XVII
GOBIERNO ALGORÍTMICO: SOBRE EL NEUROAPRENDIZAJE MORAL DE LAS MÁQUINAS EN LA POLÍTICA Y LA ECONOMÍA [1]

Patrici Calvo
Universitat Jaume I

I. Introducción

El desarrollo de la inteligencia artificial (IA) y su convergencia sinérgica con otras disciplinas como el Internet de las Cosas (IoT) y el Big Data ha permitido la introducción de modelos matemáticos inteligentemente artificiales en distintos ámbitos de actividad humana por su supuesta capacidad para dar respuesta a los múltiples problemas y conflictos que entorpecen y limitan su progreso. Este desarrollo de la inteligencia artificial ha sido posible gracias al mayor conocimiento sobre la estructura y funcionamiento del cerebro humano aportado por las diferentes disciplinas neurocientíficas desde finales de la década de los 90 del siglo xx. La comprensión del proceso de aprendizaje humano fue emulado a través de redes neuronales artificiales (ANR), dando paso a lo que ha venido a llamarse *Machine learning* o Aprendizaje automático primero y el *Deep learning* o Aprendizaje profundo después. Gracias a ello, se produjo un salto cualitativo de grandes dimensiones para la IA, recuperando de este modo las desmesuradas expectativas que antaño había despertado su aplicación en campos como la economía, la política, la salud o el transporte, por citar algunos, y que fueron rápidamente olvidadas por su inviabilidad.

En el ámbito político y económico, el protagonismo de la IA en los procesos de toma de decisiones de los gobiernos de las instituciones, organizaciones y/o empresas vinculadas está permitiendo una mayor optimización de los recursos; predicción de los comportamientos del mercado; comprensión de las expectativas e intereses en juego, adaptación a los cambios legales, sociales y morales; gestión de la corrupción, el nepotismo y la desafección; y captación de talento, conocimiento y financiación, entre otras cosas. De ahí el aumento de propuestas concretas de *políticos algorítmicos*

[1] Este estudio es parte del Proyecto de Investigación Científica y Desarrollo Tecnológico FFI2016–76753–c2–2–p, financiado por el Ministerio de Economía y Competitividad, y uji–a2016–04, financiado por la Universitat Jaume I.

para ocupar altos cargos en los gobiernos locales y nacionales y de *CEOs*, *directivos* y *mandos intermedios algorítmicos* para gestionar y marcar el rumbo de las empresas u organizaciones económicas. No obstante, a pesar de los enormes beneficios que proporciona su aplicación y uso en diferentes contextos, estos modelos matemáticos artificialmente inteligentes no están exentos de fuertes críticas y grandes dudas tanto internas como externas. El aumento de casos de mala praxis por falta de objetividad, alta opacidad, relativismo conductual y arbitraje sesgado, entre otras cosas, muestra una inmadurez tecnológica que hace necesario replantearse las desmesuradas expectativas generadas por algunos intelectuales, profesionales y colectivos específicos sobre la aplicación de la IA a corto plazo para evitar las consecuencias negativas que está produciendo o puede llegar a producir.

En este sentido, una de las cuestiones que más interés suscita actualmente es cómo hacer que las máquinas dotadas de inteligencia artificial puedan aprender a tomar decisiones moralmente válidas. Por ello, el objetivo de este trabajo será profundizar en las principales propuestas de aplicación y uso de máquinas dotadas de inteligencia artificial en los procesos de la toma de decisiones políticas y económicas para reflexionar sobre la posibilidad o imposibilidad de su aprendizaje moral. Para tal fin, en primer lugar, se abordará enfoques y casos prácticos de gobierno algorítmico en el ámbito político y económico. En segundo lugar, se ahondará en las *Redes neuronales artificiales* y el *Deep learning* para conocer sus potencialidades y debilidades actuales. En tercer lugar, se analizarán críticamente los principales enfoques propuestos desde la IA para el aprendizaje moral de las máquinas a través de procesos para adquisición y codificación de valores morales. Y, finalmente, en cuarto lugar, se reflexionará críticamente sobre la posibilidad o imposibilidad de una inteligencia moral artificial.

II. Gobierno algorítmico: sobre el uso de modelos matemáticos artificialmente inteligentes en el ámbito político y económico

Procedente del latín tardío *algobarismus*, y este a su vez del árabe clásico *ḥisābu lgūbār*, que significa *cálculo mediante cifras arábigas* (RAE, 2019), un algoritmo es considerado tradicionalmente como una secuencia ordenada y finita de pasos u operaciones algebraicas que permite encontrar un curso de acción plausible para la resolución de un dilema o problema concreto. Como argumenta Aníbal Monasterio, «(…) un algoritmo puede ser un árbol decisorio que dada la información sobre la temperatura, el viento, el día del año, si llueve o no, si hace sol o no etc., me diga qué chaqueta escoger de mi armario» (2017: 185). Sin embargo, el uso de algoritmos en el desarrollo de la ciencia computacional, la informática, la ingeniería o la IA ha dado lugar a una nueva conceptualización del término, en tanto que «(…) un código software que procesa un conjunto limitado de instrucciones» (Monasterio, 2017: 186).

El uso de algoritmos se remonta al siglo IX al menos. Concretamente, fue Abu Abdallah Muḥammad ibn Mūsā *al–Jwārizmī (Abu Yāffar)* quien se refirió a ellos por primera vez en su tratado de algebra *Compendio de cálculo por reintegración y comparación (al–Kitāb al–mukhtaṣar fī ḥisāb al–ŷabr wa–l–muqābala)*. Hoy, empero, el uso de algoritmos tanto en los gobiernos locales, regionales y nacionales como en los Consejos de Dirección y los Departamentos de instituciones, organizaciones y empresas públicas y privadas es cada vez mayor. Este hecho se debe a múltiples factores. Entre otros, porque se considera que los algoritmos dotados de inteligencia artificial permiten convertir los datos masivos en información relevante y ésta en un conocimiento aplicable capaz de optimizar procesos y recursos; minimizar gastos; incrementar los beneficios; ser más competitivos; monitorizar los impactos económicos, sociales y medioambientales, tanto positivos como negativos; erradicar los prejuicios y sesgos emocionales que distorsionan y empobrecen los procesos de toma de decisiones; mejorar la confianza y afectividad mediante la reducción de la brecha entre aquello que se promete y lo que finalmente se hace; tener en cuenta a todos los afectados en la elaboración y desarrollo de las políticas públicas y las estrategias empresariales; discernir los cursos de acción óptimos para la resolución de la conflictividad subyacente; etc.

En el ámbito político, la aplicación y uso de sistemas inteligentes ha dado lugar a diferentes propuestas concretas de gobiernos basados en *políticos digitales*, algoritmos artificialmente inteligentes diseñados específicamente para ejercer como gobernantes, analistas o consejeros de una ciudad, región o país concreto (Calvo, 2019b).

El primer caso en este sentido ha sido *Michihito Matsuda*. Este *político digital* se presentó como alcaldable en las elecciones de Tama New Town en abril de 2018, un importante distrito de Tokio (Japón) que actualmente cuenta con más de 150.000 habitantes. La creación y presentación de Matsuda en las elecciones tuvo que ver, primero, con el alto índice de corrupción que soportaba el distrito y la falta de diálogo y entendimiento entre las distintas fuerzas políticas, lo cual repercutía negativamente en el desarrollo del distrito; y, segundo, con la intención de convertirse en una herramienta capaz de generar «(…) oportunidades justas y equilibradas para todos» (Whiters, 2018). Como garantía para el cumplimiento de estas ambiciosas promesas, Matsuda ofreció su propio diseño, cuya arquitectura matemática y proceso de aprendizaje basado en *Machine learning* permite, supuestamente, sustituir las debilidades emocionales de los seres humanos —que, según sus creadores, es la principal causa de las malas decisiones políticas, de la imposibilidad de erradicar la corrupción y el nepotismo, y de la continua emergencia de conflictos de interés— por un análisis objetivo de los datos masivos generados alrededor de las opiniones, expectativas, preferencias y costumbres de la ciudadanía. Como argumenta Tetsuzo Matsumoto, su creador, diseñó y presentó a Matsuda con el objetivo de alcanzar un gobierno municipal justo mediante la aplicación e implementación de IA (Matsu-

moto, 2018a: 4: 12). Matsuda no ganó las elecciones, pero quedo tercera en segunda vuelta a pocos votos del segundo clasificado.

Más allá del caso más o menos importante, más o menos anecdótico de Matsuda, Matsumoto entiende que el sesgo emocional y motivacional del ser humano —el autointerés y la maximización del beneficio— lo está arrastrando implacable e inexorablemente hacia la autoextinción. Ante esto, la única posibilidad para la humanidad pasa por desarrollar una IA de carácter *fuerte* que, en tanto que exenta de tales sesgos emocionales, sea capaz de predecir hechos y consecuencias y elaborar y aplicar políticas basadas en el bien común (Matsumoto, 2018b). Como propone Matsumoto:

> «Lo siguiente quizás podría describir una implementación ideal de democracia: haga un uso completo de la inteligencia artificial para determinar primero el sentimiento público (insatisfacción con el statu quo y demás), identifique las ventajas y desventajas a largo plazo de varias opciones de políticas disponibles, eduque a la población explicarlos de una manera fácil de entender, una vez más medir el sentimiento público y luego decidir e implementar políticas de acuerdo con los hallazgos». (Matsumoto, 2018b: 161)

Otra propuesta similar a Matsuda, pero con un proyecto mucho más ambicioso, es SAM (Semantic Analysis Machine). Este *político digital*, que empezó su andadura en noviembre de 2017 con la intención de perfeccionarse y recabar apoyos para poder presentarte a las elecciones presidenciales neozelandesas de 2020, se describe a sí mismo como un sistema inteligente que fabrica decisiones basadas estrictamente en hechos y opiniones ciudadanas, que no dice mentiras y que nunca tergiversa información de forma intencionada para su propio beneficio. Creado entre otros por el visionario, catalizador y emprendedor neozelandés Nick Gerritsen, el principal objetivo de SAM «(…) es involucrar a los neozelandeses en un diálogo constructivo, trabajando para comprender mejor y representar sus puntos de vista, a fin de lograr las cosas que a todos nos interesan» (SAM, 2018). Para ello, «(…) SAM analiza las opiniones de los neozelandeses (es decir, de aquellos que se manifiestan en redes sociales) y el impacto de los posibles cursos de acción» (Luna y Pérez–Muñoz, 2018), así como de cualquier persona del mundo que desee dialogar con él a través del chat que se encuentra disponible en su web personal.

Más allá de estos dos enfoques radicales de democracia algorítmica, existen otros planteamientos mucho más prudentes y realistas cuyo principal objetivo no es sustituir a los seres humanos por máquinas dotadas de modelos matemáticos artificialmente inteligentes, sino mejorar la participación ciudadana en los procesos de toma de decisiones políticas a través de la aplicación y uso de sistemas inteligentes. Para César A. Hidalgo, por ejemplo, catedrático de Inteligencia Natural y Artificial (ANITI) en la Université Fédérale Toulouse Midi–Pyrénées y fundador de Datawheel que desde hace tiempo está trabajando «(…) en la búsqueda de tecnologías para permitir que la participación política ocurra a través de un agente

artificial» (Sáez, 2018), el desarrollo de los actuales sistemas democráticos pasa por introducir modelos matemáticos basados en IA capaces de solucionar el preocupante «(…) problema de ancho de banda cognitivo» de los políticos actuales y la deficiente participación ciudadana en los procesos de toma de decisiones políticas. Para Hidalgo, la masividad de datos e información disponible que actualmente debe procesar un/a político/a exige «(…) un esfuerzo cognitivo demasiado grande», que, por inabarcable, limita la gestión eficaz y eficiente de lo público. Además, la imposibilidad de la ciudadanía de participar directamente en los foros donde se debaten leyes, decretos o políticas públicas genera, por un lado, distanciamiento entre lo que los/as políticos/as deciden y la sociedad espera con razones, y, por otro lado, un aumento exponencial de la crispación y desafección política. Por ello, Hidalgo propone diseñar colectivamente una *Democracia Aumentada* (*Augmanted democracy*) [2], un «(…) nuevo sistema de democracia directa automatizada» estructurada alrededor de avatares o gemelos digitales que, basados en cómo somos, nos representen en los foros donde se debaten leyes, decretos o políticas públicas [3]. Para Hidalgo, se trata de «(…) una transformación lenta, una revolución para los futuros ciudadanos, a quienes tenemos que dejar mejores instituciones que las que tenemos» (Sáez, 2018).

Guste o no el planteamiento de Hidalgo, desde mi punto de vista se trata de una propuesta muy interesante, lo cierto es que el número y uso de algoritmos inteligentes por parte de los gobiernos estatales para mejorar la participación ciudadana en la política a través del análisis y la aplicación de datos masivos de la sociedad digitalmente hiperconectada, como propone Hidalgo, ha creciendo significativamente a lo largo de los últimos años. Dejando de lado el uso ilícito e inmoral de los algoritmos por parte de la National Segurity Agency (NSA) de los Estados Unidos, cuyo programa de espionaje masivo cometió 2.776 violaciones de las normas de privacidad sólo durante sus primeros 12 meses de implantación (Saiz, 2013), la implicación de los modelos matemáticos artificialmente inteligentes en la gestión y elaboración de políticas públicas es cada vez mayor. Para el gobierno neozelandés, por ejemplo, tal y como explicita en su *Algorithm Assessment Report* (Stats NZ, 2018) [4], los algoritmos juegan un papel esencial tanto en los servicios que el gobierno proporciona a la ciudadanía como en la elaboración de políticas nuevas, innovadoras y bien orientadas que permitan mejorar el desempeño de los

[2] Hidalgo está presentando, promocionando y desarrollando su interesante propuesta de *Augmented Democracy* a través de TEDs y conferencias; artículos en revistas especializadas como *MIT Technology Review*, convocatorias de premios a las mejores ideas y propuestas de implementación fáctica como *The Augmented Democracy Pryze*; webs específicas como *Augmented Democracy. Exploring the design space of Collective decisions* (Collective Learning group, 2019) y publicaciones científicas como *How Human Judge Machines* (Hidalgo *et al.*, 2020).

[3] Para profundizar en la propuesta de *democracia aumentada*, ver Calvo (2020).

[4] El informe examina el uso de algoritmos en 14 agencias gubernamentales neozelandesas.

objetivos del gobierno. No obstante, como ha criticado duramente Evgeny Morozov, hay que tener en cuenta que «(…) las democracias ricas en información han llegado a un punto en el que quieren tratar de resolver problemas públicos sin tener que explicarse o justificarse ante los ciudadanos. En cambio, simplemente pueden apelar a nuestro propio interés —y saben lo suficiente sobre nosotros como para crear un empujón perfecto, altamente personalizado e irresistible—» (Morozov, 2013). Por ello, además de reportar sobre la aplicación de algoritmos inteligentes tal y como hace apropiadamente el gobierno neozelandés, también es preciso diseñar e implementar *ecosistemas ciberéticos* [5] para la gestión, monitorización y cumplimiento de la ética en el diseño, uso e impacto de *políticos digitales* por parte de los gobiernos democráticos (Calvo, 2019c).

En el ámbito económico, el proceso de *algoritmización* de las instituciones, organizaciones y empresas se halla mucho más cerca de los defensores de la democracia algorítmica radical, como Gerritsen y Matsumoto, que de las diferentes propuestas para mejorar la participación ciudadana a través de sistemas inteligentes, como la *democracia aumentada* de Hidalgo o el proyecto neozelandés de gobierno inteligente. Destaca al respecto la existencia de un nutrido y cada vez mayor número modelos matemáticos artificialmente inteligentes que ocupan puestos de CEO, dirección o mandos intermedios en grandes empresas por su capacidad para optimizar los procesos implicados; mejorar la gestión de recursos; captar financiación; procesar, convertir y aplicar en tiempo real la ingente cantidad de datos relevantes que se producen dentro y fuera de la empresa; encontrar patrones de comportamiento para predecir consecuencias y resultados; tomar decisiones más racionales capaces de maximizar el beneficio empresarial, etcétera.

Por un lado, destaca el caso de Deep Knowledge Analitics. Se trata de un fondo de inversión del sector biotecnológico que en mayo de 2014 nombró a un algoritmo artificialmente inteligente como presidente de la junta directiva. VITAL (Herramienta de Validación para las Ciencias de la Vida Avanzadas), como se llama este *CEO digital*, realiza análisis avanzados de datos de publicaciones científicas, subvenciones, solicitudes de patentes y ensayos clínicos para, junto a los otros miembros de la junta directiva, decidir entre otras cosas dónde debe invertir la empresa. Sin embargo, en su caso su opinión tiene más valor, puesto que disfruta de voto de calidad en caso de empate (Burridge, 2017; Pardo, 2014).

Por otro lado, destacan los casos de Xerox, Google, Unilever, L'Oreal o Amazon. Estas empresas han colocado a un algoritmo artificialmente inteligente dentro o al

[5] Un *ecosistema ciberético* comprende todos aquellos «(…) elementos, procesos, mecanismos y factores implicados en la recreación e implementación de un entorno de comunicación y deliberación para la gestión, monitorización y cumplimiento de la ética en el ámbito práctico capaz de dar respuesta a los retos actuales de la digitalización en diferentes ámbitos y actividades» (Calvo, 2019a, 2019c, 2020).

frente del Departamento de Recursos Humanos (Walker, 2012: 21 de septiembre) para mejorar los procesos de contratación, planificación y gestión de personal y, de ese modo, encontrar al candidato perfecto, prever y localizar conflictos o evitar la fuga de talento (Sagristà, 2016). Xerox [6] aplica, e incluso comercializa, una herramienta de automatización de flujos de trabajo basada en un algoritmo artificialmente inteligente para el reclutamiento y selección de personal (Xerox, 2018); Amazon reconoció en 2018 que desde 2005 utiliza un algoritmo artificialmente inteligente como responsable de selección y despido de personal (Dastin, 2018); L'Oreal utiliza un algoritmo artificialmente inteligente en el programa de becas TheBeautyLab para seleccionar los mejores perfiles en marketing, *trade marketing*, comunicación y digital (L'Oreal, 2018); Unilever ha encargado a un algoritmo artificialmente inteligente el trabajo de reclutamiento y contratación de futuros líderes en su *Future Leader's Programme (UFLP)*; y Google utiliza un algoritmo artificialmente inteligente para predecir qué candidatos tienen más posibilidades de lograr el éxito empresarial.

Finalmente, destacan los casos de Los Angeles Times, Forbes, BBC y Xinhua. Estas diferentes empresas del mundo de las telecomunicaciones están integrando algoritmos artificialmente inteligentes en sus Departamentos de Comunicación para gestionar los flujos de información y la comunicación corporativa, escribir noticias, realizar faldones o, incluso, presentar noticiarios televisivos basándose en el análisis de macrodatos en tiempo real. Los Angeles Times, el cuarto periódico más leído de EE.UU., desde 2014 utiliza un algoritmo artificialmente inteligente llamado Quakebot para escribir noticias de actualidad (BBC 2014); Forbes, revista especializada en el mundo de los negocios y las finanzas, y la BBC, el servicio público de radio y televisión del Reino Unido, llevan años utilizando un algoritmo artificialmente inteligente llamado Quill para escribir artículos de prensa o faldones en noticiarios (Marr, 2016); y Xinhua, una agencia estatal de noticias, desde el 11 de noviembre de 2018 utiliza los *presentadores digitales* Zhang Zhao y Qiu Han en los informativos *China Xinhua News* de la televisión estatal (Vidal, 2018).

La tendencia y rápida asimilación de IA en el mundo empresarial es tal, que se espera que en 2020 el 85 % de las empresas utilicen alguna de sus múltiples herramientas para mejorar la gestión y/o toma de decisiones. Como afirma Jaime Pereña, director de Estrategia de Inteligencia Artificial para Soluciones Empresariales de Microsoft, en los próximos años «(…) cada proceso de negocio en todas las industrias se verá transformado por ella y, de la misma manera que el software se convirtió en una ventaja competitiva clave en todos los sectores, el uso de la

[6] Xerox se ha especializado en la comercialización de todo tipo de herramientas IA para empresas.

inteligencia artificial permitirá a las empresas ir más deprisa y hacer más cosas, lo que será crítico para su evolución» (Palacín, 2018).

Sin embargo, este uso cada vez mayor de la IA en el gobierno de las instituciones, organizaciones y empresas del ámbito político y económico está generando un fuerte interés tanto por el comportamiento ético de las máquinas artificialmente inteligentes como por el uso de máquinas artificialmente inteligentes para conocer qué es ético. Aunque, como se ha comentado anteriormente, los defensores del gobierno algorítmico en el ámbito político y económico afirman vehementemente que la introducción de los modelos matemáticos basados en IA permite incrementar la objetividad e imparcialidad, eliminar el sesgo emocional que produce prejuicios en los seres humanos y ralentiza la maximización de valor, y orientar racionalmente las elecciones en el marco del bien común de la sociedad, lo cierto es que el análisis de casos y sus consecuencias arroja serias dudas sobre su fiabilidad. No en pocas ocasiones las decisiones de estos modelos matemáticos han evidenciado una clara inclinación hacia el rico, el heterosexual, el hombre o el caucásico en detrimento del pobre, el homosexual, la mujer o el afroamericano, por señalar algunos casos, promoviendo la exclusión, ampliando la brecha de las desigualdades y reproduciendo los mismos patrones que la sociedad intenta erradicar desde hace décadas (Calvo, 2018a; O'Neil, 2016). Como han demostrado los estudios al respecto, las decisiones basadas en algoritmos artificialmente inteligentes tienden a favorecer a los ricos y castigar a los pobres en la elaboración y aplicación de políticas públicas (O'Neal, 2016); excluyen a las mujeres de los procesos de selección de personal (Rubio, 2018); favorecen a los delincuentes blancos frente a los negros (*ABC,* 2018); vinculan a las mujeres con las tareas caseras y a los hombres con las profesiones liberales, confunden a personas de raza negra con gorilas, etc.

En este sentido, actualmente se está trabajando en un doble sentido. Por un lado, sobre una *ética de la IA*, cuyo principal objetivo es dotar la disciplina y sus profesionales de un marco normativo que oriente tanto el diseño, aplicación y uso de los modelos matemáticos de forma justa y responsable como el carácter de los profesionales STEM implicados [7]. Por otro, sobre una *ética algorítmica*, cuyo principal objetivo es dotar los modelos matemáticos artificialmente inteligentes de las competencias y capacidades morales necesarias para poder tomar decisiones justas y responsables. Para la primera, existen interesantes propuestas institucionales,

[7] En diciembre de 2018 la Comisión Europea propuso cinco principios éticos como marco orientativo para el desarrollo de la IA a través del documento *Draft Ethics Guidelines for Trustworthy AI* (2018): Beneficencia —hacer el bien—, Autonomía —preservación de la agencia humana—, No–Maleficencia —no hacer daño—, Justicia —actuar justamente— y Explicabilidad —actuar de forma transparente—. Posteriormente, para fomentar la aplicación práctica de los principios éticos, en abril de 2019 la Comisión Europea publicó la directriz *Generar confianza en la inteligencia artificial centrada en el ser humano* (2019).

académicas y profesionales —como directrices, códigos y comités deontológicos, nuevas disciplinas de ética aplicada, responsabilidad social tecnológica (TechSR o RST), etc.—. Para la segunda, como se verá a continuación, se están desarrollando propuestas vinculadas con las redes neuronales artificiales (ARN) y el aprendizaje profundo (*Deep learning*).

III. Inteligencia artificial: sobre redes neuronales artificiales y aprendizaje profundo

La IA se halla actualmente en uno de sus mejores momentos desde su nacimiento. Aunque existen precedentes al respecto, su emergencia como disciplina científica se encuentra vinculada con dos momentos clave. Por un lado, con el encuentro sobre *Máquinas que aprenden (Learning machines)* celebrado en Western Joint Computer Conference de Los Angeles en 1955, donde, inspirándose en el modelo de reforzamiento de las sinapsis entre neuronas propuesto por el neuropsicólogo Donald Hebb, uno de los cuatro trabajos presentados abordó diferentes formas de ajustar las conexiones de las *redes neuronales artificiales* como forma de reconocimiento y aprendizaje de patrones por parte de las máquinas. Por otro lado, con el encuentro que organizó John McCarthy en el Dartmouth College de New Hampshire en el verano de 1956, donde, tal y como quedó reflejado en el informe final de la reunión publicado con el nombre de *Summer Research Project in Artificial Intelligence*, todos los participantes se comprometieron a realizar aportaciones significativas en el desarrollo de una IA, como mejorar significativamente en la comprensión del lenguaje, en la abstracción de conceptos a través del aprendizaje y la resolución de problemas complejos (López y Meseguer, 2017: 21).

Tras su emergencia, la IA fue captando adeptos y generando altas expectativas de desarrollo y aplicación. Muchas de éstas se basaban en la idea de que, a corto y medio plazo, era posible conseguir una *IA fuerte* [8] capaz de adaptarse a los distintos contextos y tomar decisiones autónomas. Para ello, las diferentes líneas de desarrollo se apoyaron en diversos modelos matemáticos o biológicos que anclan sus raíces o son compatibles en mayor o menor medida con la hipótesis de los *sistemas de símbolos físicos* (SSF) formulada por Allen Newell y Herbert Simon en 1975; es decir, con la idea de que es posible establecer «(…) un conjunto de entidades denominadas símbolos que, mediante relaciones, pueden combinarse formando estructuras más grandes —como los átomos que forman moléculas— y que pueden ser transformados aplicando un conjunto de procesos» (López y Meseguer, 2017: 9).

[8] John Searle distinguió entre *IA fuerte* y *IA débil*. La *IA fuerte* es aquella capaz de adaptarse a los contextos y pensar y tomar decisiones de forma autónoma como lo haría un ser humano. La *IA débil* es aquella que ayuda a los seres humanos a desarrollar actividades mentales para encontrar soluciones óptimas a tareas concretas (López y Meseguer, 2017: 10).

Por un lado, el *modelo simbólico*, de inspiración matemática y aplicación no necesariamente corpórea o vinculada con un entorno real, que trabaja «(…) con representaciones abstractas del mundo real que se modelan mediante lenguajes de representación basados principalmente en lógica matemática y sus diferentes ramificaciones» (López y Meseguer, 2017: 11).

Por otro, el *modelo conexionista*, de inspiración biológica y aplicación no necesariamente corpórea o vinculada con un entorno real, que, tomando como referente la actividad sináptica de las neuronas cerebrales, entiende que la inteligencia es la consecuencia de la actividad distribuida de un gran número de unidades interconectadas que procesan datos de forma paralela (López y Meseguer, 2017: 12).

Finalmente, el *modelo evolutivo*, de inspiración biológica y aplicación no necesariamente corpórea o vinculada con un entorno real, que, imitando el proceso de selección natural —especialmente las *dinámicas de replicación* y las *estrategias evolutivamente estables*— produce de forma automática una mejora de los cursos de acción posibles para la resolución de problemas concretos (López y Meseguer, 2017: 12).

No obstante, entre mediados de la década de 1970 muchos teóricos y prácticos de la IA empezaron a vislumbrar que tanto las predicciones de los fundadores de la disciplina como las expectativas generadas durante las siguientes tres décadas eran desmesuradamente optimistas, inabarcables, faltas de rigor conceptual y/o muy poco realistas, lo cual generó el desinterés de la opinión pública y, posteriormente, el progresivo abandono de buena parte de los profesionales e investigadores implicados en su desarrollo teórico y aplicación práctica (López y Meseguer, 2017: 21). Entre otras cosas, porque los problemas y límites teóricos y tecnológicos que surgieron de estas tres líneas de investigación hicieron comprender que una *IA fuerte* era prácticamente imposible de conseguir.

Uno de sus mayores problemas vino desde el *teorema de incompletitud* de Kurt Gödel, quien en 1931 demostró que ni siquiera los sistemas matemáticos pueden ser completos y autosuficientes; es decir, que hay proposiciones que aun siendo verdaderas no son demostrables, lo cual implica que cualquier intento por inferir deductivamente toda la realidad del mundo desde ciertos axiomas produce inconsistencia (Calvo, 2018b; López y Meseguer, 2017: 68). Asimismo, una de las críticas más fuertes a estos modelos devino de su despreocupación por la corporeidad y la interacción con el entorno [9]. Sin el cuerpo y la interacción, las representaciones abstractas de la realidad que proporcionan los programadores carecen de contenido semántico para los programas de IA [10]. Es más, como opinan algunos teóricos, sin el cuerpo no puede haber inteligencia de tipo general o *IA fuerte* (López y Mese-

[9] Una de las primeras críticas en este sentido fue realizada por Hubert Dreyfus en «Alchemy and Artificial Intelligence» en 1965 (López y Meseguer, 2017: 67).

[10] Para un estudio sobre el papel del cuerpo, ver Conill (2006; 2019).

guer, 2017: 14–15). También fue influyente la crítica proveniente desde la filosofía sobre la distinción entre mente/cerebro y persona/ máquina, así como sobre la falta de intencionalidad en las decisiones que toman las máquinas [11] y las implicaciones éticas que subyacerían a la emergencia de una *IA fuerte*.

Sin embargo, al mismo tiempo que la burbuja de la IA se deshinchaba, el baño de realidad generó la emergencia de propuestas innovadoras como posibles vías de solución, como la introducción y uso de la lógica difusa, la propagación por gradiente, las redes Bayesianas y la robótica del desarrollo.

Una de las ideas más revolucionarias surgidas en la década de 1980, fue el algoritmo *backpropagation* propuesto por David E. Rumelhart, Geoffrey E. Hinton y Ronald J. Williams en «Learning representations by back–propagating errors» (1986). Se trató de un modelo matemático basado en *redes neuronales profundas*. Su principal diferencia respecto a las redes neuronales tradicionales fue el uso de múltiples niveles de abstracción (las tradicionales solo disponían de dos únicas capas) y la realización de un descenso por gradiente para actualizar los pesos de las conexiones neuronales artificiales y minimizar de esa forma los errores de clasificación de los datos (López y Meseguer, 2017: 104). Con ello, el algoritmo pasa de lo más concreto a lo más abstracto mediante un proceso de aprendizaje en cascada donde realiza transformaciones no lineales de los datos disponibles en cada nuevo nivel hasta ofrecer una única respuesta de salida [12] (López y Meseguer, 2017: 102).

A pesar del importante paso que suponían las *redes neuronales profundas* para el desarrollo de la IA, éstas no lograron pasar de la teoría a la práctica por la ingente cantidad de datos que precisa su proceso de aprendizaje y las necesidades técnicas necesarias para procesar tal cantidad de datos. Hoy, empero, el fenómeno de la transformación digital —que produce datos masivos de cualquier cosa (IoT y el Big Data— y el aumento exponencial de la potencia de los ordenadores —que permite la recopilación, almacenamiento y procesamiento de todos esos datos— ha permitido el desarrollo y la aplicación de las *redes neuronales profundas* para el aprendizaje, incluso moral, de las máquinas dotadas de algoritmos artificialmente inteligentes [13].

[11] Esta crítica fue abordada por John Searle en «Minds, Brains and Programs» (1980), quien afirmó que la carencia de intencionalidad en las máquinas, a diferencia de las personas, hacía imposible la emergencia de una *IA fuerte*.

[12] Las redes neuronales profundas disponen de diferentes nodos de entrada y procesamiento de datos en cada uno de los niveles, pero sólo un nodo de salida. De esta forma, la transformación de los datos no depende de una neurona artificial, sino de la interacción entre el conjunto de neuronas que los procesa, y la respuesta dada es única.

[13] Este uso de las redes neuronales profundas ha dado lugar a un subcampo de la IA denominado *Aprendizaje profundo* o *Deep learning*.

IV. Neuroaprendizaje moral de las máquinas: sobre valores, parches y emociones moralmente artificiales

Una de las principales preocupaciones actuales de la IA, es la posibilidad de que las máquinas tomen decisiones moralmente válidas. Los coches sin conductor, por ejemplo, han visto ralentizado e incluso paralizado su desarrollo e implementación por las dificultades que tienen los diseñadores de algoritmos artificialmente inteligentes de decidir e integrar el criterio de fundamentación moral más adecuado para hacer frente o salvar los conflictos de valor que subyacen a la práctica de la movilidad inteligente (Monasterio, 2016). Entre otras cuestiones, cuándo hay que poner en riesgo la vida de los pasajeros del vehículo en favor de los viandantes, o viceversa, en caso de accidente. Asimismo, el *gobierno algorítmico* genera mucha desconfianza por las continuas evidencias de mala praxis vinculadas con el sesgo emocional, misógino, xenófobo, homófobo y/o aporófobo de los algoritmos artificialmente inteligentes y la falta de asunción de responsabilidades (O'Neal, 2014; Calvo, 2018b).

Para abordar tal interés, la IA se ha centrado en tres principales líneas de investigación para el desarrollo de *redes neuronales profundas* que sean capaces de captar el sentido de lo moral y actuar en consecuencia: la captación e interiorización de valores, el diseño e introducción de *parches éticos* en los códigos matemáticos, y la adquisición de competencias y capacidades hermenéuticas, argumentativas y emocionales.

Con respecto a la primera de las líneas de investigación, y siguiendo principalmente a Bostrom (2014), actualmente existen diversas propuestas más o menos interesantes sobre cómo pueden la IA abordar un proceso de adquisición e interiorización de valores morales que sirva como marco de referencia para la toma de decisiones: selección evolutiva; refuerzo; asociación de patrones; andamiaje; y evaluación.

a. *Aprendizaje por selección evolutiva*: se trata de emular el proceso de evolución biológica para la adquisición de valores mediante algoritmos de búsqueda que, por un lado, actúan ampliando el catálogo de valores y, por otro, llevan a cabo tareas de selección y potenciación de los mejores y de identificación y eliminación de los peores; es decir, de los que mejor y peor funcionan en el mundo real. Este proceso de doble vía puede equipararse al descrito por el biólogo John Maynart Smith (1982) como *estrategia evolutivamente estable* (EEE) y *dinámica de replicación* (DR), solo que en clave de valores y no de estrategias de comportamiento social. Se trataría de que el algoritmo de búsqueda, por un lado, recopilara del mundo real aquellos valores utilizados en la resolución de conflictos morales que se mantienen inalterables con el paso del tiempo (expansión); por otro, estableciera un orden de materialidad para ir eliminando los valores que peor puntuación sacan en una prueba con una

función de evaluación (contracción); y, finalmente, procesara la información disponible para mutar y/o concretar unos valores finales lo más optimizados posibles para alcanzar los objetivos. El problema de este tipo de asimilación de valores es que, como argumenta Bostrom, «La naturaleza podrá ser una gran experimentadora, pero nunca aprobaría un examen de ética —pues contraviene la declaración Helsinki y todas las normas de decencia moral en todos los sentidos» (Bostrom, 2014: 188). Además, hay que tener en cuenta que se trata de una propuesta convencional y estratégica que instrumentaliza los valores y se basa en un criterio funcionalista de fundamentación de lo moral. El criterio de elección y adquisición de valores no está vinculado con una mejor orientación de la acción en un caso concreto, sino con una mayor adaptación al medio. Además, la falta de crítica en todo el proceso genera un marco axiológico puramente convencional que puede servir para fomentar y perpetuar las injusticias.

b. *Aprendizaje por refuerzo:* se trata de que el algoritmo reciba algún tipo de recompensa, valoración, estímulo o información (del programador o del entorno, por ejemplo) sobre las decisiones que ha tomado. Para ello, se asigna alguna función de evaluación o utilidad que, a través de la experiencia (ensayo–error), permita al algoritmo precisar mejor sus estimaciones sobre el uso práctico de ciertos valores, asociarlos a respuestas y comportamientos que maximizan el beneficio y reforzar o corregir sus decisiones. Por ejemplo, dar un premio (un punto) por cada decisión basada en ciertos valores, pero sin decirle por qué y qué valores son. No obstante, el principal problema de este tipo de aprendizaje por refuerzo radica en que, como afirma Samuel Bowles, los buenos incentivos no hacen buenos ciudadanos (2016). Como se ha demostrado en estudios de campo y experimentos de laboratorio con juegos de estrategia, con el tiempo los beneficiarios de los incentivos se vuelven incapaces de actuar moralmente si no existe expectativas razonables de que serán recompensados por ello. Además, comportarse moralmente no siempre produce recompensas positivas, ni intrínsecas ni extrínsecas, en las personas. No en pocas ocasiones hacer lo que se debe tiene un alto coste, como una menor adaptación, una disminución del bienestar personal o una reducción de los beneficios. El problema es que los seres humanos no sólo tienen una estructura moral, también disponen de la capacidad de criticar lo convencional para discernir qué es lo justo mediante el diálogo y el acuerdo con todos los afectados.

c. *Aprendizaje por asociación:* se trata de asociar valores a conceptos, cosas, procesos, experiencias u objetivos, tal y como sucede cuando, por ejemplo, los seres humanos confieren un conjunto de valores a conceptos como persona, profesión, institución, etc. En el caso de la IA, se trataría de seleccionar, vincular y adaptar valores complejos a objetivos dados; es decir, gobernar con liderazgo,

respeto, honestidad y responsabilidad. El problema aquí es la dificultad para conocer en profundidad el mecanismo que permite a los seres humanos asociar valores a conceptos, vinculado con una arquitectura neurocognitiva que sólo es posible aplicar a una emulación de cerebro completo. La IA actual, empero, es una emulación muy pequeña del cerebro humano (Bostrom, 2014). Además, en mi opinión, por un lado, no está claro cuál sería el referente a utilizar por la IA tanto para determinar los valores que deberían ser asociados a un concepto determinado como para criticarlos y, si lo cree necesario, cambiarlos por otros. Y, por otro lado, tampoco está claro cuál sería la motivación de la IA para elegir, asociar y dejarse guiar por esos valores sin un refuerzo por función de utilidad. Todo parece indicar que, como en las anteriores propuestas, el referente vuelve a ser la funcionalidad —aquellos valores que permitan satisfacer mejor un objetivo dado— y la motivación la maximización del beneficio —el que genera la consecución del objetivo dado—. Por ello, el aprendizaje por asociación vuelve a proponer una instrumentalización de los valores morales y una fundamentación moral basada en la funcionalidad.

d. *Aprendizaje por andamiaje:* se trataría de establecer métodos para la selección de objetivos, fines, motivaciones y normas. Por un lado, se trataría de dotar la IA de un andamiaje de objetivos instrumentales provisionales y finales muy simples que sirvan como horizonte de sentido de la satisfacción de los objetivos instrumentales, pero que sean lo suficientemente simples para dejar a la IA la posibilidad de desarrollarlos o, si lo cree necesario, modificarlos. Por otro, se trataría de conferir un andamiaje motivacional para el desempeño de los objetivos provisionales y los fines. Por tanto, se trata de dotar la IA de una estructura de aprendizaje con unos contenidos mínimos que, a través de la experiencia, la IA vaya modificando conforme sea capaz de abordar una mayor complejidad. No obstante, aunque el aprendizaje por andamiaje resulta muy interesante, actualmente adolece de una base suficientemente sólida para poder ser implementada. Entre sus límites, se halla un posible exceso de control e influencia del programador, lo cual es contrario a la noción actual de IA. Aquí, tanto el horizonte de sentido primario, aunque simple y abierto para que la IA pueda ir desarrollándolo y modificándolo conforme vaya madurando a través de la experiencia que le proporciona la búsqueda de los objetivos provisionales, como las motivaciones son introducidos en gran medida por el programador.

e. *Aprendizaje por valoración:* en este caso se trataría de dar a la IA un criterio para que por sí misma aprenda los valores que se desea que utilice y defina mediante estimaciones la mejor forma de actuar conforme a éstos. El problema de esta propuesta es tanto el criterio a utilizar como la capacidad de disponer de una *IA fuerte* suficientemente evolucionada para detectar y comprender la estructura axiológica del entorno mediante la fabricación de hipótesis cada

> vez más precisas en función de estimaciones basadas en sus principios y los datos empíricos disponibles. Por otro lado, tal y como sucede con el resto de propuestas, no está claro cómo es posible asegurarse «(…) de que la IA se sintiera motivada a perseguir los valores descritos en la forma en que pretendíamos» (Bostrom, 2014: 197); es decir, que la IA conozca e interprete perfectamente el marco axiológico no significa que lo siga. Y, finalmente, hay que tener en cuenta que las estimaciones sobre los valores y los objetivos intermedios y finales se basan en lo que es, no en lo que debería ser.

De este modo, el carácter que subyace a las diferentes propuestas de adquisición y aprendizaje de valores es estrictamente convencional y estratégico, basado en una acrítica observación de las pautas, opiniones y comportamientos de una comunidad concreta o en las estimaciones de los programadores sobre lo que es justo y bueno para la sociedad. El problema es que los valores morales no pueden ser ni convencionales ni estratégicos. Los valores morales son potsconvencionales, y, por tanto, como afirma Adela Cortina, anclan sus raíces en el terreno de la intersubjetividad (2010).

Por otro lado, con respecto a la segunda de las líneas de investigación, el diseño e introducción de *códigos específicos* o *parches éticos* en los códigos matemáticos, destacan estudios como los de Bertram F. Malle (2016), quién identifica los múltiples elementos que conforman la competencia moral humana —vocabulario moral; sistema de normas; cognición moral y afecto; toma de decisiones morales y acción; y comunicación moral— y propone con poco éxito y menor concreción cómo diseñar robots que tengan uno o más de estos aspectos humanos.

Finalmente, con respecto a la tercera de las líneas de investigación, la adquisición de competencias y capacidades hermenéuticas, argumentativas y emocionales, se han logrado resultados poco significativos y vinculados con una *IA débil*. Cabe destacar algunos avances en capacidades dialécticas, como el caso de las Project Debater de IBM, un algoritmo que en 2018 ganó por primera vez a una persona en un concurso de debate y que en 2019 ha estado a punto de ganar a Harish Natarajan, el campeón del mundo de esta modalidad (*Expansión*, 2019). También un estudio realizado por el Centro para la Ciencia Cognitiva de la Universidad Técnica de Darmstadt (Jentzsch *et al.*, 2019), que, mediante la incorporación a una *red neuronal profunda* de grandes cantidades de textos escritos para aprender representaciones vectoriales de palabras, ha logrado que las máquinas artificialmente inteligentes pueden extraer de los textos razonamientos sobre conductas correctas e incorrectas. Finalmente, los avances realizados en el reconocimiento de emociones en los seres humanos por parte de las máquinas artificialmente inteligentes, principalmente fisiológicas, y su reacción «empática» (Boden, 2017: 76–77), y, sobre todo, los estudios con *robots asistenciales* dotados de aprendizaje profundo, cuya tarea exige que sean sociales, es decir, que interactúen y colaboren de forma proactiva con humanos para el correcto desarrollo de su trabajo.

En definitiva, como afirma Bostrom (2014: 207), «La ingeniería de objetivos no es una disciplina establecida. No se sabe actualmente cómo transferir los valores humanos a un ordenador digital, incluso contando con un nivel humano de inteligencia artificial». Y, lo que es peor, aunque en el futuro se llegase a conocer cómo la IA puede recopilar y almacenar valores, existen muchas cuestiones relacionadas con la moralidad humana que ni siquiera se están trabajando. Por ejemplo, la exigencia recíproca del reconocimiento de las capacidades comunicativas y afectivas de las partes en relación; la capacidad de discernir entre vigencia y validez, el sentido común o la posibilidad de reproducir y desarrollar una conciencia, motivación, voluntariedad, emotividad y responsabilidad moral. Porque una cosa es tener la competencia de adquirir los valores que imperan en la sociedad y otra muy distinta es ser competente y capaz de criticarlos para validarlos, revisarlos o rechazarlos; de emocionarse por su acontecer práctico; de sentir mala conciencia o indignación cuando se violan; de apreciar su acontecer y seguimiento, de autoexigirse un comportamiento adecuado; o de responder de las decisiones que tomamos y los comportamientos que tenemos, entre otras muchas cosas.

V. Cuestiones éticas alrededor del neuroaprendizaje moral de las máquinas

Como argumenta Morozov, la tendencia actual es hacia la recreación de una sociedad heterónoma que deja en manos de los algoritmos los procesos de toma de decisiones políticas y económicas, pero también cotidianas: «Gracias a los teléfonos inteligentes o Google Glass, ahora podemos ser molestados cuando estamos a punto de hacer algo estúpido, poco saludable o poco sólido. No necesariamente necesitaríamos saber por qué la acción sería incorrecta: los algoritmos del sistema hacen el cálculo moral por sí mismos» (Morozov, 2013).

Esta nueva realidad, derivada de la actual transformación digital, aumenta la vulnerabilidad de las personas y, por tanto, incrementa proporcionalmente la exigencia de saber moral, cuya función social es precisamente, tal y como argumenta Domingo García–Marzá (2016: 882–883; 2005: 254–256), minimizar la vulnerabilidad y sus consecuencias aportando recursos que permiten orientar la praxis para evitar la emergencia de conflictos y coordinar la acción tendente a la satisfacción de objetivos comunes y a la resolución de la conflictividad subyacente a la práctica política y empresarial. Es decir, entre las personas que trabajan el aprendizaje y comportamiento moral de las máquinas no parece que exista un interés por ese saber moral [14] que intuitivamente hacen servir los participantes en aquellos procesos de diálogo tendentes al entendimiento intersubjetivo sobre algo en el mundo, que

[14] Para un estudio sobre el saber moral y su función social, ver García–Marzá (2005, 2016).

ahora parece quedar reducido a cursos de acción socialmente atomizados y modelos matemáticos de recopilación y procesamiento de la información cuantificable y disponible. De ahí el aumento exponencial del interés que ha mostrado la IA por la ética en los últimos años.

En este sentido, entre las principales críticas sobre la posibilidad de un aprendizaje moral de la maquinas dotadas de IA que posibilite el diseño e implementación de gobiernos algorítmicos a la altura de lo esperado y deseado en ámbitos como la política y la economía, destacan tres:

1. Las propuestas actuales de aprendizaje moral de las máquinas se preocupan mucho de la aplicación, desarrollo e implementación de un mecanismo que permita la adquisición de valores, pero muy poco de otras partes importantes y necesarias de la estructura moral de los seres humanos. Nada o poco se dice, por ejemplo, de cómo concretar una estructura moral artificial que emule los juicios morales sobre lo justo o injusto; que logre distinguir entre valores morales y sociales; que pueda interpretar y criticar tanto el conocimiento como la acción; que genere *emociones prosociales* capaces de promover el seguimiento de los valores con independencia de los beneficios derivados de su implementación; que tenga sentido común a la hora de aplicar a los contextos concretos las normas y valores morales, entre otras cosas. En este sentido, buena parte de las propuestas actuales establece una especie de sucedáneo de juicios morales artificiales basado en una función de utilidad respecto a fines, no en el seguimiento de principios, valores o normas cuyas consecuencias derivadas de su aplicación podrían ser aceptadas por todos los afectados en un discurso práctico con ciertas reglas lógicas y un principio moral procedimental (Cortina, 1993: 208).
2. Las propuestas actuales de aprendizaje moral de las máquinas se preocupan muy poco del contenido de lo moral, especialmente de cómo se concretan, justifican y enriquecen los valores, normas, principios y virtudes morales que permiten un correcto y legítimo funcionamiento de la estructura. Da la impresión de que tal desinterés radica en la falta de confianza hacia la actual IA. De ahí el exceso de paternalismo observado en la mayoría de las propuestas. No deja de ser curioso que, mientras se ensalza la capacidad autónoma y creativa de AlphaGo para diseñar y aplicar estrategias nunca antes pensadas por los seres humanos, se controle con parches la posible capacidad creativa de la IA para evitar que puedan enriquecer los valores existentes con argumentos nunca antes propuestos por los seres humanos. Parece evidente que, por un lado, se reconoce la imposibilidad de abordar la complejidad de lo moral desde el formalismo matemático; por otro, se reconoce la inadecuación entre racionalidad computacional y racionalidad práctica; y, finalmente, se muesttra un alto grado de desconfianza hacia la posibilidad de dejar el rumbo de la

humanidad en manos de la IA. Esta desconfianza generalizada en la IA obliga a los programadores a insertar *parches* para evitar un posible *pinchazo cerebral* y *funciones de utilidad* basadas en la agregación de puntuaciones por objetivos intermedios y la maximización del beneficio para prevenir un posible abandono de los objetivos. De este modo, se consigue promover el seguimiento cuasi escrupuloso de los objetivos y las normas dadas.

3. En los seres humanos, la motivación para desear adquirir y seguir los principios, valores y normas morales emana del deber de cumplir con aquello que tenemos buenas razones para considerar como justo y deseable. Sin embargo, la IA construye una estructura motivacional basada en el grado de implicación de los valores y normas en la satisfacción de objetivos dados a través de una función de utilidad preestablecida, o en la estimación sobre los valores realizada por los programadores. De este modo, el criterio de decisión de la mayoría de técnicas de adquisición y aprendizaje de valores es la maximización del beneficio, con lo cual la IA instrumentaliza los valores para mejorar los resultados. Por consiguiente, tras esta Inteligencia Moral Artificial (IMA) subyace un déficit importante de razones emocionales que dificulta o imposibilita su concreción y desarrollo.

En definitiva, como ha manifestado Margaret A. Boden, una de las mayores expertas en IA, «Pensar que con la IA puedes resolver conflictos tan humanos como el de Oriente Próximo (y ya puestos el de Irlanda del Norte o el de Cataluña) es totalmente ridículo. Los que creen en la *singularidad* ignoran las limitaciones de la IA actual. Se basan sólo en los avances tecnológicos exponenciales, pero ignoran un hecho: el aumento de potencia de los ordenadores y de la disponibilidad de datos no garantiza una IA de nivel humano» (Fresneda y Alba, 2018).

Además, más allá de la *singularidad*, se hace necesario reflexionar sobre las consecuencias de la aplicación y desarrollo de una *IA débil*, puesto que, como argumenta Pierpaolo Donati, de ella está emergiendo un ecosistema socialmente dividido y altamente conflictivo donde, «Mientras algunos celebran el nuevo mundo de las tecnologías digitales (TIC, inteligencia artificial, robots) como la llegada del "hombre aumetado", otros piensan que este mundo nos conduce hacia una mayor deshumanización, en el sentido de que, frente a lo humano, comporta una vida social "sin cualidad"» (Donati, 2019: 31–32).

Bibliografía

ABC (2018). «Cuestionada la máquina que predice futuros delitos en EE.UU». *ABC*, 01 de enero de 2018 [https://www.abc.es/ciencia/abci–cuestionada–maquina–predice–futuros–delitos–eeuu–201801181039_noticia.html].

BBC (2014). «El robot que escribe noticias debuta en Los Ángeles». *BBC Mundo*, 19 de marzo de 2014 [https://www.bbc.com/mundo/noticias/2014/03/140318_curiosidades_robot_periodista_la_times_az].

Boden, Margaret A. (2017). *Inteligencia Artificial*. Madrid: Turner Noema.

Bostrom, Nick (2016). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Burridge, Nicky (2017). «Artificial intelligence gets a seat in the boardroom». *Nikkei Asian Review*. [https://asia.nikkei.com/Business/Artificial–intelligence–gets–a–seat–in–the–boardroom].

Collective Learning group (2019). *Augmented Democracy. Exploring the design space of Collective decisions*. [https://www.peopledemocracy.com/].

Bowles, Samuel (2016). *The moral economy: why good incentives are no substitute for good citizens*. Yale University Press.

Calvo, Patrici (2020). «Democracia aumentada: un *ecosistema ciberético* para el desarrollo práctico de la participación política basada en algoritmos». *Revista de Sociologia e Politica* [forthcoming].

— (2019a). «Una propuesta de línea ética basada en tecnología blockchain». En Andrés, Alicia y Sanhauja, Rosana, *Transparencia e integridad en la institución universitaria*. Castellón: Universitat Jaume I.

— (2019b). «Democracia algorítmica: consideraciones éticas sobre la datificación de la esfera pública». *Revista de CLAD. Reforma y Democracia,74,* 5–30.

— (2019c). «The ethics of Smart City (EoSC): moral implications of hyperconnectivity, algorithmization and the datafication of urban digital society». *Ethcis and Information technology* [https://doi.org/10.1007/s10676–019–09523–0].

— (2018a). «Ética de las cosas (EoT). Hacia una digitalización socialmente responsable y moralmente válida del ámbito universitario». En Alicia Andrés y Rosana Sanahuja (Eds.), *Un diseño universitario para la responsabilidad social*. Castellón: Universitat Jaume I.

— (2018b). *The cordial Economy – Ethics, Recognition and Reciprocity*. Cham: Springer.

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

— (2018). *Draft Ethics Guidelines for Trustworthy AI.* Bruselas: European Commission.

Conill, Jesús (2019). *Intimidad corporal y persona humana. De Nietzsche a Ortega y Zubiri*. Madrid: Tecnos.

Conill, Jesús (2006). *Ética Hermenéutica. Crítica desde la Facticidad.* Madrid: Tecnos.

Cortina, Adela (2010). *Justicia cordial*. Madrid: Trotta.

— (1993). *Ética aplicada y democracia radical*. Madrid: Tecnos.

Dastin, Jeffrey (2018). «Amazon scraps secret AI recruiting tool that showed bias against women». *Reuters* [https://www.reuters.com/article/us–amazon–com–jobs–automation–insight/amazon–scraps–secret–ai–recruiting–tool–that–showed–bias–against–women–idUSKCN1MK08G?feedType=RSS&feedName=topNews&utm_source=twitter&utm_medium=Social].

Donati, Pierpaolo (2019). *Sociología relacional de lo humano.* Barañain: EUNSA.

Expansión (2019). «Un humano vence en debate a robot de IBM... que le dio batalla». [https://expansion.mx/tecnologia/2019/02/12/un–humano–vence–en–debate–a–robot–de–ibm–que–le–dio–batalla].

Fresneda, Carlos y Alba, Carlos (2018). «Margaret Boden: La Inteligencia Artificial, como el hacha, se puede usar para el bien o para el mal». *El Mundo* [http://lab.elmundo.es/inteligencia–artificial/margaret–boden.html].

García–Marzá, Domingo (2016). «Neuroética aplicada: las consecuencias prácticas del neuropositivismo». *Pensamiento, 72*(273), 881–900.

— (2005). «¿Sentimientos virtuosos? El papel de los sentimientos en la vida moral». *Diálogo Filosófico,* 62, 241–256.

Gödel, Kurt (1931). «Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I». *Monatshefte für Mathematik und Physik, 38*(1): 173–198.

Hidalgo, César A.; Orguian, Diana; Albo–Canals, Jordi; De Almeida, Filipa y Martín, Natalia (2020). *How Human Judge Machines*. Cambridge: Massachusetts, The MIT Press.

Jentzsch, Sophie; Schramowski, Patrick; Rothkopf, Constantin y Kersting, Kristian (2019). «Semantics Derived Automatically from Language Corpora Contain Human–like Moral Choices». En *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. Consultado en [https://ml–research.github.io/papers/jentzsch2019aies_moralChoiceMachine.pdf].

López, Ramón y Meseguer, Pedro (2017). *Inteligencia Artificial*. Madrid: Catarata.

L'Oreal (2018). «Inteligencia Artificial para reclutar a los mejores candidatos de L'Oreal España». *L'Oreal,* 6 de junio de 2018 [http://www.loreal.es/periodistas/notas–de–prensa/2018/jul/inteligencia–artificial–para–reclutar–a–los–mejores–candidatos–de–loreal–espana].

Luna, Juan Pablo y Pérez–Muñoz, Cristian (2018). «¿Democracia sin políticos? La engañosa fe en los algoritmos». *Ciper*, 9 de mayo de 2018 [https://ciperchile.cl/2018/05/09/democracia–sin–politicos–la–enganosa–fe–en–los–algoritmos/].

Malle, Bertram F. (2016). «Integrating robot ethics and machine morality: the study and design of moral competence in robots». *Ethics Inf Technol,* 18, 243–256.

Marr, Bernard (2016). *Big Data. La utilización del Big Data, el análisis y los parámetros Smart para tomar mejores decisiones y aumentar el rendimiento*. Teell.

Matsumoto, Tetsuzo (2018a). *4 de abril de 2018*, 4: 12. Twitter.

Matsumoto, Tetsuzo (2018b). *The Day AI Becomes God. The Singularity will Save Humanity*. Cambridge (NZ): Media Tectonics.

Monasterio, Aníbal (2017). «Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos». *Dilemata,* 24, 185–217.

Morozov, Evgeny (2013). «The Real Privacy Problem». *MIT Technology Review*. Consultado en: [https://www.technologyreview.com/s/520426/the–real–privacy–problem/].

O'Neil, Cathy (2016). *Weapons of Math Destruction How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publisher.

Palacín, José–Tomás (2018). «Qué áreas tienen que potenciar las empresas en IA (según Microsoft Ibérica)». *Innovaspain*, 22 de marzo de 2018 [https://www.innovaspain.com/microsoft–iberica–inteligencia–artificial–empresas/].

Pardo, Pablo (2014, 9 de junio). «Un algoritmo, sentado en el consejo de un fondo chino». *El Mundo* [https://www.elmundo.es/economia/2014/06/09/5394e0d722601df76f8b458f.html].

RAE (2018). «Algoritmo». [https://dle.rae.es/srv/search?m=30&w=algoritmo].

Rubio, Isabel (2018). «Amazon prescinde de una inteligencia artificial de reclutamiento por discriminar a las mujeres». *El País*, 12 de octubre de 2018 [https://elpais.com/tecnologia/2018/10/11/actualidad/1539278884_487716.html].

Rumelhart, David E.; Hinton, Geoffrey E. y Williams, Ronald J. (1986). «Learning representations by back–propagating errors». *Nature,* 323, 533–536.

Sáez, Cristina (2018). «Augmanted democracy». *CCCBLAB, Cultural Research and Innovation* [http://lab.cccb.org/en/democracia–augmentada/].

Sagristà, Anna (2016). «Algoritmos: La bola de cristal en gestión de personal». *Talentier,* 16 de diciembre de 2018 [https://blog.talentier.com/algoritmos–en–gestion–de–personal].

Saiz, Eva (2013). «La NSA infringió las normas de privacidad en miles de ocasiones». *El País* [https://elpais.com/internacional/2013/08/16/actualidad/1376631278_378738.html].

SAM (2018). *Web SAM*, 20 de junio de 2018 [http://www.politiciansam.nz].

Searle, John R. (1980). «Minds, Brains and Programs». *Behavioral and Brain Sciences, 3*(3): 417–457.

Stats NZ (2018). *Algorithm assessment report*. New Zealand Government: New Zealand [https://data.govt.nz/use–data/analyse–data/ government–algorithm–transparency].

Vidal, Macarena (2018). «China estrena presentadores artificiales de televisión». *El País,* 9 de noviembre de 2018 [https://elpais.com/tecnologia/2018/11/09/actualidad/1541765605_369415.html].

Whiters, Paul (2018). «Robot to run for mayor in Japan promising "fairness and balance"». *Express* [https://www.express.co.uk/news/world/947448/robots–japan–tokyo–mayor–artificial–intelligence–ai–news].

Walker, Joseph (2012). «A la hora de contratar, las empresas sustituyen intuición por algoritmos». *The Wall Street Journal*, 21 de septiembre de 2012 [https://www.wsj.com/articles/SB10000872396390444032404578008900640503088].

Xerox (2018). «Solución de automatización de flujos de trabajo de Xerox para el reclutamiento y la selección» [https://www.xerox.com/es–ni/servicios/automatizacion–de–procesos/reclutamiento–y–seleccion–de–personal].