un elemento de jerarquía y distinción, que aísla a los individuos de sus disímiles en capacidad cognitiva. Así, la elevación en el nivel cognitivo, en lugar de permitir una mejor comprensión del diferente, crea una barrera entre las personas más instruidas y aquellas con menor nivel de estudios. Es la bondad del corazón, y no la mayor habilidad cognitiva, la que permite a la persona identificarse con el diferente. Pero será solo posible una unión moral,

nunca intelectual (Schopenhauer 2003, 147). En este sentido, parece que los estudios sobre polarización política dan la razón a Schopenhauer.

Una salida escapista a esta evidencia aportada por el estudio de la polarización sería interpretar los datos como una correlación casual fruto de la metodología empleada en las encuestas que miden la polarización. Sin embargo, esta es una salida escapista, propia del paciente de la terapia cognitiva que se mantiene aferrado a creencias centrales asumidas que distorsionan su conocimiento de la realidad. Por ese motivo, creo necesario someter a terapia académica esta relación asentada en la academia entre educación, civilidad y tolerancia. Esta creencia central en la academia puede ser sometida a terapia académica partiendo de estudios realizados en la ciencia política y la sociología. Estos explican por qué las personas con mayor nivel académico están más polarizadas.

Desde la sociología política se subraya, por un lado, que los sujetos con mayor nivel educativo poseen una mayor sofisticación ideológica (Abramowitz 2010, 23). Es decir, su mayor entrenamiento cognitivo les hace tener un mayor nivel de abstracción. Los individuos con estudios superiores pueden comprender mejor los argumentarios que apoyan las posiciones sobre los diferentes temas. Esto permite a los sujetos identificarse más estrechamente con los posicionamientos ideológicos de los partidos sobre los distintos asuntos y contraargumentar a favor de su partido al recibir evidencias que refutan sus posiciones. En cambio, los sujetos con menor nivel educativo, al tener un menor entrenamiento cognitivo, realizan un acercamiento a la política más liviano. Pueden tener una fuerte identificación emocional con un partido, pero al tener una menor sofisticación ideológica, carecen de las herramientas que les permitan enfrascarse en la batalla política.

Por otro lado, Jason Brennan (2018, 92) incide en la relación entre nivel educativo, y consumo de información política. El consumo de información política es mayor entre las personas con estudios universitarios que entre gente de menor nivel académico. La clave reside, como señala Brennan, en que el consumo de información política representa un factor de socialización entre la gente con estudios superiores. Es decir, las personas con estudios superiores se relacionan con gente similar, y un tema de conversación habitual ente la gente con estudios superiores es la actualidad política. Estar al día de la actualidad política es necesario para socializar en ámbitos como el académico. Pensemos también cómo un tema habitual de conversación entre la gente con estudios superiores son las series de las plataformas de *streaming*. En cambio, entre la gente con baja cualificación académica priman otros temas de conversación, como puede ser el fútbol o los *realities*. Esta es una tesis que desde la academia se puede interpretar como prejuiciosa o clasista, pero lo realmente clasista y prejuicioso es no ser consciente de esta diferencia de consumo televisivo. Porque es un prejuicio academicista pensar que todo el mundo se interesa por la política o que está al día de la actualidad política.

En realidad, la élite académica e intelectual vive en una burbuja informativa ajena a la realidad de una parte de la población. Para comprobarlo basta acudir a los datos recogidos por Brennan en su libro *Contra la democracia* relativos al nivel de información política de la ciudadanía. Entre amplias capas de la población el conocimiento de la política es escaso, aunque luego acudan a votar en las elecciones. Por ese motivo, su consumo de información política también es escaso, ya que conversar sobre política no es un factor de socialización

en su entorno. Esa gente se interesa más por otro tipo de contenidos que son los que representan un elemento de socialización en su ámbito.

Esta tesis se ve corroborada también por otros datos relativos al consumo de información política *online*:

Por un lado, podemos citar un interesante trabajo basado en el estudio del historial de búsqueda de usuarios de la red: de los 1,2 millones de historiales de búsqueda analizados, sólo 173.450 (14% del total) leyeron al menos 10 noticias de política, y sólo 50.383 (4% del total) habían leído dos artículos de opinión a lo largo de los tres meses del estudio (Flaxman, Goel y Rao 2016, 301).

Segundo, un reciente estudio del Pew Research Center sobre el consumo de información política en Twitter revela, por un lado, que esta red social es la que más se dedica a la política: 1/3 de los tuits que se publican abordan contenido político. Pero, sobre todo, el estudio arroja que el 70% de los tuits sobre política los publican o retuitean graduados universitarios (Pew Research Center 2022, 16). Es decir, el consumo de información política en Twitter se produce en gran medida entre universitarios. Por el contrario, la gente sin estudios universitarios sigue en Twitter mayoritariamente contenidos no políticos.

En esta misma línea, resulta también curioso que los youtubers o influencers más famosos a nivel mundial no hablan precisamente de política, sino de videojuegos, moda o viajes (Siurana, 2021). Estos datos nos dan una idea de cuánta gente realmente usa internet para informarse de política. Por ese motivo es un error pensar que todo el mundo se interesa por la política.

2.2. El sujeto racional y autoconsciente

Por su parte, la psicología evolucionista explica esa relación entre polarización política y nivel educativo a partir de una hipótesis señalada por autores como Haidt (2019, 122) o Mercier y Sperber (2017, 257): la búsqueda de estatus como origen del razonamiento. Esta es una hipótesis sugerente que pone en cuestión otra idea ampliamente asumida en la academia: la del ser humano como sujeto racional y autoconsciente.

Son muchos los pensadores que han defendido que el ser humano está sometido al influjo de pasiones o impulsos irracionales que invaden y someten inadvertidamente la consciencia individual. Ejemplos paradigmáticos son Maquiavelo, Hume, Spinoza, Nietzsche, Schopenhauer y Freud. Sin embargo, en filosofía ha predominado la idea del ser humano como sujeto racional, asentado en un yo autoconsciente. La facultad de la razón, desarrollada en sus diversas formas, desde el *lógos* griego hasta la razón cartesiana o la kantiana, será la característica que diferencie cualitativamente al hombre del resto de animales. De ahí que los autores que acentúen la dimensión inconsciente en el hombre frente al yo autoconsciente sean vistos tradicionalmente como "la cara B" de la filosofía. Algunos de ellos son incluso denominados despectivamente como "filósofos de la sospecha".

Pero será con el desarrollo de las neurociencias en las últimas décadas cuando recobren impulso el inconsciente y las emociones. Así, las teorías del procesamiento dual explican la cognición como fruto de dos procesos, uno intuitivo o inconsciente (sistema 1), y otro cognitivo o consciente (sistema 2). Es más, el problema hoy día es comprender cómo

funcionan realmente los procesos cognitivos conocidos como sistema 2, dado el fenómeno de los sesgos cognitivos (Mercier y Sperber 2017, 182). Es precisamente este elemento de los sesgos cognitivos el que hace necesario someter a una terapia académica la idea del yo racional y autoconsciente. Para ello me apoyaré en algunas evidencias presentadas desde la psicología que arrojan resultados inquietantes:

Hasta los años noventa del siglo pasado la psicología consideraba que la razón tiene como función propia mejorar la cognición individual, esto es, permitir al individuo conocer la realidad. En este punto la psicología no se apartaba de la visión aristotélica del hombre como animal con *lógos* que tiene una tendencia natural al conocimiento de la realidad (Aristóteles, 2000, 980b25). La teoría darwiniana de la evolución no cambiará este enfoque en gran medida, solo que desde entonces la razón es considerada una facultad surgida evolutivamente. Pero su función seguirá siendo la misma: mejorar la cognición individual para favorecer la supervivencia (Mercier y Sperber 2017, 179). Por ese motivo, los sesgos cognitivos serán vistos como deficiencias de la razón al realizar la función que le es propia.

Sin embargo, Haidt, con su propuesta del intuicionismo social y Hugo Mercier y Dan Sperber, con su teoría interaccionista, darán una explicación revolucionaria a los sesgos cognitivos, como el de confirmación, el de razonamiento motivado o el de endogrupo/exogrupo. Para estos autores, estos sesgos no son errores de programación de nuestra mente en su intento de conocer la realidad, sino respuestas desarrolladas evolutivamente para garantizar la supervivencia de los sujetos en el entorno social. Si la razón hubiera surgido para conocer la realidad, estaría sesgada a favor de la búsqueda de contraargumentos, para que así el juicio fuera lo más objetivo posible. Pero sucede lo contrario: el razonamiento está sesgado hacia una evaluación favorable de las evidencias que apoyan el punto de vista del sujeto, frente a las evidencias que hay en contra. Por ese motivo, mantienen estos autores que el razonamiento no surgió para conocer la realidad, sino para mantener la supervivencia dentro de un grupo buscando razones y evaluando las evidencias de un modo favorable al sujeto (Haidt 2019, 139; Mercier y Sperber 2017, 331). Porque evolutivamente tener un juicio objetivo puede resultar desastroso socialmente, tanto para el propio individuo como para su grupo.

Ahora bien, esta relectura de los sesgos cognitivos en clave evolutiva y grupalista socava la idea asumida en filosofía sobre el ser humano como sujeto autoconsciente y racional. Más concretamente, las propuestas de Haidt y Mercier y Sperber, en la línea de otros desarrollos de la psicología evolucionista (Malo 2021, 169), sacan a la luz un rasgo de la naturaleza humana prácticamente ignorado en filosofía: la propensión de los individuos a buscar y mantener un estatus en el grupo. Este elemento lo podemos rastrear en la historia de la filosofía en el deseo de gloria señalado por Hobbes en el *Leviatán* (Hobbes 2014, 102) y, de una forma especial en Rousseau, cuando, en el *Discurso sobre los orígenes y fundamentos de la desigualdad entre los hombres*, señala cómo en el estado de naturaleza los individuos monitorizaban el estatus propio y el ajeno en sus primeras celebraciones comunitarias (Rousseau 2000, 283-284). Sin embargo, el paradigma moderno del reconocimiento de la igual dignidad, en sus diversas formas (Taylor, Ricoeur, Honneth) opacará el elemento de la búsqueda de estatus como un impulso presente en la naturaleza humana. No obstante, en sociología sí se atiende a este elemento. Veblen desarrolla magistralmente esta idea en referencia al consumo en *Teoría de la clase ociosa*. Allí incluso dice que la propensión a la emulación "constituye un rasgo omnipenetrante de la naturaleza humana" (Veblen 1963, 114).

Este elemento del estatus aparece de forma clara en el experimento de conformidad de Asch. En él la percepción de una opinión mayoritaria (aunque falsa) y el miedo a quedar aislados del grupo, llevan a un 36% de los sujetos a manifestar una opinión contraria a su propia percepción de la realidad (Asch 1956, 10). De ahí que, según la interpretación común, estos sujetos subordinen su juicio objetivo al sentimiento de aceptación por parte del grupo (Ross, Bierbrauer y Hoffman 1982, 64).

Sin embargo, a mi parecer el escenario creado por Asch no presenta meramente a un sujeto racional acogotado por una opinión externa que inhibe cualquier oposición. Más bien ese experimento revela que los sujetos tienen una identidad más compleja que la supuesta por gran parte de la tradición filosófica. Los individuos tienen una identidad social y una necesidad de estima y reconocimiento por parte de los demás. El individuo no es meramente un yo racional que conoce la realidad. Es también un ser social que procura mantener un estatus dentro de su grupo. Por eso el individuo evalúa las consecuencias sociales que tendrá su expresión pública. Porque, como defienden Haidt, Mercier y Sperber, el razonamiento evolucionó en una dimensión social como forma de garantizar la supervivencia individual (y el estatus) dentro del grupo. En este sentido, es el estatus social que el sujeto desea mantener el que le lleva a expresar el juicio tribal que garantiza su estatus frente al juicio racional de objetividad fáctica.

Por supuesto, la idea de que el hombre es un ser social que forja su identidad por el reconocimiento de los demás ya está anticipada por Aristóteles o Hegel. Del mismo modo, la dimensión emocional de la razón la encontramos en la razón cordial de Cortina (2007, 161). Pero el matiz que abren Haidt, Mercier y Sperber con la dimensión social del razonamiento y la búsqueda de estatus es importante. Así como la razón cordial reconoce situaciones de injusticia (Cortina 2007, 192), la psicología evolucionista defiende que en el hombre existe un sentido latente que rastrea contextos en los que el propio estatus se ve amenazado y, en esas ocasiones, ese impulso emerge para llevar al sujeto a decir o hacer cosas que mantengan su estatus. Así, el impulso por el estatus arroja una nueva luz sobre fenómenos como la aporofobia (Pérez Zafrilla en prensa). En determinados contextos, la conducta aporófoba viene dada por un rechazo a aquello que pueda incomodar, amenazar o perturbar el estatus de las personas. Por ejemplo, en el caso de la persona que habla a sus amigos sobre su hija médica, pero calla sobre su hijo camarero, se evidencia cómo la persona reconoce un contexto en el que su estatus está siendo monitorizado. Justamente porque la persona sabe que revelar que su hijo es pobre es algo que rebajará su estatus, la persona calla sobre su hijo camarero y sólo habla sobre su hija médica, como forma de mantener el estatus ante sus amistades.

Así pues, esta hipótesis de la búsqueda de estatus de Haidt, Mercier y Sperber como origen del razonamiento representa una terapia académica que pone en cuestión la idea de sujeto racional y autoconsciente, asumida en gran medida por la academia. La terapia académica hace aflorar la búsqueda de estatus como un motor latente de la conducta personal, ya que mantener el estatus era un requisito necesario para mantener la supervivencia en los grupos a lo largo de la evolución.

Ahora bien, ¿y qué relación guarda esta búsqueda de estatus con la conexión entre polarización política y el nivel educativo de los sujetos? Pues que el nivel educativo de los sujetos, en lugar de constituir un elemento indicativo de racionalidad, autonomía individual, objetividad y mesura en el juicio, representa un acelerador de la búsqueda de estatus. Las

personas con mayor estatus socioeconómico y de mayor nivel de estudios están más preocupadas por su estatus que las personas con menor nivel socioeconómico. Porque, como señala Henderson (2021), las personas de mayor nivel socioeconómico son las que más tienen que perder con la pérdida de estatus dentro del grupo. Esta situación hace que las personas de mayor nivel educativo, como decía en la sección anterior, empleen sus mayores habilidades cognitivas para encontrar argumentos que refuercen su posición y refuten la del adversario. Esto explica que las personas de mayor nivel académico (y con mayor estatus) evalúen las evidencias de una forma más sesgada que las que tienen menos estudios y un estatus más bajo. Porque un mayor estatus que mantener gracias al mayor nivel educativo hace que las personas se empleen con mayor vehemencia en la defensa de sus posiciones como forma de mantener su estatus en su grupo. En cambio, reconocer que el adversario tiene razón en un punto o admitir un error en el propio planteamiento conllevará el rechazo del propio grupo y, por ende, una pérdida de estatus, como el planteamiento de Haidt, Sperber y Mercier pone de manifiesto. Por todo ello, las personas de mayor nivel de estudios emplean sus habilidades cognitivas para polarizar (esto es, reforzar) su posición, en lugar de para abrirse a la comprensión del otro y admitir los propios errores. Esta es una táctica dirigida a mantener el estatus.

De la terapia académica realizada a estas ideas fuertemente asentadas en la tradición filosófica emerge la búsqueda de estatus como un elemento presente en la naturaleza humana que estaba en gran medida olvidado por las diferentes corrientes filosóficas. Otro aspecto que revela la terapia académica, como se refleja en el experimento de Asch, es que, en ocasiones, en el contexto social surge un conflicto entre la búsqueda de la verdad y el afán de estatus. Este es un conflicto que, por desgracia, se produce también en la academia, y a él dedicaré la última sección.

3. La investigación científica, a terapia

La academia tiene como uno de sus objetivos el conocimiento de la realidad, o, dicho de otro modo, la búsqueda de la verdad. Pero esa búsqueda no se produce de una forma objetiva. Los científicos, en su búsqueda de la verdad, también tienen sesgos. Buena muestra de ello da Kuhn en *La estructura de las revoluciones científicas*. Los científicos (y por extensión los académicos) buscan la verdad desde la asunción de un paradigma teórico. Así, cuando surge un paradigma nuevo que pone sobre la mesa objeciones difíciles de rebatir, los científicos, lejos de actuar humildemente y reconocer su error, actúan desde el sesgo de confirmación restando validez a las objeciones presentadas, y desde el sesgo de razonamiento motivado, buscando nuevas teorías que refuten las objeciones y confirmen el propio paradigma ahora cuestionado por otros (Kuhn 2001, 53).

Sin embargo, la ciencia se caracteriza por tres elementos, apuntados por Haidt y Lukianoff, que le permiten emplear los sesgos cognitivos a su favor. El primero es el hecho de que los científicos, mediante la discusión y la falsación de teorías, pueden anularse mutuamente sus sesgos. El segundo elemento es que, para que pueda producirse esa anulación mutua de los sesgos cognitivos en el mundo académico, es condición necesaria que exista un ambiente en el que reine la libertad investigadora y el pluralismo de puntos de vista. El tercer elemento

es que, en la ciencia, la búsqueda de la verdad y el estatus, lejos de estar separados, van de la mano. En la ciencia, el estatus lo da precisamente la búsqueda de la verdad: aquella teoría que es contrastada con evidencias y explica mejor la realidad gana en verdad, pero también en estatus, al desplazar a la teoría perdedora (Haidt y Lukianoff 2019, 174).

Por ese motivo, precisamente, el problema surge cuando alguno de estos pilares de la investigación científica falla. Cuando el ecosistema dentro de un ámbito no es plural, sino que prima la homogeneidad de perspectivas, se quiebra también la unión de búsqueda de la verdad y el estatus. En ese contexto, la búsqueda de la verdad se lleva a cabo desde el sesgo compartido de confirmación dentro de una determinada perspectiva, que se refuerza más y más, produciéndose la conocida polarización de grupo de Sunstein (2002, 179). Esta falta de pluralidad no pasa de ser una circunstancia azarosa que merma el progreso de la investigación si en el ámbito académico todos se guían por la búsqueda de la verdad. Si todos los científicos están convencidos de un paradigma, ese paradigma seguirá reforzándose en la academia mientras no surja una idea alternativa.

El problema surge cuando en ese ámbito homogéneo el estatus se asocia precisamente a la defensa del paradigma dominante. Cuando los miembros de un ámbito saben que su estatus depende de defender el paradigma reinante, se produce la quiebra de la relación entre estatus y búsqueda de la verdad. Esto se manifiesta de determinadas formas: se prima la investigación sobre temas que refuerzan el paradigma dominante y se relegan otros temas que puedan llevar a cuestionar el paradigma; sólo se contemplan hipótesis explicativas acordes a los esquemas del paradigma dominante, desechando otras alternativas; se destacan los resultados coincidentes con el paradigma y se oscurecen otros que lo puedan cuestionar; se enmarcan los hechos de un modo acorde a ese paradigma para así lograr su publicación en revistas más prestigiosas, sabedores de que los evaluadores y editores apoyan el paradigma dominante y rechazarán más fácilmente los estudios que rebatan ese paradigma… Todo ello hace que los resultados de las investigaciones realizadas y publicadas refuercen el paradigma dominante (Clark y Winegard 2020, 18).[1]

El punto clave está en que en todos estos comportamientos late la escisión entre búsqueda de estatus y búsqueda de la verdad. Los académicos, como personas de alto nivel educativo, son conocedores de que, en un ecosistema homogéneo, su estatus como investigadores depende de la defensa del paradigma imperante y que apartarse del mismo les hará caer en desgracia (Malo 2021, 318). El estatus se adquiere defendiendo el paradigma dominante, ya que es el método más fácil para ser citado, y las citas son hoy el valor que marca el estatus en la academia, a un doble nivel: por un lado, porque el hecho de ser citado se considera un factor objetivo de la obtención de estatus. Esta medición del estatus por las citas asimila la

1 En esta línea, un reciente estudio que analiza artículos de diversas disciplinas muestra que no hay una relación entre artículos más citados y replicabilidad del experimento (Serra-García y Gneezy, 2021). Es decir, los estudios más citados son precisamente los que no se replican. En cambio, los trabajos que se replican apenas son citados. Esto muestra que un sesgo presente en la academia es que se citan aquellos artículos que refuerzan el paradigma, por el hecho de que lo refuerzan, aunque no sean replicables, ya que ello permitirá que al autor del nuevo trabajo se le citará también. En cambio, los artículos que pueden ser replicados, si contradicen el paradigma, son condenados al olvido. Esta situación da lugar a sesgos académicos como el sesgo de publicación o el de citación (publicar o citar sólo estudios que confirman el paradigma), o el de subutilización de evidencias (desechar evidencias que contradicen el paradigma) (Leng y Leng, 2020).

academia a las redes sociales con los *likes* (Siurana 2021, 228). Pero existe una diferencia: de forma paradójica las citas negativas (las críticas) cuentan también como *likes*. Por otro lado, las citas se emplean para determinar el índice de impacto de las revistas de prestigio, entendiendo que las citas son un criterio objetivo de calidad de esas publicaciones, cuestión esta también discutible (Feenstra y Pallarés-Domínguez 2021). En todo caso, publicar en revistas posicionadas en un cuartil elevado (medido por el número de citas) es el factor principal que se sigue para la asignación de acreditaciones y de sexenios de investigación, siendo la obtención de acreditaciones y sexenios elementos imprescindibles para progresar en la carrera académica y, con ello, para la obtención de estatus dentro de la Universidad. De esta forma, para ser citado y obtener el estatus académico derivado de las citas (acreditaciones, sexenios), el elemento necesario es publicar en revistas de prestigio en las que, evidentemente, será más fácil publicar si se siguen unos parámetros establecidos por ese paradigma hegemónico y que todos los investigadores conocen. Por ese motivo, la academia es un ámbito en el que es fácil que, como en el experimento de Asch, el mantenimiento del estatus se anteponga en ocasiones a la búsqueda de la verdad, corrompiéndose así la institución académica. Porque en el contexto actual anteponer la verdad al estatus tiene consecuencias nefastas. Buena prueba de ello son los casos de linchamientos y cancelaciones a profesores en las universidades, sobre todo americanas (Haidt y Lukianoff 2019, 165). Esta realidad representa una verdad dolorosa que es necesario reconocer y asumir como terapia en el mundo académico.

4. Conclusión:

La terapia académica realizada en las páginas precedentes ha sacado a la luz realidades que contradicen algunas de las ideas más firmes asentadas en la academia. El elemento principal es la búsqueda de estatus como un motor latente del razonamiento y la acción de los sujetos, especialmente entre aquellos de mayor estatus socioeconómico. Este fenómeno explica por qué un mayor nivel educativo se correlaciona con una menor tolerancia ideológica. Pero también permite comprender cómo los sesgos presentes en la cognición humana socavan, al menos en parte, la idea de sujeto racional.

Reconocer y asumir estas realidades en el mundo académico representa una terapia necesaria. Solo así será posible encontrar vías para que las mayores habilidades cognitivas desarrolladas en los sujetos y los valores difundidos en la educación, como la tolerancia, la crítica y la apertura de mente a nuevas perspectivas, ayuden a los sujetos instruidos a ser realmente más tolerantes y humildes.

Referencias

Abramowitz, A. (2010). *The disappearing center: Engaged citizens, polarization and American democracy.* Yale: Yale University Press.

Aristóteles (2000). *Metafísica*. Madrid: Gredos.

Asch, S. E. (1956). Studies of Independence and Conformity: I A Minority of One against a Unanimous Majority. *Psychological Monographs: General and Applied*, *70* (9), 1-70.

Brennan, J. (2018). *Contra la democracia*. Madrid: Deusto.
Cortina, A. (2007). *Ética de la razón cordial*. Oviedo: Nobel.
Cortina, A. (2013). *Para qué sirve realmente la ética*. Barcelona: Paidós.
Clark, C. & Winegard, B. M. (2020). Tribalism in war and peace: The nature and evolution of ideological epistemology and its significance for modern social science. *Psychological Inquiry*, *31*, 1-22.
Ellis, A. (1980). *Razón y emoción en psicoterapia*. Zarauz: Itxaropena.
Feenstra, R. y Pallarés-Domínguez, D. (2021). Las dimensiones éticas de los sistemas de valoración y difusión científica en el área de filosofía moral. *Daimon. Revista Internacional de Filosofía*, *83*, 37-55.
Flaxman, S.; Goel, S. y Rao J. M. (2016). Filter bubbles, echo chambers and online news consumption. *Public Opinion Quarterly*, *80*, 298-320.
Freud, S. (1989). *Psicopatología de la vida cotidiana*. Madrid: Alianza.
Haidt, J. (2019). *La mente de los justos*. Barcelona: Deusto.
Haidt, J. y Lukianoff, G. (2019). *La transformación de la mente moderna*. Barcelona: Deusto.
Henderson, R. (2021). Persuasion and Prestige paradox: Are high status people more likely to lie? *Quillette*, 3 April. Accesible en: https://quillette.com/2021/04/03/persuasion-and-the-prestige-paradox-are-high-status-people-more-likely-to-lie/ Consultado el 5 de julio de 2022.
Herreras, E. (2021*). Lo que la posverdad esconde. Medios de comunicación y crisis de la democracia*. Barcelona: MRA.
Hobbes, J. (2014). *Leviatán*. México: F.C.E.
Kuhn, T. (2001). *La estructura de las revoluciones científicas*. México: F.C.E.
Leng, G. y Leng, R. I. (2020). Unintended consequences: The perils of publication and citation bias, *The MIT Press Reader*, 15th September. Disponible en:
https://thereader.mitpress.mit.edu/perils-of-publication-and-citation-bias/. Acceso: 17 de octubre de 2022.
Malo, P. (2021). *Los peligros de la moralidad. Por qué la moral es una amenaza para las sociedades del siglo XXI*. Barcelona: Deusto.
Mercier, H. y Sperber, D. (2017). *The enigma of reason: a new theory of human understanding*. London: Allen Lane.
Pérez Zafrilla, P. J. (2020). "Polarización política: estado de la cuestión y orientaciones para el análisis". En C. Santibáñez (Ed.). *Emociones, argumentación y argumentos*. Lima: Palestra, 97-124.
Pérez Zafrilla, P. J. (en prensa). "El reverso de la aporofobia: la protección del estatus como patología social", *Daimon. Revista Internacional de Filosofía*. Disponible en: https://revistas.um.es/daimon/libraryFiles/downloadPublic/12151
Pew Research Center (2016). "A wider ideological gap between more and less educated adults". April 2016. Accesible en: https://www.pewresearch.org/politics/2016/04/26/a-wider-ideological-gap-between-more-and-less-educated-adults/. Consultado el 13 de julio de 2022.
Pew Research Center (2022). "Politics on Twitter: One-Third of Tweets From U.S. Adults Are Political". June 2022. Accesible en: https://www.pewresearch.org/politics/2022/06/16/

politics-on-twitter-one-third-of-tweets-from-u-s-adults-are-political/. Accedido en 19 de junio de 2022.

Ross, L., Bierbauer, G. y Hoffman, S. (1982). El papel de los procesos de atribución en la conformidad y el disentimiento: Reencontrando la situación de Asch. *Estudios de Psicología*, *10*, 63-78.

Rousseau, J. J. (2000). "Discurso sobre el origen y los fundamentos de la desigualdad entre los hombres". En *Del Contrato social. Discursos*. (pp.229-316). Madrid: Alianza.

Schopenhauer, A. (2003). *El mundo como voluntad y representación*, vol.II. Madrid: F.C.E.

Serra-García, M. y Gneezy, U. (2021). Non-replicable publications are cited more than replicable ones", *Science Advances*, 7 (21), eabd1705.

Siurana, J. C. (2021). *Ética para influencers*. Madrid: Plaza y Valdés.

Sunstein, C. (2002). The Law of Group Polarization. *The Journal of Political Philosophy*, *10* (2), 175-195.

RESEÑAS

Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 177-209
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico)

COECKELBERGH, M. (2021). *Ética de la inteligencia artificial*. Madrid: Cátedra (183 pp.)

La admiración y el miedo, respuestas humanas muy comunes ante los avances tecnológicos y presentes en diferentes contextos culturales, residen en el corazón de los futuristas y los catastrofistas. Por ello, el ejercicio que demanda el análisis de estos comportamientos debe ponerse en manos de un constante cuestionamiento ético. Y es que no son tanto las ideas futuristas las que se escapan de esta conveniente reflexión, sino que son aquellas que se encuentran en los lugares más comunes, carentes de todo atisbo de duda, las que resultan más difíciles de cuestionar con detenimiento, mirada austera y amplitud de miras. De este modo, nos encontramos en dominios en donde se asumen ciertos presupuestos sin someterse a juicio, lo que resulta ser un terreno muy fructífero para la mirada inquisidora de quien pretende hacer filosofía. Y es este papel de vocación interrogadora el que desempeña Mark Coeckelbergh en las páginas de *Ética de la inteligencia artificial,* para conducirnos hacia la duda y constante interrogación, es decir, para enseñarnos *«que lo que pensamos los seres humanos tiene también varias caras»* (Coeckelbergh, 2021).

El presente ensayo dedica sus esfuerzos al propósito de presentar de forma completa, y sistemática, el estado de la cuestión de las problemáticas éticas que se derivan de las tecnologías de inteligencia artificial (IA) en diferentes disciplinas, integrando y agregando claridad al trabajo en este campo. La evaluación de los posibles avances en su desarrollo ha producido una cierta desconfianza que ha llevado al autor a la necesidad de repensar y discutir los problemas éticos y sociales que suscitan estos monstruos del apocalipsis. El autor los define de este modo haciendo referencia al monstruo de Frankenstein, una figura que aporta una valiosa mirada acerca de la ambivalente posición de estas nuevas tecnologías.

Dos de los temas más recurrentes del ensayo y que producen mucha inquietud académica y social son la superinteligencia y el transhumanismo. Quizás, resuenen mucho en la actualidad, pero para sorpresa de muchos desde bien temprano el ser humano sueña con alcanzar el poder de un dios y, hacerse con atributos divinos tales como la inmortalidad. Esta aspiración se ha trasladado a la medicina actual en forma de medidas integrales contra el envejecimiento. Su origen no es otro que el miedo producido, por la fragilidad humana y su capacidad de cometer errores, lo que nos ha llevado a mejorar al ser humano hasta llegar a comprenderlo como un *Homo deus*: humanos que han ascendido a la categoría de dioses (Harari, 2016). Muchos piensan que la biología humana necesita mejorarse para no arriesgarnos a quedar como «la parte lenta y cada vez más ineficiente de la IA (Armstrong, 2014)».

Actualmente, tal como menciona Coeckelbergh, existen dos caminos hacia la superinteligencia. Uno consiste en que la IA se automejore continuamente, diseñando y mejorando versiones de sí misma, que, a su

vez, diseñarían versiones mejoradas exponencialmente: un acontecimiento llamado La Singularidad Tecnológica, que consiste, en el escenario en el que los seres humanos serán superados por máquinas inteligentes o inteligencias, o inteligencias biologías cognitivamente mejoradas o ambas.

Si bien tales avances pueden resultar escalofriantes, por suerte para algunos o para desgracia de otros, los avances en neurociencia todavía nos sitúan lejos de estas posibles aplicaciones. Y, aunque pueda relajarnos, no significa que debamos permanecer exentos de preocupaciones. Existen numerosas aplicaciones de software inteligente que nos acompañan en nuestra vida cotidiana. Y son el origen de múltiples problemáticas que afectan a la forma en la que estamos y convivimos en el mundo. Una de las áreas que más gravemente se ha visto afectada recientemente y que hemos dejado pasar por alto, además de que se haya instaurado como modelo económico, han sido los datos, datos que forman parte de nuestra privacidad y que, como bien define Carissa Véliz (2022) «la realidad es tan colectiva como es personal». Y que como define Shoshana Zuboff (2021), el capitalismo tiene el poder de impulsar la difusión de la conexión y la vigilancia en línea, de modo que los espacios sociales que antes permanecían dentro de un círculo social cerrado, ahora se están abriendo a corporaciones dirigidas a la obtención de beneficios y o la regulación de conductas. Este último fenómeno conlleva el riesgo de que los usuarios puedan ser manipulados y explotados. La IA tiene múltiples aplicaciones y, puede ser utilizada para manipular nuestras necesidades y recoger datos muy personales acerca de nosotros, es decir, creencias religiosas, ideologías políticas u, orientación sexual o datos referentes a la salud. Recabar este tipo de datos, en muchos países se considera un delito, ya que permanecen altamente protegidos jurídicamente. Y dentro del contexto capitalista, la herramienta más utilizada para recabar datos es a través del uso de las redes sociales. «Este tipo de tecnologías de aprendizaje automático y ciencias de datos puede conducir a nuevas formas de manipulación, vigilancia y totalitarismo, no necesariamente bajo la apariencia de regímenes autoritarios, sino de manera subrepticia y altamente efectiva» (Coeckelbergh, 2021, p.88). Las múltiples aplicaciones de la IA introducen numerosas problemáticas que implican graves injusticias, como es el caso de los sesgos algorítmicos, a la hora de conceder una ayuda financiera, calcular la factura de la luz, predecir la reincidencia de un preso, etc. Nuestro éxito para enfrentarnos con los grandes problemas de este periodo como predice Coeckelbergh, viene determinado por «una combinación de inteligencia abstracta (humana y artificial) y sabiduría pragmática concreta desarrollada sobre la base de experiencia humana situacional y práctica (incluyendo nuestra experiencia con la tecnología)» (2021).

Esta obra combina de forma satisfactoria un lenguaje accesible para el público general que pretenda conocer las inquietudes y problemáticas éticas que derivan de la IA con un tratamiento detallado de los diferentes debates que conforman este campo de investigación. En este sentido, profundiza de forma resumida en cada una de estas temáticas a través de doce capítulos breves, cada uno de los cuales se ocupa de desglosar minuciosamente distintas problemáticas. La presentación y exposición de las posiciones y argumentos es fluida y favorece una lectura ágil y clara. La forma en la que trata los temas que se discuten, y su pretensión de abordar reflexivamente una variedad de debates y discusiones de especial relevancia filosófica y social, dota al lector de una serie

de capacidades y herramientas para comprender de mejor forma los desafíos éticos planteados por la IA.

Por todo lo anterior, las páginas que conforman *Ética de la inteligencia artificial* se erigen, en definitiva, como ejemplo de cómo la ética y la ciencia poseen espacios de reflexión y pensamiento común y por qué es importante ahondar en ellos. Algo de agradecer habida cuenta de que necesitamos medios conceptuales para afrontar los nuevos retos de las sociedades tecnológicas.

Referencias

Armstrong, S. (2014). *Smarter Than Us: The Rise of Machine Intelligence*. Machine Intelligence Research Institute.

Coeckelbergh, M. (2021). *Ética de la inteligencia artificial*. Ediciones Cátedra.

Harari, Y. N. (2016). *Homo Deus: A Brief History of Tomorrow*. Random House.

Véliz, C. (2022*). Privacidad es poder: Datos, vigilancia y libertad en la era digital*. Penguin Random House Grupo Editorial.

Zuboff, S. (2021). *La era del capitalismo de la vigilancia: La lucha por un futuro humano frente a las nuevas fronteras del poder*. Ediciones Culturales Paidos.

Jorge Couceiro Monteagudo
(Universidad Complutense de Madrid. jorgecouceiromonteagudo@gmail.com)

GONZÁLEZ-ESTEBAN, ELSA y SIURANA, JUAN CARLOS (eds.) (2023). *Inteligencia Artificial: concepto, alcance y retos*. Valencia: Tirant Blanch.

El libro *Inteligencia Artificial: Concepto, Alcance, Retos*, editado por Elsa González y Juan Carlos Siurana y dividido en tres grandes bloques, es un trabajo focalizado en la creciente penetración de la Inteligencia Artificial (IA) en todos los aspectos de la vida humana y sus impactos sobre la sociedad *hiperdigitalizada* e *hiperconectada* y sus diferentes esferas funcionales, como la economía, la política, la comunicación o la investigación científica, entre otras.

Entre las múltiples aportaciones del libro, destaca su propuesta de un diseño institucional y una infraestructura ética que permita orientar la gestión y toma de decisiones en las organizaciones en un sentido responsable, justo y equitativo y garantizar la acción y supervisión humanas, entendidas éstas desde la ética del discurso como participación y búsqueda deliberativa de acuerdos por parte de todos los afectados.

El **Bloque I. Inteligencia Artificial: concepto** consta de cuatro capítulos escritos por Juan Arana, Jesús Conill, Dieter Sturma y Pedro Jesús Teruel respectivamente. Este bloque explora diferentes aspectos del concepto de la IA y su relación con la mente humana y la tecnología. En él se discuten los enfoques naturalistas de la mente humana, se analizan las críticas de los neurocientíficos al paradigma computacional y se exploran las posibilidades de la IA en diversos campos.

El primer capítulo, "Algunas críticas de los neurocientíficos al paradigma computa-

cional", presenta una discusión sobre las críticas que los neurocientíficos han hecho al paradigma computacional en la explicación de la mente humana. Entre otras cuestiones, se argumenta que la concepción naturalista de la mente humana defiende la tesis de que toda ella puede ser cabalmente explicada con los medios y procedimientos de la ciencia natural, ahora mismo o en un futuro previsible. Entre sus partidarios, destacan los representantes de la IA fuerte, los filósofos de la mente de orientación materialista y los neurocientíficos que intentan aclarar con el paradigma neuronal todos los aspectos relevantes del psiquismo. Sin embargo, se presentan críticas a esta concepción naturalista, argumentando que la mente humana es mucho más compleja de lo que se puede explicar con la IA y el paradigma computacional.

El segundo capítulo, "Inteligencia artificial e inteligencia corporal y sentiente (en perspectiva zubiriana)", aborda la relación entre la IA y la inteligencia corporal y sentiente desde una perspectiva zubiriana. En este capítulo se argumenta que la IA no puede ser considerada como una forma de inteligencia comparable a la inteligencia corporal y sentiente, ya que la IA carece de la capacidad de sentir y experimentar el mundo de la misma manera que lo hace un ser humano. Se presenta también la perspectiva zubiriana, que sostiene que la inteligencia corporal y sentiente es fundamental para la comprensión del mundo y la toma de decisiones, y que la IA no puede reemplazar esta forma de inteligencia.

El tercer capítulo, "La extensión de la mente. Inteligencia artificial, tecnología y ser humano como forma de vida", explora cómo la IA y la tecnología pueden ser consideradas una extensión de la mente humana y cómo esto afecta a nuestra forma de vida. Se presenta la idea de que la tecnología y la IA pueden ser vistas como una herramienta para mejorar nuestra calidad de vida y nuestra capacidad para comprender el mundo que nos rodea. Sin embargo, también se plantean preocupaciones sobre cómo la dependencia de la tecnología y la IA puede afectar a nuestra capacidad para pensar y tomar decisiones de manera autónoma.

El cuarto y último capítulo de este Bloque I, "Recrear el mundo. Inteligencia artificial y emociones: una ejercitación neurofilosófica", se centra en la relación entre la IA y las emociones, y cómo esto puede ser abordado desde una perspectiva neurofilosófica. En este capítulo, se argumenta que la IA puede ser programada para simular emociones humanas, pero que esto no significa que la IA realmente experimente emociones de la misma manera que lo hace un ser humano. Se presenta la perspectiva neurofilosófica, que sostiene que las emociones son una parte fundamental de la experiencia humana y que la IA no puede reemplazar esta experiencia emocional.

En resumen, el Bloque I. Inteligencia artificial: concepto proporciona una introducción sólida al concepto de la IA y su relación con la mente humana y la tecnología. Los cuatro capítulos presentan diferentes perspectivas y enfoques para abordar estos temas, desde las críticas de los neurocientíficos al paradigma computacional hasta la relación entre la IA y las emociones. En general, el Bloque I es una lectura interesante y esencial para cualquier persona interesada en comprender mejor el concepto de la IA y sus implicaciones para la sociedad y la humanidad en general.

El **Bloque II**. **Inteligencia Artificial: alcance** consta de cinco capítulos escritos por Domingo García-Marzá, Francisco Arenas Dolz, Alexander Kriebitz y Christoph Lütge, Javier Gracia y Pedro Jesús Pérez Zafrilla respectivamente. Esta sección

aborda diferentes temas relacionados con la IA y su impacto en la sociedad como son: la crítica de la razón algorítmica, la democracia en la era de la IA, las consecuencias de la IA para los derechos humanos, la ventaja de la IA en la educación y las conexiones entre la IA y la psicología evolucionista.

En el primer capítulo, "Crítica de la razón algorítmica: contra la neutralidad como ideología", se realiza una crítica a la razón algorítmica y su neutralidad como ideología. Entre otras cuestiones importantes, se argumenta que la razón algorítmica no es neutral, ya que está influenciada por los intereses de quienes la diseñan y utilizan, y puede promover y perpetuar la discriminación y la desigualdad en la sociedad. Por ello, se propone una reflexión crítica sobre la razón algorítmica y su impacto en la sociedad.

En el segundo capítulo, "Más allá del capitalismo de la vigilancia. La democracia en la era de la inteligencia artificial", se analiza el impacto de la IA en la democracia y la sociedad. En él se argumenta que la IA está transformando la sociedad y la economía, y que el capitalismo de la vigilancia es una amenaza para la democracia. Dadas estas circunstancias, se propone una reflexión crítica sobre la IA y su impacto en la democracia, así como la necesidad de una regulación ética de la IA.

En el tercer capítulo, "La inteligencia artificial y sus consecuencias para los derechos humanos: una evaluación desde la ética de la empresa" se examinan las consecuencias de la IA para los derechos humanos. Destaca al respecto cómo la IA puede tener un impacto negativo en los derechos humanos, como la privacidad, la libertad de expresión y la igualdad. Por lo cual, se propone una evaluación ética de la IA desde la perspectiva de la empresa, así como la necesidad de una regulación ética y legal de la IA capaz de proteger los derechos humanos.

En el cuarto capítulo, "Inteligencia artificial: ¿una ventaja para la educación?", se profundiza en el impacto de la IA en la educación. En él se subraya las ventajas y potencialidades que la IA puede tener para la educación, ya que ésta puede mejorar la calidad de la enseñanza y personalizar el aprendizaje. Por esta razón, se sugiere una reflexión crítica sobre la IA y su impacto en la educación, así como la necesidad de concretar una regulación ética y pedagógica de la IA en el ámbito educativo.

En el quinto capitulo, "Conexiones entre inteligencia artificial y psicología evolucionista", se examinan las vinculaciones existentes entre la IA y la psicología evolucionista. En este texto se señala que la IA puede ser una herramienta útil para entender la mente humana y la evolución de la inteligencia. Por consiguiente, se propone una reflexión crítica sobre las conexiones entre la IA y la psicología evolucionista, y se sugiere la necesidad de una colaboración interdisciplinaria entre la IA y la psicología evolucionista.

En conclusión, el Bloque II. Inteligencia Artificial: alcance, presenta una reflexión crítica sobre la IA y su impacto en la sociedad. A lo largo de los cinco capítulos los autores proponen una regulación ética y democrática de la IA para proteger los derechos humanos y promover una sociedad más responsable, justa y equitativa.

En el **Bloque III. Inteligencia Artificial: Retos** consta de cinco capítulos escritos por Agustín Domingo Moratalla, Philip Brey, Elsa González Esteban, Juan Carlos Siurana Aparisi, Francisco Fernández Beltrán y María José Senent Vidal y Asunción Ventura Franch respectivamente. En esta parte del libro se acomete una reflexión crítica sobre una serie de temas éticos relacio-

nados con la IA y su impacto en diferentes ámbitos de la sociedad. Destaca al respecto la preocupación actual sobre la necesidad de acometer un desarrollo de la IA que tenga en cuenta las implicaciones sociales y éticas de la tecnología. Además, los contenidos del bloque proponen diversas estrategias para garantizar que la IA se utilice de manera responsable, justa y equitativa.

En el primer capítulo, "El reto del discernimiento en la inteligencia artificial: tiempo, atención y apropiación", se reflexiona sobre los desafíos que plantea la IA en términos de nuestra capacidad para discernir y tomar decisiones informadas. Al respecto, se argumenta que la IA puede afectar a nuestra capacidad para prestar atención y tomar decisiones de manera autónoma, así como que es necesario desarrollar estrategias para fomentar el discernimiento y la apropiación crítica de la tecnología.

En el segundo capítulo, "Hacia una estrategia multilateral para la ética de la inteligencia artificial", se analiza el contexto digital para proponer una estrategia multilateral desde la cual abordar los desafíos éticos de la IA. Destaca al respecto la necesidad y posibilidad de establecer un marco ético común que permita a los diferentes actores involucrados en el desarrollo y la implementación de la IA trabajar juntos con el objetivo de garantizar que la tecnología se utilice de manera responsable y ética.

En el capítulo tercero, "Investigación e innovación en inteligencia artificial: responsabilidad y confianza", se reflexiona sobre la importancia de la responsabilidad y la confianza en la investigación y la innovación en IA. Entre las principales cuestiones, el capítulo acentúa la necesidad desarrollar una cultura de responsabilidad y confianza en la investigación y la innovación en IA para garantizar que la tecnología se utilice de manera sociablemente responsable y ética.

En el capítulo cuarto, "Ética para la aplicación de la inteligencia artificial a la mercadotecnia influyente", se profundiza en los principales desafíos éticos que plantea la aplicación de la IA a la mercadotecnia influyente. Al respecto, se argumenta que es necesario establecer un marco ético que permita a los profesionales de la mercadotecnia utilizar la IA de manera responsable y ética, y que se deben establecer límites claros para evitar el uso manipulativo de la tecnología.

En el capítulo quinto, "Implicaciones éticas del impacto de la inteligencia artificial en la comunicación pública", se analiza el impacto de la IA en la comunicación pública y los desafíos éticos que plantea. Entre otras cuestiones, el capítulo subraya la necesidad de desarrollar una ética de la comunicación pública que tenga en cuenta el impacto de la IA en la sociedad, así como la exigencia de establecer límites claros para evitar el uso manipulativo de la tecnología.

En el sexto y último capítulo del Bloque III, "Aplicación de la perspectiva de género a la regulación de los sistemas de inteligencia artificial", se profundiza en el papel que debe desarrollar la perspectiva de género en la regulación de los sistemas de IA. En el capítulo se remarca la necesidad de tener en cuenta las cuestiones de género en el desarrollo y la implementación de la IA. Asimismo, también se recalca la exigencia de establecer medidas para garantizar que la tecnología no reproduzca o refuerce las desigualdades de género existentes.

En conclusión, el Bloque III. Inteligencia Artificial: retos aborda una serie de temas éticos relacionados con la IA y su impacto en diferentes ámbitos de la sociedad. Los autores reflexionan sobre los desafíos que plantea la IA en términos de nuestra capacidad para discernir y tomar decisiones informadas, y proponen estrategias para fomentar el discernimiento y la

apropiación crítica de la tecnología. Por un lado, en el bloque se discute la importancia de establecer un marco ético común que permita a los diferentes actores involucrados en el desarrollo y la implementación de la IA trabajar juntos para garantizar que la tecnología se utilice de manera responsable y ética. También se reflexiona sobre la importancia de la responsabilidad y la confianza en la investigación y la innovación en IA y la necesidad de analizar y encontrar soluciones plausibles a los desafíos éticos que plantea la aplicación de la IA a la mercadotecnia influyente. Finalmente, se profundiza en los impactos de la IA en la comunicación pública y las posibles medidas a tomar para garantizar que la tecnología no reproduzca o refuerce las desigualdades de género existentes.

En resumen, este libro es una contribución valiosa al debate sobre la ética de la IA y su impacto en la sociedad y constituye una lectura esencial para cualquier persona interesada en comprender los desafíos éticos de la IA y en contribuir a un desarrollo tecnológico responsable, justo y equitativo.

Carlos Saura García
(Universitat Jaume I de Castellón)

GAMERO CABRERA, Isabel G. (2021): *La paradoja de Habermas. ¿Qué sucede cuando se aplica la teoría de la acción comunicativa a debates actuales?* Madrid: Dado Ediciones. 333 pp.

A través de la peculiar estructura que configura esta obra la autora, Isabel G. Gamero Cabrera, ofrece un acercamiento a la filosofía desarrollada y defendida por Jürgen Habermas. Valiéndose de un tono ameno y cercano, va presentando la teoría habermasiana acompañada de las problemáticas derivadas del aparato conceptual que el filosofo ha ido configurando, todo ello aludiendo a la actualidad de los temas abordados al tratar de aplicarlos en la compleja sociedad vigente. En este sentido, la autora viene a conformar una amplia imagen del pensamiento del filósofo desde un punto de vista crítico, atendiendo a las complejidades y carencias que ha logrado identificar, enfocando, al mismo tiempo, su mirada en las profundas crisis recientemente acontecidas.

Recurriendo a la organización propia de esta obra, se pueden identificar claramente dos partes que, de forma simétrica, se van complementando capítulo a capítulo, permitiendo al lector elegir entre diversos modos de lectura. Tal y como señala Isabel Gamero, se podría optar por una lectura tradicional, comenzando por la parte A, dedicada a la así denominada *racionalidad comunicativa,* finalizando por la B, dedicada a abordar las cuestiones relativas al *mundo de la vida*. De igual modo, se podría elegir hacerlo justo al contrario, comenzando por la parte B y finalizando por la A, o se podría optar por una tercera opción, elegida para la elaboración de esta reseña, consistente en hacerlo de forma paralela, avanzando paso a paso por ambas secciones (p. 9). Así, se abordará el primer capítulo de cada una de las mitades, dedicado en ambos casos a presentar el aparato conceptual fundamental, por un lado, introduciendo la noción de raciona-

lidad habermasiana como aquello que nos caracteriza a los seres humanos, diferenciándonos del resto de especies (p. 19). En este sentido, la autora expone la forma en la que, para Habermas, tiene una importancia clave el componente práctico, refiriéndose a la *racionalidad comunicativa* como una capacidad humana que necesita ser aprendida y mejorada socialmente. Teniendo en cuenta esta necesidad de entreno y perfeccionamiento, así como los imperfectos contextos sociales en los que nos instauramos, las formas de racionalidad predominantes serían las identificadas por el autor como imperfectas, exponiendo su *racionalidad comunicativa* como una "aspiración que se frustra con mucha frecuencia en la realidad cotidiana" (p. 23).

Por otro lado, en las páginas introductorias de la, así denominada, parte B, la autora explica el otro concepto clave que hilará esta sección del escrito, a saber, el *mundo de la vida*, que consistiría en "el sustrato común o base compartida por distintos hablantes que permite que se entiendan entre sí y puedan llegar a ponerse de acuerdo" (p. 165). Tal y como explica Isabel, Habermas utiliza este concepto de forma complementaria a su noción de racionalidad, evitando, de este modo, el exceso de abstracción, logrando acercar su propuesta filosófica a la realidad cotidiana "compartida y vivida por los hablantes" (p. 167). Siguiendo la caracterización inicial presentada, se puede identificar, a través de las nociones conjugadas, la visión conciliadora de Habermas que permeará toda su obra, así como el intento de solución de las dificultades comunicativas propias de las sociedades contemporáneas (p. 170).

Retrotrayéndonos de nuevo a la primera parte del escrito, localizaríamos el capítulo uno, donde la autora aborda las primeras formulaciones de la propuesta filosófica de Habermas, incidiendo en la génesis del concepto de *racionalidad comunicativa*, entendido primeramente como "la proyección de una situación ideal" (p. 29) inalcanzable en los imperfectos contextos sociales actuales. Siguiendo las aclaraciones presentadas, se entendería que la *racionalidad comunicativa* sirve como una guía para mejorar los procesos comunicativos vigentes. Pero la complejidad que envuelve este modelo se refleja en las críticas recogidas al final del capítulo, enfocadas especialmente en la cuestión de la dualidad aparente; por un lado, tendríamos el concepto de *racionalidad comunicativa* teórico-filosófico y, por otro, la aplicación práctica del mismo, acompañada de una fuerte connotación ético-crítica. De esta forma, Isabel Gamero nos ofrece una concisa explicación de las críticas que Habermas recibió en aquella primera etapa –dirigidas en su mayoría a un supuesto alejamiento con respecto a la realidad sociocultural en la que nos encontramos– haciéndose cargo, igualmente, de las propuestas resolutivas y aclarativas que el autor formuló. En el apartado complementario o conjugado, localizado en la parte B del libro, se encuentra la "primera comprensión" (p.171) o aproximación al *mundo de la vida*, incidiendo en la influencia que Habermas recoge de la propuesta filosófica desarrollada por Husserl, así como en su reformulación propia. Las críticas recibidas, en este sentido, comparten ciertas similitudes con las aportadas en lo referente a la *racionalidad comunicativa*, apuntando a las influencias que el autor recoge de la fenomenología y del idealismo kantiano. Justamente, por las cuestiones señaladas, en las siguientes obras se producen algunos cambios de carácter importante, aclarando que desde el principio se mostró crítico con el idealismo, a través de la exposición del carácter realista y práctico de su propuesta al

valerse del concepto de *sentido común* formulado por G. E. Moore y de las así denominadas *certezas* propuestas por el segundo Wittgenstein (p. 180). De forma fluida, la autora aborda estas cuestiones, así como el viraje indicado, en el segundo capítulo de esta misma sección (p. 181), exponiendo la adaptación que Habermas realiza aplicada a su nueva comprensión del *mundo de la vida*. Del mismo modo, encontramos recogidas las importantes dificultades propias de la reinterpretación que el filósofo realiza, la cual vendría a asumir una concepción de las certezas fija y fundacionalista, entendidas como "firmes, incuestionables y más que verdaderas que forman la base del mundo de la vida" (p. 197) alejándose de forma radical de lo que un autor como Wittgenstein entendería por certeza.

Siguiendo esta trayectoria, en el segundo capítulo de la parte A, Isabel Gamero expone las variaciones que el concepto habermasiano de *racionalidad comunicativa* va sufriendo, todas ellas vinculadas con una necesidad práctica de alejamiento con respecto a los horizontes ideales criticados. De este modo, la noción quedaría reubicada como la condición de posibilidad del lenguaje y la comunicación humana (p. 41). Recurriendo, para esta nueva explicación, a un plano teórico alejado de cualquier imposición normativa con respecto a las actividades de comunicación cotidianas, generalmente imperfectas y en constante reconstrucción. Las críticas en esta etapa se enfocan, especialmente, en esa toma de distancia con respecto a la cotidianidad o praxis humana, la cual podría ser entendida como una explicación "desde una posición teórica de autoridad" (p. 58), configurando una suerte de alejamiento con respecto a las realidades humanas fácticas. Lo que genera una nueva reinterpretación de su concepto, recogido por la autora en el siguiente capítulo de esta misma sección, dedicado a la así denominada *tercera comprensión*, ahora entendida como condición suficiente, contrafáctica y vinculada al aprendizaje lingüístico y ético, recurriendo, de este modo, a elementos relevantes propios de la sociología y la pedagogía (p. 60). Por su parte, en el tercer capítulo de la cara B encontramos, de un modo similar, esa nueva aplicación de la sociología, en este caso centrando su mirada en el *mundo de la vida*, ahora entendiendo tal concepto como aquello que articula los procesos comunicativos generados en las sociedades humanas. Es decir, haciendo referencia al conjunto de saberes que comparten las personas pertenecientes a una comunidad, manteniendo, así, el carácter abstracto y estático por el que será nuevamente criticado (pp. 209-212).

Tomando como punto de partida todas las variaciones y críticas acontecidas a lo largo de su personal recorrido filosófico, Habermas introduce en sus siguientes obras un fuerte carácter pragmatista que le permitirá, por fin, encaminar el aparato conceptual que ha ido configurando hacia una aplicación eminentemente práctica de la *racionalidad comunicativa*. Así, en el capítulo cuatro de la parte A, la autora nos ofrece un amplio análisis de este periodo, atendiendo a las profundas novedades introducidas, como la aceptación del falibilismo en lo referente al conocimiento humano, sumado a la posibilidad del cambio y la mejora (pp. 77-79). Alejándose, de este modo, de la posición estática que previamente había defendido, pero manteniendo una clara preferencia por una postura tendente al universalismo. Rechazando, de esta forma, la salida rortiana del etnocentrismo al incidir en la necesidad de una suerte de conexión ideal y común a todos los seres humanos (p. 100). En lo referente al así denominado *mundo de la vida* se produce un viraje similar, enfocado en la necesidad de elaborar una

postura denominada por Habermas "postmetafísica" (p. 243), cercana al pragmatismo y caracterizada por un intento de alusión a la experiencia vivida por las personas y las relaciones sociales generadas entre ellas, pero sin descartar la pretensión de universalismo previamente descrito, negando la inconmensurabilidad entre las diversas formas de vida referidas (pp. 265-266).

A partir de este capítulo, Isabel Gamero muestra cómo la conjugación entre la *racionalidad comunicativa* y el *mundo de la vida* se intensifica, compartiendo una serie de problemáticas análogas al intentar darle una aplicación práctica y sociopolítica a todo el bagaje filosófico que el autor había venido conformando. De este modo, en el quinto capítulo de ambas secciones encontramos el resultado de esta necesidad habermasiana de resolución de los conflictos humanos reales, centrando su mirada en los conflictos derivados de la complejidad propia de nuestras plurales sociedades multiculturales. Así, la autora nos presenta el concepto de "sociedad postsecular" (p. 101), a través del cual Habermas trata de abordar la pluralidad religiosa que en múltiples ocasiones ha derivado en actos de violencia y odio. Este modelo de sociedad se basaría en el intento comunitario de "alcanzar una situación de convivencia y respeto a las diferencias" (p.103), sin encerrarnos en nuestra propia visión y aludiendo a una necesidad universal de empatía y entendimiento mutuo, sumado al intento de evitar "una deriva fundamentalista y violenta" (p. 136). Para lograr esto, el autor se aleja del intento de elaborar una suerte de posición neutral, al contrario, se centra en la necesidad de un "posicionamiento como habitante del mundo de la vida" (pp. 270-271), así como en la importancia de fomentar el aprendizaje y el consenso mutuo. Evitando, de este modo, el rechazo hacia determinadas culturas o sectores y en favor de una convivencia a través del respeto de las creencias ajenas (p. 271).

En los capítulos finales de cada una de las secciones la autora aborda de forma ordenada algunos de los debates actuales más importantes en los que Habermas tomó partido. De esta manera, en el capítulo sexto de la sección B, tal y como indica su título, la autora se hace cargo de la posición habermasiana con respecto a la cuestión de "la independencia de algunas regiones, la polémica sobre el uso del *Hiyab* en Francia y breves alusiones sobre el Brexit" (p. 289), ofreciéndonos una exposición de la aplicación concreta de la posición de Habermas, basada en el concepto ya abordado de *racionalidad comunicativa*, así como de las pretensiones de validez de cada uno de los discursos enfrentados. Haciéndose cargo, igualmente, de las críticas y problemáticas derivadas de tales concepciones, aludiendo a la necesidad de creación de una suerte de organismo internacional regulatorio de tales discrepancias (p. 298). Por otro lado, en el capítulo sexto de la parte A, la autora nos acerca a uno de los debates actuales más controvertidos, a saber, "la legalización (o no) del aborto" (p. 137) a través de la aplicación de las consideraciones habermasianas de la *racionalidad comunicativa*, así como de su modelo social colaborativo basado en el consenso entre las diferentes perspectivas o puntos de vista (p. 139). Haciéndose cargo en todo momento de la complejidad propia de este debate concreto, puesto que, tal y como expresa el filósofo, "ambas partes poseen buenos argumentos y razones para defender su postura" (pp. 139-140) si consideramos los intereses de cada una de las partes y sus diferentes formas de afrontar la vida.

Tras todo el viaje filosófico y conceptual que hemos venido realizando, la autora nos ofrece un capítulo siete (en cada una de las

secciones) capaz de recoger y encauzar todas las complicaciones que las formulaciones de Habermas han tenido, así como las intensas críticas recibidas y las profundas carencias que, en multitud de casos, tendría su teoría al aplicarse a las problemáticas generadas en el seno de nuestra plural sociedad multicultural. De este modo, podemos identificar una suerte de perseverancia habermasiana basada en el intento de formalizar una suerte de marco universal y abstracto en el que nuestras diferencias quedarían apaciguadas, el cual, al analizar nuestro contexto sociopolítico, parece verse como una idealización o un proyecto actualmente inalcanzable. Justamente partiendo de estas importantes carencias, Isabel G. Gamero dedica unas últimas páginas en ambas secciones a repensar el origen de algunas de las debilidades más importantes de esta teoría, aludiendo a la necesidad de atender a los intentos de solución generados tras el así conocido giro pragmático y la génesis del falibilismo, junto a las postulaciones de la nueva teoría crítica que desde hace décadas se viene configurando.

M.ª de los Ángeles Pérez del Amo
(Universidad Complutense de Madrid)

SONGEL, F. (2021). *El arte de leer las calles*. Valencia: Barlin Libros.

El opúsculo de Fiona Songel (2021) (que, además de autora de libros y artículos, es traductora y propietaria de una pequeña librería en Valencia, llamada "La Primera")[1] es un interesante título para presentar e introducirse en la figura del *flâneur* a través de Walter Benjamin. Un libro que, por cierto, no pretende únicamente contribuir a la bibliografía secundaria acerca de las tesis del autor alemán, sino que además trata de acercarse a un lector más general mediante dos aportaciones fundamentales. La primera de ellas, algo que se detallará a continuación, es la de presentar a otros autores no tan conocidos que, sin embargo, constituyen acepciones distintas del concepto de caminante e influyeron decisivamente en Benjamin. Así, figuras como la de Franz Hessel (2015), van a ser nombres recurrentes a lo largo de esta pequeña obra. La segunda de ellas, que, sin duda alguna, es la más destacable e innovadora, es su actualización al debate sobre la cuestión del "género del *flâneur*", decisivo en el ámbito de la crítica cultural y literaria contemporánea, así como del feminismo.

El arte de leer las calles arranca con un sucinto prólogo de Anacleto Ferrer, que, en pocas palabras, nos introduce en la figura del *flâneur* a partir de un breve recorrido histórico por diferentes autores. Este acaba aterrizando en, como no podía ser otra manera, Charles Baudelaire, cuya contribución es decisiva para la estética y crítica de arte. Baudelaire, es bien sabido, ejercerá

1 Así, dice Anacleto Ferrer en el prólogo de esta misma obra: "[…] como librera. Y este es, curiosamente, el frente más próximo a la *flâneire*, entendida, al menos, a la manera de Benjamin […]." (Ferrer, 2021, pg. 13).

gran influencia en Benjamin. Sin embargo, como antes se anticipaba, no es la única: de hecho, la mayor influencia la vamos a encontrar no tanto en él como en Hessel (2015), en el cual Benjamin ve claramente representada la figura del *flâneur*.

Tras el mencionado prólogo comienza nuestro viaje o el paseo por la ciudad, donde volverán a hacer acto de presencia los dos autores mencionados. Como ya se anticipó, pese a que en el texto se vaya a seguir fundamentalmente el planteamiento de Walter Benjamin, no se van a descuidar nunca, tal y como se hace ya desde el prólogo, sus influencias. Así, Songel vuelve a convocar, por un lado, desde Berlín, a Hessel (2015), en el que ve la "personificación" del *flâneur*. Hessel (2015), en resumidas cuentas, describe al *flâneur* como una persona con la mirada puesta hacia lo contingente, hacia lo que esta oculto, y que a través del deambular desvela los secretos de la ciudad. Igualmente, y, por otro lado, desde París, a ojos de la autora, no podemos obviar la influencia de las llamadas "fisiologías" (Songel, 2021, pg. 50), en la lectura de Benjamin sobre Baudelaire. Destacando la siguiente idea, en palabras de la autora:

"Dicha lectura le permite exponer los aspectos que hacen de esta ciudad [París] la "capital del siglo XIX", como lo son […] la relación de incomodidad que sufren las masas urbanas con una ciudad que se encuentra en constante transformación, […]." (Songel, 2021, pg. 51).

El *flâneur*, nos propone Songel, debe ser entendido como un caminante de la época moderna. Un caminante que se mezcla entre las multitudes, pero que mira "desde arriba" y en solitario todo lo urbano. Así, no podemos entender el concepto sin enmarcarlo dentro de la modernidad; es decir, sin tener en cuenta los cambios del siglo XIX y principios del XX, épocas en las que empieza a abrirse un hueco. En este sentido, debemos traer al frente las palabras del sociólogo Richard Sennet (1994) y, con ellas, poner el centro del escenario el término que para muchos es la otra cara de la modernidad: a saber, la ciudad. Ésta, en su acepción moderna, se caracteriza por tres cosas: la "individualidad", la "soledad cívica" y la "velocidad" (Sennet, 1994, pg. 360). Justamente esta idea es lo que, para la autora, crea el marco perfecto para la aparición del caminante. En su interior, éste se funde con la multitud y es a través de un mirar, detallado y calmado, que describe todo tipo de incomodidades y de observaciones. Así, el *flâneur* no solamente describe la arquitectura urbanística, sino que también las incomodidades, la "velocidad" y, en suma, los problemas de la nueva ciudad moderna.

No podemos dejar de lado el segundo hecho fundamental en la obra de Fiona Songel: esto es, la cuestión sobre el "género del *flâneur*". Si atendemos a la etimología de la palabra, tal y como ella hace, descubrimos inmediatamente que ésta es masculina. El término, además, se entiende de la siguiente manera: "un paseante callejero urbano, ocioso, e intelectualmente activo […]" (Songel, 2021, pg. 20). A partir de esta idea del *flâneur* como figura masculina y privilegiada que apunta al hombre burgués surge un interesante debate que la autora no duda en traer el frente. Así pues, en este punto, y en diálogo con Lauren Elkin (2017), refiere a la figura de la *flâneuse*. ¿Existió un concepto tal o no? ¿Se puede hablar de la contrapartida femenina del masculino caminante? Para Elkin (2017), podemos enfrentarnos a este debate optando por dos "vías" diferentes. Por un lado, podemos "ajustar la figura femenina al concepto"; y, por otro lado, tratar "redefinir el concepto en sí". La autora del texto que aquí se reseña, a través de un riguroso análisis de la postura

de Elkin (2017), llega a una nueva idea, en la que niega la primera posibilidad, aquella que apunta a que la *flâneuse* sea una simple caminante que se ajusta al modelo burgués-patriarcal. Más al contrario, considera que "es necesario para reclamar el sitio de la *flâneuse*, revisar el concepto y actualizarlo" (Songel, 2021, pg. 112), lo cual representa una de las grandes novedades del libro y constituye la clave de bóveda para entender su postura dentro del mencionado debate.

El texto de *El arte de leer las calles* representa un estudio que narra con suma riqueza su nacimiento, desarrollo y presencia en la obra de Walter Benjamin sin descuidar nunca sus influencias. Solo por esto su lectura es recomendable para un público general y obligatoria para aquel que en el mencionado concepto se quiera especializar. Su verdadero componente innovador reside, sin embargo, en el particular acceso que representa a la cuestión del "género del *flâneur*". Como hemos visto, en él no solo se repasan las últimas actualizaciones del debate, sino que, además, se ofrecen nuevas problemáticas que surgen a partir de ellas, así como una postura particular dentro de la discusión.

Bibliografía

Elkin, L. (2017). *Flâneuse. Una paseante en París, Nueva York, Tokio, Venecia y Londres*. Barcelona, Malpaso

Hesserl, F. (2015). *Paseos por Berlín*. Madrid, Errata Naturae

Sennet, R. (1994). Individualismo urbano, en *El cuerpo y la ciudad en la civilización occidental*. Madrid, Alianza.

Songel, F. (2021). *El arte de leer las calles*. Valencia, Barlin Libros.

Eduardo Torres Morán
(Universidad Autónoma de Madrid)

ESQUIROL, J. (2021), *Humano, más humano: Una antropología de la herida infinita*., Barcelona: Acantilado.

Con un lenguaje preciso y sencillo y con la generosidad de guiarnos por los senderos de una reflexión, en la obra se presentan los diferentes matices que integran lo humano. El título es una respuesta a las promesas del transhumanismo, que con la ayuda de la tecnología busca superar las limitaciones humanas. Para el autor no se trata de ir más allá de la humano, como propone el transhumanismo, sino profundizar en lo humano. En lugar de querer ir más lejos, mirar hacia adentro, ser cada día más humanos.

Este ensayo forma parte de una trilogía que inicia con *La resistencia íntima* (2015) y continua con *La penúltima bondad* (2018), En cada una de estas obras el autor profundiza sobre la condición humana, y nos invita a pensarnos a nosotros mismos desde lo más íntimo.

Su propuesta de intensificación de la humano la presenta como un itinerario, un viaje, que nos lleva a apreciar lo cotidiano desde un ángulo muy distinto, pues lo más profundo se encuentra en la sencillez de la vida que vivimos la mayoría de los días,

redescubriendo la belleza del amanecer, la fragancia en el aire, la forma de las copas de los árboles. Éste camino de aprendizaje no se limita a las sensaciones que experimentamos con el mundo que nos rodea, también atiende a la relación única que entablamos cuando nos encontramos con otro ser humano. Un encuentro, que a pesar de que ya ha sido tratado por otros filósofos, tiene un sello muy particular que lleva la firma del autor, en el reconocimiento del regalo que recibimos con la presencia del otro.

El autor tienen la virtud de mostrarnos la complejidad de la naturaleza humana de manera diáfana. Esto lo logra gracias a que nos presenta los elementos que integran esta filosofía de vida y aclara que cada elemento es parte de una constelación. La figura estelar solo se puede formar con la conjunción de todos los elementos. Y el nombre de esta constelación es la filosofía de la proximidad, que no solamente plantea una forma de pensar y entender el mundo, sobre todo describe una forma de ser y de estar en el mundo.

El trabajo minucioso de ir desmenuzando cada una de las partes que integran esta constelación, va dibujando una actitud ante la vida. Algunos de los conceptos son: alguien, pronombre de lo humano; la intemperie, descripción de la situación en la que nos encontramos; el reencuentro, esperanza de unión y el canto, la palabra que vibra en nosotros. A esta lista se suman la herida infinita, el amparo, la compañía y la resistencia íntima, entre otros.

Cada persona es alguien, es decir, es un ser humano con un nombre que manifiesta su singularidad. Cuando nos dirigimos a una persona pronunciando su nombre con tacto y cuidado la reconocemos. En su propuesta de filosofía de vida, el autor nos invita a cambiar la lógica de nuestros tiempos, que se fundamenta en la acumulación y búsqueda del poder a una lógica que priorice la relación humana. En su opinión, lo que más necesitamos es fortalecer los lazos que nos unen con los demás. Transformar las relaciones con el gesto humano que ofrece protección y amparo y, aprender en nuestro camino de humanización, modelos de acogida y calidez.

Como un rasgo distintivo de la condición humana, Esquirol identifica cuatro heridas infinitas: el gozo de la vida, el asombro del mundo, el regalo del tú y el miedo a la muerte. El gozo de la vida es el disfrute que proporciona el banquete de la vida. El asombro del mundo representa la curiosidad que siempre ha guiado a la filosofía. El regalo del otro se manifiesta en los espacios de vida que se amplían por la presencia de otra persona. El miedo a la muerte se contrapone al gozo de la vida y es el recuerdo de que somos finitos y tenemos un destino inevitable.

Esquirol nos propone una hermenéutica del sentido de vida. Nuestra razón de ser es cuidar y acompañar. El cuidado será necesario para evitar que algo se dañe, desgaste o lastime. El acompañamiento fortalecerá los lazos humanos. En el centro de esta brújula del vivir, se encuentra el corazón, expresión de la sensibilidad, de la calidez y de estar al lado de los demás.

El lenguaje también tiene una gran relevancia en esta filosofía de vida. Con el lenguaje se busca no sólo el entendimiento mutuo, se habla para cuidar, proteger y consolar al otro. El saludo es la manifestación de la preocupación por el otro. Y se buscará dar abrigo y amparo con las palabras. Cada palabra que vibre en nosotros, manifestará la celebración de la belleza o el consuelo ante las vicisitudes.

La humanidad se manifiesta en las palabras dulces que cuidan y protegen. En contraste, donde no hay palabra, donde impera el silencio, donde no hay eco de la voz humana, se encuentra el vacio. Esto lo entiende perfec-

tamente Esquirol al hacernos ver que la mayor muestra de inhumanidad es la frialdad, la indiferencia, la insensibilidad ante el otro. Por ello, la palabra da vida a lo humano y, mientras más cálida y dulce sea, mayor nuestra posibilidad de tener un encuentro con el otro.

En la visión de Esquirol, una sociedad más humana busca construir un hogar, en vez de dominar al resto del mundo. En su constelación, la poiética del mundo tiene como meta la belleza y la justicia. La poiética de la vida se define por los verbos amar y pensar. Y la poiética del sentido busca articular y crear más sentido. El horizonte de desarrollo que se propone está en clave franciscana: hacer del mundo una casa y del otro un compañero. Construir un nosotros y vivir juntos con espíritu fraterno compartiendo el pan y el canto.

En un mundo incierto, lleno de preguntas y desconcierto, vemos una luz de esperanza que nos alienta a seguir andando. La obra de Esquirol es reveladora y revitalizante. El autor nos anima a no desesperar y establecer un horizonte de futuro. Así como valorar la pluralidad y singularidad de las vidas humanas. Nótese el plural de la vidas humanas, esto es porque hemos pasado de lo humano como concepto abstracto, al humano como persona concreta, singular y única.

Finalmente, me gustaría destacar que en términos muy prácticos el autor nos invita a vivir con lo esencial en vez de vivir con demasiado. El tener demasiado agota y desvía la atención. En esta filosofía de vida, se necesita poco: pan, casa y canto. El pan representa el gozo de la vida, la casa es el espacio donde nos sentimos seguros y el canto es la palabra que celebra y ampara.

La filosofía de la proximidad nos recuerda que aprendemos a ser humanos, a estar en una relación, a vivir y a convivir. Y este es un camino que debemos transitar con una guía del buen vivir que nos oriente y nos de perspectiva.

La obra constituye una hermenéutica del sentido de la vida, que otorga dirección y propósito. En estos tiempos de incertidumbre requerimos más que nunca una orientación para el buen vivir que nos permita elegir con sabiduría y aprender a vivir de forma más atenta.

Bibliografía

ESQUIROL, J. (2015), *La resistencia íntima*., Barcelona: Acantilado.

ESQUIROL, J. (2018), *La penúltima bondad*., Barcelona: Acantilado.

Miriam Molinar Varela
(Tecnológico de Monterrey, México)

GONZÁLEZ FERNÁNDEZ, Martín (2019), *Michel de Montaigne (1533-1592): La filosofía como ensayo. (Defensa de los animales)*. Madrid, Sindéresis, 430 págs.

Martín González Fernández, profesor titular de la Universidad de Santiago de Compostela (Galicia), compone una obra no impermeable a la crítica y tan revolucionaria como pertinente en el escenario de hoy día: *Michel de Montaigne (1533-1592): La filosofía como ensayo. (Defensa de los animales)* constituye un punto cardinal y,

a la vez, nodal para el estudio del pensador bordelés. Respecto al título de la obra, en el prólogo de Oscar Parcero Oubiña se escribe con prudente razón: «El ensayo pone y nos pone a prueba, mostrando como para un *verdadero* filosofar ningún resultado es satisfactorio, pues todo resultado que lo fuere no sería sino la fatal interrupción del quehacer filosófico, su cancelación, en nombre —eso sí, vanamente— de la verdad. […] en el sentido de un *exagium* latino —del que se origina la palabra—, es decir, un sopesar, un ponderar cada cosa, incluyendo lo más variopinto» (González Fernández, 2019, p. 8). La defensa del prologuista sobre el género y el estilo ensayístico, bien de Montaigne, bien del autor, González Fernández, es ante todo esencial en el debate actual que encara al academicismo (esto es, el estilo puramente académico-intelectualista ejercido en los despachos de las universidades) con lo divulgativo (escritos que, si bien no son puramente académicos, se allanan y amoldan en aras de su comunicación, su difusión).

La obra se compone de tres grandes bloques. El primer capítulo, «La filosofía como ensayo», es una panorámica en clave ensayística del pensamiento filosófico de Montaigne, que plasma «tensión»; «contradicción», a mi juicio, meditada y justificada; y «hosca gravedad», en palabras de González Fernández. El ensayo nos sitúa al borde del acantilado que supone el arte de escribir. Relaciona en estas hojas la filosofía (en concreto, la ontología) de Montaigne con pensadores contemporáneos suyos, como Nicolás Maquiavelo o Giordano Bruno. Sin embargo, como se reconoce, «no todo será filosofía propiamente dicha. Habrá arte, literatura, recortes de prensa diaria, de G. F. Watts hasta Jorge L. Borges, la genética o la biotecnología, el humanismo y el poshumanismo» (p. 13). Un elenco de temáticas, sin aspirar a ser sentenciosas, de las que brotan serias e intrigantes preguntas que se despliegan hasta el bloque que sigue.

En el segundo capítulo, «Metamorfosis: "*et [il] portoit à Augusta un bonnet fouré par la ville*"», se prosigue la profundización de las temáticas anteriores y se plantan, a fin de hacerlas crecer con vigor, interrogantes de la filosofía más práctica del pensador bordelés. Asimismo y sobre este asunto, dos fuentes originales llevadas a estudio en la obra suponen un acercamiento al viaje: *Journal de voyage en Italie par la Suisse et l'Allemagne en 1580 et 1581*, «donde narra, a veces con pluma ajena, la de un escribano que le acompaña, a veces por propia mano, a veces en francés, a veces en italiano, su experiencia de viaje […] que, en ruta, se verán sometidos luego a la censura romana, y que, según propia confesión, le cambiará la vida» (p. 169), y *De la vanité* (1588), donde se «esconde toda una teoría acerca del arte de viajar» (*Ibíd.*). Viajes por la *no man's land*, por el sentimiento errante del viajero que no vislumbra fronteras en el horizonte.

El tercer capítulo, «*Como gatos y perros*. El alma de los brutos en el Renacimiento: escépticos y libertinos», recupera «la mirada, histórica y filosófica, sobre la cuestión todavía abierta de la racionalidad de los animales no-humanos. Y pararnos a reflexionar […] si los humanos nos comportamos realmente como perros y gatos, en qué reside la diferencia entre nosotros y ellos. […] No se trata aquí de ofrecer argumentos en una dirección u otra, y pudiera encontrársele al final esta utilidad, sino tocar un tema que históricamente preocupó a los filósofos» (p. 235). Defensores y detractores observarán una colorida palestra para sus personales pinceladas. En el presente capítulo se crea una valiosa lectura sobre *animal rights*, *animal ethics*, desde la Antigüedad hasta nuestros días, tauroma-

quia, religión, sentido de justicia, lenguaje, ecología, animalismo, «revolución neurológica (António Damásio)» o, como se dijo, racionalidad. La observación cautelosa del autor pone de relieve un asunto de actualidad. Estudio del problema animal. Razón por la cual el libro es más que enriquecedor desde diversos ángulos.

La obra cierra magistralmente con un epílogo de Jorge Cendón Conde, quien confirma cuán actual es la figura de Michel de Montaigne. El autor de la obra que aquí nos ocupa, «manifiesta el síntoma de esta nueva historia y, a la vez, de esta nueva filosofía, de esta nueva historia filosófica, libre de ataduras finalistas, de cargas ideológicas limitadoras, de compromiso de escuela, en torno a la figura de un clásico […] El ensayo, el viaje y los animales, trinidad de bolsillo» (p. 422), escribe Cendón Conde. El libro es, a todas luces, una oda irresistible, un monumento ejemplar, a la memoria y al pensamiento de un Michel de Montaigne vivo, muy vivo en la actualidad.

Alejandro G. J. Peña
(Universidad de Sevilla)

PRO VELASCO, M. L. (2021). *Introducción a la ética de Robert Spaemann*. Granada: Editorial Comares.

La Profesora María Luisa Pro tiene entre sus principales intereses filosóficos el estudio de la filosofía de Robert Spaemann (1927-2018) sobre quien escribió su tesis doctoral en la Universidad de Salamanca (*Presupuestos e implicaciones de la ética de Robert Spaemann*). Desde su primer libro (2017, *Relación entre persona y felicidad en la obra de Robert Spaemann*. Ávila) no ha dejado de investigar y publicar sobre problemas abordados por el eminente pensador: educación, ética, ecología, bioética, persona y razonabilidad de la existencia de Dios, entre otros. Unas 10 publicaciones, incluida la presente, número significativo, dada su corta carrera docente.

En este caso nos ocupamos de su último libro, publicación que se ha beneficiado del intercambio epistolar mantenido con el filósofo alemán mientras redactaba su investigación (p. 4). Esta publicación es la segunda monografía que aparece en nuestra lengua, después de la de Ana Marta González. (1996). *Naturaleza y dignidad. Un estudio desde Robert Spaemann*. Pamplona: EUNSA. Desgraciadamente, en nuestro idioma no abundan las publicaciones sobre Spaemann, así que nos felicitamos por esta iniciativa, dada la talla filosófica de Spaemann.

El objetivo es claro: elaborar un perfil de su concepción filosófica [ética] (p. 1). El lector debe esperar, pues, que se le informe de los problemas éticos que abordó, de su posición filosófica ante ellos y de las dificultades que pueden encontrarse. Y en este objetivo el libro no defrauda porque el lector asiste a una exposición sistemática de problemas, soluciones y alternativas.

Esta *Introducción* consta de cuatro partes. En la primera se nos presenta el contexto histórico y filosófico del autor.

Estas páginas se pueden leer como una biografía intelectual y son imprescindibles para comprender las preocupaciones de Spaemann. Encontraremos también las razones que pudieron llevarle a la filosofía y, en especial, a la ética, aunque pareciera que sus intereses vitales apuntaban en otras direcciones. Con sobriedad y con amenidad asistimos al desfile de personas y acontecimientos que marcaron el camino filosófico que se nos relata en los siguientes tres capítulos.

Establecido el marco histórico, la autora, con acierto, nos propone los presupuestos fundamentales de la ética de R. Spaemann. Hay que agradecer de veras el esfuerzo sistemático que aquí se realiza, pues, cualquiera que se haya acercado a los textos de este filósofo compartirá, como el mismo Spaemann reconoció, que "sus pensamientos solían ser algo desordenados y que solo por medio de la escritura llegaban a ordenarse como si de un puzle se tratara" (p. 4; p. 53). La autora, queriendo completar la tarea iniciada por Nissing y Buchheim (p. 31) propone cinco presupuestos: teología racional, redescubrimiento de la teleología racional, antropología, problemática de la persona en la actualidad y presupuestos metafísicos. Podría malinterpretarse el término "presupuesto" en el sentido de "prejuicio" (véase p. 32, segundo párrafo). Pero no es el caso. Spaemann es consciente de que, en filosofía, hasta los mismos presupuestos exigen justificación y, en especial, el presupuesto "teología racional", decisivo para comprender tanto al filósofo como a su filosofía (p. 44). Por eso la autora se encarga de presentar algunos argumentos decisivos para sostener la pertinencia, más que plausible, de que Dios, no solo es una constante antropológica, sino también la verdad y el sentido del hombre en tanto que persona. Quien niegue este "presupuesto" ha de aceptar la carga de la prueba para dar razón de lo real.

Si el presupuesto de la teología natural es el principal, la rehabilitación del pensamiento teleológico es una de las aportaciones de mayor interés para la discusión filosófica actual (p. 45). Esta aportación cobra todo su relieve en "la filosofía de la persona, la herencia más destacable que nos legó este pensador" (p. 49; 59).

Esta segunda parte del libro está muy bien pensada, pues la autora nos hace pasar de un presupuesto a otro con naturalidad, mostrando de qué modo el siguiente nace y está entrelazado con el anterior.

Respondiendo al título del libro, el capítulo central está dedicado a la ética. Para nosotros es el más interesante (tal vez por "formación profesional"). También aquí la autora ha hecho un esfuerzo de síntesis que hay que agradecer y en el que trata de aunar la perspectiva histórica (secuencia seguida por el autor) y la teórica (temas relevantes que han ido apareciendo en su reflexión filosófica). De estos cuatro temas, dos se formulan en forma de crítica y dos al modo de propuestas (la parte más extensa): crítica al relativismo ético, en qué consiste la vida lograda, crítica al consecuencialismo y la propuesta positiva de una ética de la benevolencia como categoría ética fundamental (p. 73; p. 93. 96-97). En esta parte se nota el interés de la autora por la educación, que en sentido estricto no puede ser más que educación *ética*. En muchas páginas la Profesora María Luisa resalta consecuencias educativas de la ética de Spaemann de las que se ha ocupado en otras publicaciones.

Este capítulo ayuda a situar las tesis del autor en la perspectiva histórica del debate con otras corrientes y filósofos. Por su *claridad*, subrayo el apartado dedicado a la crítica del consecuencialismo

(p. 91-93). Por su *importancia*, señalo las páginas dedicadas a una categoría decisiva, aunque de difícil tratamiento filosófico: el amor como benevolencia (p. 93-99). Sería de interés estudiar los débitos y las aportaciones de Spaemann (p. 95) en relación con otros dos grandes filósofos que hicieron del amor una categoría central: San Agustín y Max Scheler (p. 95, tal vez sin olvidar la intuición de Leibniz, p. 94). Y por su *actualidad*, se apreciarán especialmente las páginas dedicadas a las consecuencias que esta concepción del amor tiene para los fundamentos filosóficos del cuidado del mundo natural (99-100).

Esta actualidad de la obra que reseñamos se hace más patente en su última parte que se centra en el debate de Spaemann con diversos filósofos, siempre en el ámbito de la ética. El mero elenco de nombres da idea del tamaño intelectual de Spaemann. En efecto, este capítulo comienza señalando los autores con los que se midió Spaemann, y cuyo debate solo se esboza, remitiendo al lector a los textos correspondientes de Spaemann. Se trata de Peter Sloterdijk a quien criticó por negar límites en la manipulación genética; Richard Dawkins, frente al que reivindicó la no reducción del hombre a lo que las ciencias nos dicen de él; Derek Parfit, ante quien subrayó la identidad humana, más allá de la autoconciencia actual; y, por último, Norbert Hoerster, quien ha llevado al ámbito jurídico la noción de persona de cuño lockeano, con resultados paradójicos: el cuerpo de la madre se ha convertido en el lugar más inseguro del mundo (p. 105).

Pero este capítulo se centra en la respuesta explícita y crítica que Spaemann dio a dos influyentes pensadores coetáneos: Singer y Dennet. Debido a esta confrontación, declara la autora, este capítulo "es el más rico gracias a los diálogos y confrontaciones con estos autores contemporáneos" (p. 105).

La primera parte se dedica a la confrontación con P. Singer. Son de agradecer por el estudioso de Spaemann las concordancias entre las páginas de Singer y las posiciones de Spaemann.

En apariencia, las posiciones de ambos son muy dispares, pero ¿podrían encontrarse puntos de encuentro? Esta perspectiva de investigación y los resultados a que llega son, a nuestro juicio, un punto de vista de gran interés por la actitud filosófica que la autora manifiesta (p. 108-114).

Es claro que la fundamentación consecuencialista de la ética de Singer le aleja esencialmente del planteamiento de Spaemann (p. 106-108). Pero esta constatación no hace imposibles ciertos paralelismos entre ambas éticas: En primer lugar, ambos abogan por una ética práctica en el sentido griego del término, es decir, una ética que sea guía para la vida (p. 108). Además, ambos argumentan a favor del cuidado de la naturaleza en su conjunto (p. 109-112). Tal vez la coincidencia más importante y clara de todas está en el acuerdo sobre "la paradoja del hedonismo": "quien busca ser feliz a toda costa, no lo consigue" (p. 113). Hasta aquí los puntos de encuentro.

Las discrepancias son, sin duda, más profundas que las semejanzas. La diferente noción de "persona", junto con el consecuencialismo, son los quicios en torno a los cuales giran las divergencias (pp. 106 y 119): en Singer, siguiendo a Locke, definición empirista de persona (son personas aquellos miembros de cualquier especie biológica que muestren ciertas propiedades). En Spaemann, en cambio, la persona no consiste en sus propiedades constatables, sino que es la portadora de tales propiedades. De estas diferencias se sigue,

como escolio, que aborto y eutanasia serán legítimos para Singer, no para Spaemann. La autora nos ofrece en estas páginas los argumentos de Singer y los contraargumentos de Spaemann en una panorámica que nos permite orientarnos en el debate.

Al final de estas páginas (p. 126-127) la autora señala otra raíz de las discrepancias: creer o no en la existencia de Dios. ¿Es la existencia de Dios el presupuesto que hace comprensibles las discrepancias y que abre cosmovisiones irreconciliables entre sí? Es esta una afirmación que va más allá de la descripción de semejanzas y diferencias, argumentos y contraargumentos. Hablar de cosmovisiones irreconciliables, ¿no puede llevar a hablar de "filosofías" irreconciliables, lo que implicaría la imposibilidad de todo diálogo verdaderamente filosófico?

El apartado dedicado al diálogo con Dennet, trabajado por la autora desde su época de estudiante (p. 127, nota 52), se centra en la noción de persona. En *Conditions of Personhood*, Dennet sostiene que calificamos de personas a aquellos seres humanos o no humanos que exhiben seis características. Enumeradas dichas propiedades, el análisis se centra en aquellas que fueron blanco de Spaemann: *Las personas dependen de nuestra actitud hacia ellas* (condición tercera), por lo que algo llegaría a ser alguien, si así lo decidiésemos; *las personas deben ser capaces de reciprocidad* (condición cuarta), algo que ni Spaemann, ni siquiera Singer acepta; *las personas morales deben poder hablar* (condición quinta) y deben ser *autoconscientes* (condición sexta). Expuestas las seis condiciones para ser considerado o considerada como persona y mostrado el orden de su interdependencia, Dennet trata de explicar "por qué nos resulta tan difícil aceptar que estas seis tesis son condición necesaria y suficiente para la cualidad moral de la persona" (p. 134-135). La razón aportada es que la noción de persona es normativa y actúa de ideal regulativo, es un "deber ser", no un "es (hecho)". Pero esta explicación, apuntada por Dennet, no acierta en lo más importante. De acuerdo con el último apartado: "Apuntes críticos de Robert Spaemann" (135-140), no llamamos "persona" a quien exhibe estas seis cualidades (u otras que quisieran proponerse), sino a su portador (la persona es el hombre, no la cualidad del hombre, p. 140). Esta es, sin duda, una clave de lectura del último capítulo del libro de Spaemann *Personas*. Esta tesis encierra la posibilidad de que existan otras especies biológicas, diferentes de *homo sapiens sapiens*, que sean también personas. Y, de hecho, la antropología empírica confirma que otras especies de homínidos, ya extintas, han sido personas, como prueban sus herramientas líticas, indicios de manejo de fuego y representaciones artísticas, algo de lo que nos informa la arqueología.

Las últimas páginas (136-140) antes del epílogo argumentan también seis tesis con las que Spaemann muestra las debilidades de la concepción empirista de la persona de Dennet (y también de Singer, Parfit y Hoerster).

Por el método (expositivo y sintético) y por los contenidos expuestos, no es difícil bosquejar los tipos de lectores a los que se dirige la obra que estamos reseñando. En primer lugar, el estudiante universitario que necesita hacerse con una idea global de temas, preocupaciones e interlocutores de Spaemann, en tanto que filósofo moral. En segundo lugar, es de interés también para el profesor universitario que quiere profundizar en la ética spaemanniana. Para él, este libro cuenta con una buena herramienta: la bibliografía final. El primer apartado de

la bibliografía es el de mayor interés para el investigador, pues recoge bibliografía actualizada hasta 2020. Es útil, además del elenco de fuentes en español, la literatura secundaria que, con buen criterio, ha sido clasificada por temas. La bibliografía cuenta, además, con una "bibliografía complementaria" y una "Webgrafía" de otras obras citadas y de autores relacionados con Spaemann, sea porque se ocupan de temas que él abordó o porque él ha debatido con ellos. Todo este aparato bibliográfico es de enorme importancia para quien desee profundizar en el filósofo alemán que nos ocupa.

En consecuencia, esta obra puede emplearse con provecho por alumnos, como medio de un primer acercamiento a la ética de Spaemann y por profesores universitarios, como herramienta para preparar seminarios sobre los temas o autores estudiados.

En síntesis, de este libro nos parece adecuado el contenido, el método, el planteamiento general y la disposición de las partes. La edición está muy cuidada en su maquetación y carece prácticamente de erratas. Para próximas ediciones, sería bueno cuidar el modo de decir, así como limar expresiones que pueden sonar extrañas o duras en nuestra lengua.

El índice es detallado, pero echamos de menos, sin embargo, un glosario de términos que multiplicaría sin duda la utilidad de la obra, tanto para quienes se inician como para quienes desean adentrarse en la ética de Spaemann.

En esta *Introducción a la ética de Robert Spaemann,* echamos de menos más apuntes críticos o sugerencias de desarrollo de las tesis del filósofo alemán. La crítica más explícita está en p. 139 donde se dice que tal vez Spaemann pueda contradecirse al sostener que la persona es la estructura de un desarrollo que, como tal, contiene la posibilidad (¿potencialidad?) de desarrollar determinadas cualidades.

La antropología de Spaemann, pensada, sobre todo, desde el individuo (incluso cuando apela a la necesidad de un tú, pp. 137-138; 143, por ejemplo), merecería ser completada y repensada a la luz de una antropología de inspiración fenomenológica, que advierte que el ser humano es una unidad dual en reciprocidad, esto es, el ser humano existe, no solo como individuo, sino también como varón y como mujer, ambos con idéntica dignidad, pero distintos (no distantes), llamados a una reciprocidad donde lo humano se entiende a la luz de la categoría de la comunión (¿una versión del amor benevolente?). Este rasgo, inscrito teleológicamente en la naturaleza humana, muestra, en la línea de Spaemann, que el ser humano es un hápax en el mundo natural, lo que tiene consecuencias filosóficas de enorme calado y que convendría explotar.

En conclusión, la opción de la autora ha sido publicar un libro expositivo. Este es su valor y también su "límite". Por eso, el lector queda con ganas de más y mejor. Pero no hay que olvidar que se trata de una *Introducción* y que, como tal, ha logrado el principal de sus objetivos: precisamente que el lector se quede con ganas *de más y mejor* y vaya a beber a las fuentes.

Jesús Manuel Conderana Cerillo
(Universidad Pontificia de Salamanca)

CAMPOS, Ricardo. *La sombra de la sospecha. Peligrosidad, psiquiatría y derecho en España (siglos XIX y XX)*. Madrid, Catarata, 2021, 255 páginas [ISBN: 978-84-1352-197-9].

En un artículo publicado en 1978 en la revista *Journal of Law and Psychiatry* y titulado "La evolución del concepto de «individuo peligroso» en la psiquiatría del siglo XIX", el filósofo e historiador Michel Foucault reflexionaba a partir de un juicio concreto sobre los problemas que planteaba la aplicación de los códigos legales en ciertos casos. Y es que, en efecto, tras la comisión de un delito, al acusado se le exige en la corte no ya que confiese sus actos contra la Ley, sino que además recapacite sobre sí mismo, que comparta sus reflexiones y confidencias, en fin, que diga *quién es* y qué le ha llevado a su situación actual. Se trata de una especie de ejercicio espiritual, de un autoexamen psicológico que no hace sino apuntar, precisamente advierte el francés, a esa "psiquiatrización" del ámbito de lo penal que fue extendiéndose por toda Europa y América desde principios del siglo XIX. Es el momento, precisamente, en el que la nueva psiquiatría impregna los principios de la reforma penal, y donde se pregunta por la responsabilidad de un sujeto que no es capaz de dominarse, que está sometido por la locura, en fin, que es un *individuo peligroso*.

En estas líneas temáticas cabría situar ciertamente el libro de Ricardo Campos que aquí se reseña, *La sombra de la sospecha. Peligrosidad, psiquiatría y derecho en España (siglos XIX y XX)*, una obra llamada a ser de referencia infranqueable y que, en el ámbito español, se entronca con otros trabajos herederos en cierta medida de la obra foucaultiana, pero esencialmente preocupados por los desarrollos históricos en España de ciertos problemas tratados por el francés en otros contextos: *Miserables y locos. Medicina mental y orden social en la España del siglo XIX*, de Fernando Álvarez-Uría (1983); *Ciencia y marginación. Sobre negros, locos y criminales*, de José Luis Peset (1983); *La razón y la sinrazón: asistencia psiquiátrica y desarrollo del Estado en la España contemporánea*, de Josep Mª Comelles (1988); *La invención del racismo. Nacimiento de la biopolítica en España (1600-1940)*, de Francisco Vázquez (2009); o *El discurso psicopatológico de la modernidad: ensayos de historia de la psiquiatría*, de Enric Novella (2018).

En este orden de problemas, el investigador del Instituto de Historia del CSIC hace gala de unos extensos conocimientos de la historia de la psiquiatría en España, para plantear las relaciones entre medicina mental y derecho penal en nuestro país, pero elaborando con ello un ejercicio virtuoso de historia del presente que entronca los problemas planteados con nuestra más radical actualidad. Es esta exigencia la que nos permite conectar su trabajo con los anteriormente citados –entre otros–, y la que impulsa las cuestiones centrales de *La sombra de la sospecha*: la pregunta por la propia noción de peligrosidad y sus implicaciones jurídicas, policiales, administrativas, penales, etc.; el papel de la psiquiatría en los discursos penales desarrollados en España desde mediados de la década de 1850; la primacía de un derecho fundamentalmente orientado al castigo del autor, y menos atento a las circunstancias en las que se cometieron los actos delictivos; los ejercicios y discursos parapsiquiátricos mucho menos atentos frecuentemente a los conceptos científicos que a las exigencias políticas y las preocupaciones sociales del momento; etc.

Cuestiones, en efecto, que se conectan con las dos grandes preguntas de las que parte el ensayo, según el propio autor señala en la introducción: por un lado, averiguar cómo una disciplina como la psiquiatría impulsó la patologización del crimen y la identificación de la enfermedad mental con la peligrosidad, a pesar precisamente de sus iniciales proclamas humanistas y filantrópicas; y, por otro lado, discernir cómo la psiquiatría logró influir en el derecho penal, incluyendo en el código penal determinadas propuestas sobre la peligrosidad social, e incluso –ya quedó anotado aquí– desplazar el foco de atención sobre la personalidad del infractor, por encima incluso de los hechos cometidos. Por supuesto, advierte el Dr. Campos, no se trata tanto de proponer una línea genealógica que dibuje una solución de continuidad en el desarrollo de esas imbricaciones entre la psiquiatría y el derecho penal, sino de trazar precisamente las continuidades y discontinuidades de esta historia, recalcando al tiempo los aspectos ideológicos de los escritos analizados y los discursos políticos que se fueron tomando en cada momento.

Para responder a estas preguntas el autor adopta una cronología larga, que se extiende desde mediados del siglo XIX con la emergencia de conceptos como el de monomanía y sus controversias asociadas, hasta la aprobación del Código Penal de 1995, que derogaba la anterior Ley de Peligrosidad y Rehabilitación Social franquista de 1970, mostrando al tiempo la radical actualidad de la investigación. Estudio de *long durée* que además, y esto nos parece especialmente loable, se funda en un extenso material histórico donde las fuentes son extraordinariamente heterogéneas y diversas: desde códigos legislativos hasta artículos y monografías psiquiátricas, referencias literarias y actas del Diario de Sesiones de las Cortes, expedientes de los Tribunales de Vagos y Maleantes y de Peligrosidad Social después, monografías criminológicas y jurídicas, etc. Además de cuantiosas referencias provenientes de la historia social, la literatura, las llamadas disciplinas "psy" o bibliografía secundaria especializada en los distintos temas analizados.

Con este arco cronológico y sobre esta riqueza de fuentes primarias y secundarias, *La sobra de la sospecha* se divide en siete capítulos que indagan en las cuestiones planteados según cortes históricos oportunamente justificados. En cada uno de ellos, así, se analizan las imbricaciones entre los discursos psiquiátricos y los desarrollos penales, prestando especial atención tanto a las disputas generadas como a las distintas propuestas de intervención, y todo ello sin olvidar el contexto histórico y los conflictos sociales donde se enmarcan los problemas tratados. Es de hecho este contexto el que justifica los estratos de tiempo analizados en cada capítulo, donde vemos aparecer y desplazarse, silenciar o encumbrar determinados discursos y alianzas entre juristas y psiquiatras, lógicas agónicas que dibujan precisamente las idiosincrasias propias de cada momento.

Así, en el primer capítulo, titulado "La sospecha: enajenados que se confunden con los cuerdos", el autor adopta como eje vector la cuestión de la monomanía y su influencia en varios casos criminales de gran impacto en su momento, cuestiones que le permiten mostrar la importancia, controversias y retroalimentaciones entre las viejas categorías alienistas, los discursos degeneracionistas y las cada vez más presentes referencias a la criminología italiana con las teorías lambrosianas al frente. Eran precisamente esos cruces los que habían otorgado al psiquiatra el cometido de diagnosticar la peligrosidad del acusado, reformulando así la teoría del "delincuente nato"; esto es, del "individuo peligroso". En el segundo capítulo, "Vagos y trabajadores", el

autor se centra en los diagnósticos higiénicos y médico-sociales que vinculaban ya desde mediados del siglo XIX la pobreza con la criminalidad, en clara referencia a esa Teoría de la Defensa Social que perseguía limitar las desastrosas consecuencias sociales derivadas de los procesos de industrialización y depauperización de los nuevos trabajadores obreros y sus condiciones de vida en las ciudades. Aliados con los discursos degeneracionistas y el llamado darwinismo social, los teóricos de la medicina social y los médicos y psiquiatras eran los encargados de diseñar programas de intervención política que impidieran la degeneración de la raza. Este análisis de los programas de la medicina social fundados en las prácticas y discursos científicos es continuado en el capítulo tres del libro que aquí se reseña, titulado "La Horda", un apartado centrado precisamente en el lugar de acción privilegiada para estos saberes: la gran ciudad. En torno a las reflexiones y encuestas sobre la "mala vida", el autor se centra aquí en los discursos referidos al submundo urbanita, recurriendo de nuevo a una gran variedad de fuentes –estudios criminológicos, fuentes literarias, notas policiales, prensa de sucesos, etc.–, y estableciendo además paralelismos con los discursos y prácticas desplegados en otros lugares de Iberoamérica y Europa. El personaje del "malviviente" aparece así como un individuo potencialmente peligroso, dibujado entre discursos pseudocientíficos y soflamas políticas que encuentran en Madrid y Barcelona sus lugares privilegiados, y que reactualizan la vieja distinción entre el vago y el trabajador honrado.

Centrado quizá algo más en la historia y la sociología de la psiquiatría, en el capítulo cuarto, titulado "La era de la higiene mental", Ricardo Campos nos ofrece una espléndida panorámica del ámbito de la psiquiatría desde 1915 hasta el inicio de la Guerra Civil, un momento donde la escuela de Ramón y Cajal y los Lafora, Sanchís Banús, Sacristán, Mira o Fuster componen una incomparable generación de psiquiatras. Es en este contexto cuando nacen tanto la Asociación Española de Neuropsiquiatras y la Liga de Higiene Mental, iniciativas que buscaban en cierta medida minimizar la población nosocomial proyectando la medicina mental con nuevos tintes sociales y reformadores. Con avances y retrocesos dependiendo de las tumultuosas coyunturas políticas del periodo, la búsqueda de una medicalización efectiva de la enfermedad mental no consiguió afianzarse, mientras que las categorías de peligrosidad y las perspectivas de la defensa social se mantuvieron casi intactas. Este panorama continúa siendo analizado en el siguiente capítulo, "Las Reformas Republicanas", donde el autor analiza los avances de la psiquiatría durante los cortos años de vida de la Segunda República, un nuevo régimen que asentó la primacía del criterio médico sobre la enfermedad mental, creó nuevas cátedras e impulsó la apertura del primer dispensario en Madrid, pero no eliminó la categoría de peligrosidad de su legislación ni desvinculó la psiquiatría de la teoría de la defensa social. Es conocida en este sentido la promulgación en 1933 de la famosa Ley de Vagos y Maleantes, quizá las expresión legislativa española más clara respecto a la línea del Estado interventor de aquellos años, y que de hecho incluía el supuesto de la criminalidad predelictual.

La instrumentalización política de esta Ley y de los propios discursos sobre la peligrosidad encontrarían especial predicamento durante "La Larga Noche del Franquismo", título del capítulo seis de *La sombra de la sospecha*, donde Ricardo Campos nos muestra la aniquilación del renovador movimiento de higiene mental republicano, así como la estrategia patologizadora del adversario político que reinó en la psiquiatría de la posguerra española. Con Vallejo Nágera y López

Ibor al frente, la psiquiatría oficial de aquel primer franquismo limitó no obstante las derivas eugenésicas de otros Estados totalitarios debido a sus manifiestas influencias católicas, lo que no impidió más tarde ampliar los potenciales individuos disciplinados bajo la renovada Ley de Vagos y Maleantes a homosexuales, gamberros o vagabundos. El concepto de "individuo peligroso" seguía por tanto plenamente efectivo en nuestras prácticas y discursos penales, tanto que sería incluido en la también famosa Ley de Peligrosidad y Rehabilitación Social de 1970, tratada ya en el séptimo y último capítulo del libro, titulado "La Peligrosidad en el Tardofranquismo y la Transición Democrática". En este capítulo, que arranca en la España del Desarrollismo y concluye a principios de los años ochenta, se continúa el análisis de las retroalimentaciones entre psiquiatría y derecho penal, en un momento en el que se evidencian tendencias internacionalistas –como en tantos otros ámbitos en aquellos años–, se recuperan los principios de la Higiene Mental ensayados durante la República y se concede progresivamente cierto espacio para los métodos procedentes del psicoanálisis, teoría considerada subversiva en el periodo anterior. Todo ello, no obstante, en un contexto en que las transformaciones socioculturales provocados por el aumento de los niveles de vida y bienestar de los españoles, la extensión del consumismo o los cambios de costumbres derivados del turismo masivo y la emigración a Europa –normalmente de vuelta en el caso español–, erosionaban los "principios del régimen" y provocaban una creciente desafección entre los jóvenes. En este contexto, la nueva Ley no solo certificaba la vieja vinculación entre peligrosidad y enfermedad mental, sino que además redoblaba la versión puramente punitiva de una legislación con tintes claramente clasistas y moralizadores. El libro concluye, no obstante, con unos valiosos apuntes sobre las críticas hacia esta Ley de Peligrosidad y Rehabilitación Social ya desde el tardofranquismo y muy especialmente tras la muerte de Franco, cuando un movimiento creciente tanto antipsiquiátrico como de pacientes reclamaban una transformación profunda de la asistencia de la enfermedad mental en España.

Como ya señalamos al inicio de estas líneas, este magistral recorrido por las múltiples hibridaciones, conflictos e interrelaciones entre la psiquiatría y los discursos y prácticas penales en torno a la noción de peligrosidad, constituye sin duda un brillante ejercicio de historia del presente. Se trata, en efecto, de una muestra de esa ontología de nosotros mismos que nos permite permanecer atentos en este caso a las reactualizaciones de aquella vieja conexión entre enfermedad mental y peligrosidad, más allá de cualquier consideración social, cultural o acaso curativa. El propio Ricardo Campos advierte al inicio mismo de este ensayo, al recordar el informe que el grupo de Ética y Legislación de la Asociación Española de Neuropsiquiatría emitía en 2013 mostrando su rechazo a la modificación propuesta por el entonces ministro de Justicia Alberto Ruiz-Gallardón, cómo la equiparación de la enfermedad mental con la peligrosidad sigue siendo una tentación demasiado palpable entre nuestros legisladores. Después de todo, lo señala el autor, las relaciones entre psiquiatría, derecho y la consideración del enfermo mental como peligroso, lo vemos en las páginas y secciones bellamente narradas y pertinentemente enlazadas de este libro, se han situado en el centro y han marcado el devenir mismo de la psiquiatría desde hace más de 200 años.

Salvador Cayuela Sánchez
(Universidad de Murcia)

CAYUELA SÁNCHEZ, Salvador y RUIZ RODRÍGUEZ, Paula Arantzazu (2022). *Foucault y la medicina. La verdad muda del cuerpo*. Madrid: Morata, 226 páginas.

De entre todos los libros escritos sobre la obra de Michel Foucault el que aquí pasamos a reseñar constituye un esfuerzo conjunto por dar sistematicidad a la relación de este pensador con la disciplina médica. En el fondo, se podría decir que dicha relación se encuentra atravesando, como poco, buena parte de sus ideas principales. Sin embargo, en los estudios en español es difícil encontrar un análisis demorado y riguroso de todas las aristas de su impacto y sus consecuencias. Lo cual resulta casi inusitado si entendemos que para Foucault no hay poder ni tampoco libertad sin un cuerpo que, a la vez, padece (es moldeado y subjetivado) y resiste (es capaz de un cuidado de sí emancipador). Por tanto, el *agón* que el pensador estableció entre los resortes disciplinarios y de control de la dupla saber/poder, de un lado, y la posibilidad del devenir autotransformador por fuera de dichos resortes, de otro, no se entiende sin la centralidad de un cuerpo que desde el surgimiento de la ciencia médica en la modernidad pero, en general, desde la emergencia misma del *médomai* (pensar, meditar, cuidar), voz ineludiblemente ligada a la de *logos* si atendemos a su significado en el griego antiguo, es nódulo reticular de la consabida *microfísica del poder* a la que dio forma a través del estudio genealógico y arqueológico de los saberes. Por tanto, estamos ante un libro imprescindible si queremos estar en condiciones de abarcar en profundidad el pensamiento foucaultiano. Pensamiento del o sobre el cuerpo y pensamiento que es *ya* cuerpo, es decir, subjetividades que se moldean y se producen por mor de su encarnación corporal, biológica, y que, también por ella, escapan y desbordan la tenaza de los dispositivos.

Editado por Salvador Cayuela y Paula Arantzazu Ruiz, autores a su vez de la introducción y el capítulo final, el libro abarca tres grandes áreas temáticas: la primera, titulada *Sobre la locura y la perversión* y centrada en la psiquiatría, su acontecer histórico y, por tanto, sus transformaciones a lo largo del tiempo, incluye tres capítulos. Así, abre la sección Fernando Álvarez-Uría Rico con el texto "Brujería, medicina y procesos de subjetivación. La tarea de fundamentar una moral laica de la ciudadanía". El catedrático de la Universidad Complutense fija su atención en el proceso secularizador de la modernidad bajo el cual la psiquiatría se habría convertido en deudora de las concepciones religiosas. Calando ampliamente en las teorías científicas acerca de la enfermedad mental y emparentándola con la moralidad, la instancia religiosa se habría transformado en instancia estatal prolongando el gobierno pastoral y trocando trascendencia por inmanencia. De ahí la necesidad de encaminarnos hacia un laicismo que, apunta el autor, es deducible de la obra de Foucault.

El segundo capítulo de este bloque, firmado por el también catedrático Francisco Vázquez García (Universidad de Cádiz) y denominado "Foucault y la Medicina de las perversiones", se ocupa de la cuestión de las supuestas desviaciones de la conducta sexual en términos psicopatológicos desde una perspectiva arqueológica en la que no sólo se involucra la obra de Foucault sino en la que, más ampliamente, se aborda la reconstrucción histórico-teórica de los discursos en torno a las "perversiones" que de

un modo u otro se hicieron eco de las tesis construccionistas foucaultianas. Por último, para cerrar el apartado, tenemos la aportación de Enric Novella, "La locura, el sueño y la existencia. El joven Foucault y la psicopatología fenomenológica". El profesor Titular de Historia de la Ciencia (Universidad de Valencia) establece allí la relevancia del análisis existencial y la psiquiatría fenomenológica de Ludwig Binswanger, fundamentalmente, como parte idiosincrásica de las preocupaciones foucaultianas más reconocidas: desde su atención sobre el sueño o la locura, hasta la ontología del presente que anuda el necesario estudio del pasado, pasando por el despliegue del cuidado de sí.

En la siguiente área temática que estructura el libro bajo el epígrafe *Sobre biopolítica y bioética*, encontramos otros tres capítulos. En primer lugar, el capítulo cinco elaborado por Josep M. Comelles y Joan Guix Oliver y titulado "Covid-19, entre el riesgo la biopolítica y la medicalización. El caso de Cataluña", constituye un riguroso acercamiento a la actualidad de los temas foucaultianos en torno a la biopolítica. En un contexto de pandemia global, la lectura biopolítica es desde luego fundamental, más aun si se aborda con el rigor crítico que exponen los doctores en Medicina Comelles y Oliver en su texto. Anudando el fenómeno vírico-biológico del SARS-CoV-2 con las condiciones tecnológicas, sociales, culturales, políticas y ecológicas de nuestro mundo globalizado que lo configuran como una "sindemia", los autores se aplican al caso concreto de Cataluña para mostrar la posible irreversibilidad de la rotura con el "modelo de *auctoritas* médica" (p. 107).

En segundo lugar, ocupando el capítulo número seis, Richard Cleminson, Catedrático de estudios Hispánicos en la Universidad de Leeds, firma "La política de la salud en el movimiento libertario lusófono: Portugal y Mozambique, 1910-1935". Biopolítica colonial, salud e "historia y práctica del anarquismo" obrero (p. 117) funcionan aplicando la "caja de herramientas" foucaultiana como puntos nodales de una retícula de biopoderes que encuentran su tensión en el establecimiento de un tipo de biomedicina nacional-capitalista propia de las metrópolis, con sus consiguientes categorizaciones jerárquicas raciales y económicas, y una forma alternativa de entender el conocimiento y la praxis médica que ponía en jaque aquellas mismas dinámicas capitalistas de explotación. En tercer lugar, el séptimo capítulo y último de este gran apartado nos lo trae Diego José García Capilla con "Bioética: una perspectiva a través de la obra de Michel Foucault". A pesar de la ausencia, al menos explícita, de esta temática en el pensador de Poitiers, tal y como reconoce el propio García Capilla, el médico y filósofo de la Universidad de Murcia ofrece una investigación sistemática de las implicaciones que la dupla saber/poder tuvo para el surgimiento de la bioética en los años setenta, así como para su posterior desarrollo. Implicaciones teóricas y prácticas en su avance conceptual y en su proceso de institucionalización a las que la aplicación de los términos fundamentalmente biopolíticos de Foucault ayuda, sin lugar a dudas, a arrojar luz.

Por último, con el tercer área temática, *Sobre la norma, la desviación y la discapacidad*, se cierra esta obra con dos capítulos más. El capítulo ocho queda elaborado por Melania Moscoso Pérez, Científico Titular en el Instituto de Filosofía del CSIC. Con "Foucault y los *Disability Studies*: aproximaciones a una relación conflictiva", Moscoso logra un acercamiento crítico a la noción de discapacidad desde los desarrollos teóricos que, partiendo de Foucault, han visto en esta noción su dimensión biopolítica irreductible. La discapacidad así vista no es otra cosa

que una categoría que logra cristalizar la producción de subjetividades determinadas por ella bajo el binarismo normal/patológico, resultado del devenir histórico hacia la preponderancia de la normalización. Moscoso emprende, pues, una genealogía del concepto que llega hasta las actuales teorías capacitistas, las cuales, según reflexión, habrían abandonado en parte la profundidad foucaultiana (y canguilhemiana) necesaria.

Cierra la obra el título "Un silencio que interpela. Interpretación biopolítica de la desviación física", por Salvador Cayuela Sánchez y Paula A. Ruiz Rodríguez, doctores y profesores de la Universidad de Murcia. En línea con el capítulo anterior, el texto señala el trayecto que conformó la noción de discapacidad mediante su encuadre en el marco del biopoder contemporáneo hasta la actualidad de la gubernamentalidad propia del neoliberalismo capitalista. Así, "transformando lo negativo en positivo, el gobierno neoliberal de la discapacidad es capaz de regular y tornar en espacio de actividad social y productiva el fenómeno de la discapacidad" (p. 214). Trasunto inequívoco de ese poder que incluye a la vida en sus cálculos mediante una *anatomopolítica* de los cuerpos y una biopolítica de las poblaciones.

En definitiva, como ya adelantamos al inicio, estamos ante un libro fundamental en la medida en que reconstruye arqueológica y críticamente el vínculo de Foucault con la medicina a través de los estudios humanísticos y sociales mostrando de manera crítica su actualidad y el calado del debate en torno a cuestiones nucleares: biopolítica, bioética, sexualidad, locura, salud, patología, discapacidad, resistencia, libertad, etc. Conceptos que desfilan en este estudio para dar cuenta, como se afirma en el primer capítulo firmado por los editores a modo de introducción, de la transformación en la obra y la vida de Foucault del "escalpelo en pluma (...) tomando la hoja de papel como el cuerpo de los demás" (p. 10). Seguramente, también el suyo propio.

María García Pérez
(Universidad de Valladolid).

RODRIGUEZ, R. y JARAN, F. (eds.). (2021) *El proyecto de una antropología fenomenológica*. Madrid: Guillermo Escolar. 318 páginas.

La relación entre la antropología filosófica y la fenomenología es el objetivo del presente texto. Bien como inclusión, confluencia, divergencia, etc. las relaciones entre ambas vienen marcadas por el veto fenomenológico de Husserl y Heidegger. Así, los especialistas que confluyen en el libro abordan si es posible, y en qué sentido, el proyecto de una antropología fenomenológica.

Ramón Rodríguez presenta la evolución de la antropología filosófica en el pensamiento de Heidegger. En un primer momento, la antropología fenomenológica radical trataría de retroceder a las experiencias originarias de las que surgieron los conceptos de la antropología occidental con los que pensamos la vida fáctica. Después, en la etapa de la ontología fundamental de *Ser*

y Tiempo Heidegger establecería un primer rechazo explícito de la antropología filosófica por sustancializar conceptualmente lo que tendría que transparentar el ser a través de una existencia dada en actos intencionales. Pero también podría darse el caso de que una analítica existencial completa funcionara como a priori existencial de la antropología filosófica, cosa que no aparece en la elaboración final de *Ser y tiempo*. En los tres años siguientes a *Ser y Tiempo,* Heidegger habría abordado la indeterminación de la antropología filosófica atendiendo tanto al método como a su inicio y final, asumiendo que sólo desde la responsabilidad que comporta la comprensión del ser podría superarse la finitud esencial de la antropología filosófica. Por último, Rodríguez señala que, en el horizonte ontohistórico de la *Kehre,* Heidegger considera la antropología filosófica como un indicio de que en la época moderna el ser se manifiesta como humanismo y metafísica de la subjetividad: la antropología filosófica quedaría absorbida en el proyecto de una antropología total, como corresponde al destino del ser en la actual realización epocal de la metafísica.

Jean-Claude Monod se detiene en la antropología filosófica de Blumenberg para explicar cómo supera la supuesta prohibición antropológica de Husserl y Heidegger, prohibición contingente en ambos casos. Blumenberg supera la universalidad de conciencia pretendida por Husserl, independiente de determinaciones metafísicas, gracias a la fundamentación antropológica de la intencionalidad. En el caso de Heidegger, Blumenberg habría prescindido de la pregunta por el ser para centrarse en las categorías antropológicas centrales de la obra de Heidegger. Monod nos presenta el ardid de Blumenberg para reivindicar la antropología fenomenológica en el sentido en que "la posibilidad manifiesta de una antropología fenomenológica es más fuerte que la prohibición o la exclusión factual de la antropología por parte de los fenomenólogos" (p. 55). Desde aquí Monod nos presenta las características de la antropología de Blumenberg, quien, rechazando una definición fijada de hombre, apunta a la indigencia de ser (*mängelwesen*), su distancia ontológica espacio-temporal y la visibilidad de un ser que ve y es visto, resultando su antropología un conjunto de hipótesis no clausuradas sobre el hombre.

Hans Ruin expone que Heidegger se opone a las filosofías de la vida porque éstas fallan a la hora de mostrar el sentido ontológico de la existencia humana. En este trasfondo, Heidegger habría establecido la diferencia entre el hombre y el animal. Si el hombre es un configurador de mundo, el animal se revela como un ser pobre de mundo. Ruin nos muestra que, en el pensamiento de Heidegger, tal distinción no equivale a una jerarquización entitativa. El animal tiene mundo y su propia forma de habitarlo, pero lo vive en un limitado repertorio de comportamientos. Por eso, insiste Ruin, para Heidegger el animal a la vez tiene y no tiene mundo, ya que no puede adoptar un posicionamiento en cuanto tal. El hombre vive su mundo desde la condición de abierto (tener acceso al mundo en cuanto tal), lo cual se manifiesta en su condición mortal. Si podemos decir que los hombres mueren, del animal sólo puede indicarse que llega a su final.

Jean Grondin señala que Gadamer suscribe una especie de antropología negativa, sin que pueda darse una definición definitiva sobre el ser humano. El hombre es para Gadamer un ser que comprende, pero que comprende de manera limitada, pues no llega nunca a comprenderse a sí mismo. Gadamer habría tratado de recuperar la tradición humanística. La racionalidad humana

se basa no en la autorreflexión (como decía la modernidad) cuanto en la sustancia predada de la historia. Por eso, la racionalidad absoluta no sería una posibilidad de la humanidad histórica y la libertad se da dentro y gracias a la historia. Por último, señala Grondin, Gadamer rescata la lingüisticidad aristotélica para fundamentar en el lenguaje nuestra libertad y nuestra apertura al mundo.

Mª Carmen López Sanz explica el interés fenomenológico por el sentido de lo humano a través de Husserl y Merleau-Ponty. Explica, en primer lugar, que la subjetividad trascendental corresponde en Husserl con el campo de experiencia del flujo vital. A continuación explicita, de la mano de Merleau-Ponty, que tal subjetividad trascendental corresponde con el movimiento de la intersubjetividad instituyente-instituida, "no tenemos la experiencia del yo como subjetividad absoluta sino como continuamente deshecho y rehecho por el curso del tiempo y de la vida común" (p. 116). En tercer lugar, López Sanz reflexiona sobre el carácter abierto de la fenomenología en relación con las ciencias humanas, en especial la antropología estructural de Levi-Strauss. Frente al relativismo estructural, la fenomenología de Merleau-Ponty atisba una estructura fundante común previa a las distinción de lo físico, lo vital y los psíquico. Por último, señala López Sanz, esta campo fundante, carne (*chair*) en Merleau-Ponty, *Lebenswelt* en Husserl, se revela como intra-ontología, ofreciendo el sentido de los "hechos" atendidos por las ciencas humanas y dibujando así la futura integración de sentido de todas ellas.

François Jaran explica el rechazo frontal de Heidegger para presentar la pregunta por el ser como antropología. Explica Jaran que, en Heidegger, el ser, sin relación de causalidad con el ente, determina al ente en cuanto ente y aparece como horizonte desde el cual el ente es comprendido. Por eso, *Dasein* no se traduciría como ser humano, sino como sitio (*Stätte*) de la comprensión del ser. Avanza Jaran explicando así la neutralidad del *Dasein* en Heidegger: no se refiere a una especie que coincide con el *homo sapiens*, sino a una estructura trascendental desde donde el ser humano recibe su forma de ser. Pero esta estructura, a diferencia de las categorías kantianas, no es inmutable, sino modificada en la temporalidad y constituida en el encuentro con los entes. Por eso, "el *Dasein* neutro no existe como tal" (p. 162): para que el descubrimiento del ser tenga lugar se hace necesaria la ruptura de la neutralidad en la vida fáctica del *Dasein*.

Leonardo Rodríguez Duplá explica por qué *El puesto del hombre en el cosmos* de Max Scheler puede ser considerada la obra fundacional de la antropología filosófica y en qué consiste su novedad. Para Rodríguez Duplá, a pesar de la continuidad de fondo, esta obra representa en Scheler el inicio de la filosofía filosófica "en sentido estricto" (p. 169). Encontramos su novedad en tres rasgos: el intento de aclarar la esencia del hombre en contraste con los demás seres vivos, el establecimiento de un diálogo las ciencias naturales y la pretensión de convertirse en centro de la reflexión filosófica. Indica Rodríguez Duplá, cómo, en su recorrido histórico, la disciplina se enmarcó en la reflexión sobre las dimensiones invariables de la condición humana desde la filosofía de la vida y la naturaleza. Por ello, la antropología filosófica aparecerá en disputa con la filosofía de la historia, la cual considera la humanidad del hombre no como algo ya dado, sino como un ideal a realizar. Termina Rodriguez Duplá señalando que el veto de Husserl a una antropología filosófica habría sido sobrepasado desde la misma fenomenología, mientras que el veto de Heidegger, implica un antagonismo irreconciliable con Scheler.

Alejandro Escudero Pérez compara las nociones de trascendencia (Heidegger) y excentricidad (Plessner), señalando que ambos autores rechazan que pueda hablarse de una esencia humana con un contenido fijo. La trascendencia apunta en Heidegger al ser de la existencia, a la esencia del fundamento y a la apertura al mundo. Recalca Escudero Pérez que, para Heidegger, el hombre es así formador de mundo, en un doble movimiento de salir de sí mismo y volver a sí mismo. La excentricidad de Plessner se refiere a la posicionalidad del hombre respecto de su entorno: el hombre crea cultura y rompe con el determinismo que tendrían otros seres vivos, introduciendo la mediación en la relación con el medio y la autoconciencia por la que se distancia de todo, incluso de sí mismo. Para evitar el riesgo de objetivación de sus mediaciones culturales, Plessner introduce el nexo entre esta posicionalidad excéntrica y la historicidad del mundo de la cultura. Escudero Pérez se pregunta si, aun queriendo superar el dualismo moderno, las posturas de Heidegger y Plessner no habrían terminado por replicarlo a través del movimiento dual de salida y vuelta del hombre a sí mismo. Responde indicando que, gracias a la evolución y rectificación posterior de Heidegger, podríamos entender la trascendencia como el cuidado del ser, encomienda que recibe el hombre cada vez que vive en lo Abierto y lo habita, participando en el juego del comprender. De esta manera, la pregunta por el hombre se transforma en la pregunta por el acontecer del ser.

Stefano Cazzanelli compara la "analítica coexistencial" (p. 218) de Karl Löwith y la analítica existencial de su maestro Martin Heidegger. Ante el problema de la co-originalidad heideggeriana entre el ser-en-el-mundo y ser-con-el-otro, Cazzanelli atestigua que en Löwith la individuación del sujeto se produce en relación con el otro. En Löwith el mundo compartido precede a la constitución del sujeto, es decir "solo empezando por el dato antropológico de la comunidad relacional es posible en seguida esbozar el contenido ontológico de la estructura formal del ser-con" (p. 217). Frente a la neutralidad originaria del ser-ahí en Heidegger, Lowith se apega a la concreción de lo real y la facticidad de la existencia. Por ello, Cazzanelli distingue el pensamiento ontológico fundamental heideggeriano de la antropología ontológica de Karl Löwith, insistiendo en la paradoja de que Löwith es más fiel que su maestro al método fenomenológico husserliano al dar prioridad al dato y la antecedencia de lo real. Frente a una filosofía al servicio del ser en Heidegger, encontramos, en Löwith, una filosofía al servicio del hombre.

José Manuel Chillón explica que, en Heidegger, la actitud filosófica deviene condición necesaria para "entender en qué sentido la comprensión ordinaria sirve a los intereses de la ontología" (p. 233). Según Chillón, Heidegger distingue dos niveles en la cotidianeidad: el nivel primario, preteórico, preontológico y el nivel originario, teórico y ontológico. En el día a día, vivimos en el primer nivel, de una manera concreta y óntica, en la que, sin embargo, se hace posible ya cierta comprensión del ser. Más allá de Husserl, dice Chillón, la facticidad no es algo que se pueda poner entre paréntesis, sino más bien convertirse en ámbito en que el *Dasein* transparenta el ser. Este es el sentido en el que la fenomenología se transforma en hermenéutica en Heidegger. En el nivel óntico es posible que el *Dasein* malogre la posibilidad de transparentar el ser y se pierda en el modo impropio de ser en el mundo, esto es en la publicidad del Uno (*das Man*) donde no puede responder a la pregunta por el quién. Pero también en

la cotidianeidad aparecen las experiencias originarias desde las que el *Dasein* accede a la construcción del sí mismo: las cosas a la mano como *pragmata,* la solicitud hacia los otros en la experiencia del *coestar* y los propios estados de ánimo se deconstruyen en perspectiva fenomenológica para dar paso a la aperturidad auténtica. En esta aperturidad, el tiempo aparece horizonte de sentido. En este horizonte temporal el *Dasein* asume la resolución de hacerse cargo de sí mismo liberándose de la impropiedad del uno. Por eso, el *Dasein* se convierte en yo cuando encuentra su analítico existencial decisivo: el cuidado.

Teresa Álvarez realiza una crítica de la comprensión kantiana del instinto desde la fenomenología de Husserl. Álvarez detecta que Kant entiende el instinto de manera negativa, como una base innata moldeable por la praxis del hombre. La subjetividad kantiana se auto-determina tomando conciencia de las representaciones arrojadas por el instinto. El instinto aparece así como una *representación oscura* de la que la libertad, conforme al desarrollo moral del hombre, puede sustraernos. Pero se trata de una descripción epistémica, que abre una brecha entre la consciencia subjetiva que un sujeto tenga de sus representaciones y sus determinaciones fisiológicas. Husserl, sin embargo, advierte que nunca podremos saber si a una representación oscura le corresponde un objeto determinado, por lo que la indeterminación de la representación da paso a la indagación fenomenológica. En tal indagación se descubre que cualquier afección ocurre siempre de manera determinada; el puro instinto quedaría fuera de la experiencia humana del mundo. Como idea límite, el instinto "ejercerá en la fenomenología un papel descriptivo o reconstructivo de la actividad apetitiva y desiderativa de cada sujeto" (p. 282).

Miguel A. Martínez Gallego expone la respuesta de Scheler a las críticas recibidas de Heidegger en *Ser y Tiempo* donde había acusado a la antropología de Scheler de replicar una antropología esencialista de corte paleocristiano. Scheler objeta que su modo de proceder es más bien al revés: la esencia humana no es punto de partida, sino punto de llegada. Además, en su respuesta, Scheler devuelve a Heidegger la misma acusación recibida: el *Dasein* no es sino una descripción dogmática que esencializa los modos de ser. Scheler acusa a Heidegger de biologicismo vitalista, por no haber matizado la dimensión espiritual de la angustia y la cura y por no haber distinguido un centro supra-histórico en el ser humano.

Como apreciamos, cada uno de los autores se detiene en un aspecto de la problemática planteada. Si bien se trata de un texto cuidado y con una loable precisión conceptual, en algunos autores habría sido de agradecer una mayor claridad. Tal vez su alta especialización haya dificultado una exposición más simple. Podemos decir que se trata de un texto para lectores que ya estén familiarizados con el lenguaje fenomenológico. Además, habría sido deseable una selección más equilibrada de los problemas y los autores elegidos. Algunos análisis aparecen desdoblados en diferentes capítulos y, aunque no obsta para alcanzar mayor campo de visión, en ocasiones, resulta redundante. La atención dedicada a Heidegger sea, tal vez, excesiva. En este sentido, la presencia de autores como Jean-Luc Marion, Edith Stein, Von Hildebrand, Sartre, Schütz, etc. habría hecho justicia a un título tan sugerente como el de esta obra. Entiendo, no obstante, que se trata de un proyecto incipiente cuyas implicaciones y ulteriores desarrollos podrían enhebrarse con los actuales debates filosóficos de la antropología social y cultural. Ejemplo de

este carácter provisorio es la ausencia de un epílogo o conclusiones en que tales líneas se hubieran siquiera esbozado. No sería de extrañar, pues, un futuro proyecto editorial que se adentrara en tales cuestiones. Con todo, es un texto altamente recomendable que aborda una cuestión tan abierta como necesaria en el actual debate fenomenológico sobre la antropología filosófica.

Julián García Labrador
(Universidad Rey Juan Carlos)

NORMAS DE PUBLICACIÓN

La finalidad de *Daimon - Revista Internacional de Filosofía* es publicar trabajos de investigación en filosofía. *Daimon* es, desde 2001, una publicación cuatrimestral. Algunos de los números son monográficos y otros no. Los números monográficos son anunciados con antelación suficiente (al menos un año) mediante la correspondiente *llamada para aportaciones (call for papers),* en la que se anuncia el tema del monográfico y el nombre de la persona encargada de coordinarlo. En el caso de que un monográfico no reciba originales suficientes para completar el volumen (actualmente tenemos fijado un límite en torno a las doscientas páginas), se completará con una sección de artículos variados.

Formato de los originales: Véase en https://revistas.um.es/daimon/about/submissions

El texto de los artículos y de notas críticas que sea enviado para revisión NO debe contener datos personales del autor o autores, ni en el propio texto, ni en las propiedades del archivo informático, ni en las citas bibliográficas (en este último caso, cada cita de trabajos del autor ha de ser sustituida por la palabra "Autor" y el año de la publicación referida).
Las citas bibliográficas han de hacerse de acuerdo con el ESTILO APA a partir de *Publication Manual of the American Psychological Association*, *7th edition*, de 2020 (https://apastyle.apa.org/style-grammar-guidelines/index). Resumen en español de la 7ª ed. de estas normas en http://www.um.es/analesps/informes/APA7ed-resumenNormas-v10febr2021.pdf.

Derechos de autor:

Las obras que se publican en esta revista están sujetas a los siguientes términos:

1. El Servicio de Publicaciones de la Universidad de Murcia (la editorial) conserva los derechos patrimoniales (copyright) de las obras publicadas, y favorece y permite la reutilización de las mismas bajo la licencia de uso indicada en el punto 2.
   © Servicio de Publicaciones, Universidad de Murcia, 2011
2. Las obras se publican en la edición electrónica de la revista bajo una licencia Creative Commons Reconocimiento-NoComercial-SinObraDerivada 3.0 España (texto legal). Se pueden copiar, usar, difundir, transmitir y exponer públicamente, siempre que: i) se cite la autoría y la fuente original de su publicación (revista, editorial y URL de la obra); ii) no se usen para fines comerciales; iii) se mencione la existencia y especificaciones de esta licencia de uso.
   
3. Condiciones de auto-archivo. Se permite y se anima a los autores a difundir electrónicamente las versiones pre-print (versión antes de ser evaluada) y/o post-print (versión evaluada y aceptada para su publicación) de sus obras antes de su publicación, ya que favorece su circulación y difusión más temprana y con ello un posible aumento en su citación y alcance entre la comunidad académica.

Procedimiento: Véase en http://revistas.um.es/index.php/daimon/about/submissions

Daimon. Revista Internacional de Filosofía
Publicación cuatrimestral. Número 90. Septiembre-Diciembre 2023

Artículos

Sección 1ª: Inteligencia artificial, datos y objetividad. ¿El regreso del naturalismo dataísta?
Editores: Ariel Guersenzvaig y David Casacuberta

Sección 2ª: Ética aplicada para una Inteligencia artificial confiable
Editores: Elsa González-Esteban y Domingo García Marzá

edit.um
EDICIÓN DE LA UNIVERSIDAD DE MURCIA