La ética de los agentes cibernéticos (Una ética de plástico para seres de plástico)*

The ethics of cybernetic agents (A plastic ethics for plastic beings)

Miguel L. Lacruz Mantecón

Profesor Titular de Derecho Civil en la Facultad de Derecho de la Universidad de Zaragoza

DOI: 10.14679/2036

Resumen / Abstract: Entre las cuestiones que suscita la nueva tecnología de la IA tenemos la llamada Ética de la IA. De hecho, la reciente Resolución UE *Marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas* de 20 de octubre 2020, declara que las cuestiones de carácter ético y jurídico relacionadas con la inteligencia artificial deben abordarse a través de un marco regulador del Derecho de la Unión efectivo, global y con visión de futuro que refleje los principios y valores de la Unión consagrados en los Tratados y en la Carta de los Derechos Fundamentales. Es decir, se está pidiendo una reglamentación de la ética europea de la IA. Además, se adjunta como Anexo una Propuesta de Reglamento del Parlamento Europeo y del Consejo sobre los principios éticos para el desarrollo, el despliegue y el uso de la IA, en 24 artículos. Este trabajo explora el significado de esta Ética para robots, Roboética o Ciberética, distinguiendo la ética que se dirige a los seres humanos en su relación con las máquinas de la que se programa en el sistema inteligente como medida de control y seguridad.

* El presente trabajo se ha realizado al amparo del Proyecto «Derecho e inteligencia artificial: nuevos horizontes jurídicos de la personalidad y la responsabilidad robóticas», IP. Margarita Castilla Barea, (PID2019-108669RB-100 / AEI / 10.13039 / 501100011033).

Among the issues raised by the new AI technology we have the so-called Ethics of AI. In fact, the recent EU *Resolution Framework on the ethical aspects of artificial intelligence, robotics and related technologies* of October 20, 2020, declares that ethical and legal issues related to artificial intelligence must be addressed through a framework an effective, comprehensive and forward-looking regulator of Union law that reflects the principles and values of the Union enshrined in the Treaties and in the Charter of Fundamental Rights. In other words, a regulation of the European ethics of AI is being requested. In addition, a Proposal for a Regulation of the European Parliament and of the Council on the ethical principles for the development, deployment and use of AI, in 24 articles, is attached as an Annex. This work explores the meaning of this Ethics for robots, Roboethics or Cyberethics, distinguishing the ethics that addresses human beings in their relationship with machines from those that are programmed in the intelligent system as a control and security measure.

Palabras clave / Keywords:

Inteligencia artificial; Ética; Robots; Ciberética; Nuevas tecnologías.

Artificial intelligence; Ethics; Robots; Cyberethics; New technologies.

1. La Ética como uno de los retos jurídicos para la IA

El planteamiento de una regulación de la tecnología de la IA implica para muchos autores el surgimiento de una nueva disciplina jurídica; en este sentido, Barrio Andrés coordina un *Derecho de los robots*[1], cuya razón principal de ser está en brindar un marco de reglas jurídicas que pueda conferir certeza respecto de los deberes y de las responsabilidades "…de los actores involucrados en el proceso de innovación robótica". Pero, por otro lado, este Derecho debe también garantizar, añade, "… el desarrollo de la robótica en un entorno que respete los valores propios del ordenamiento jurídico europeo, con pleno respeto a los derechos fundamentales consagrados en la Carta de los Derechos Fundamentales de la Unión Europea (CDFUE) y en el Convenio Europeo para la Protección de los Derechos Humanos y de las Libertades Fundamentales (CEDH)". Lograr este objetivo es lo que se proponen los juristas europeos mediante la implementación de una ética de la IA, siendo fundamentales los textos europeos centrados en los aspectos éticos de la IA, por lo que resulta imprescindible una referencia a los recientes textos europeos[2].

Ya el denominado *Robolaw Project*, que se inserta en el Séptimo Programa Marco de Investigación y Desarrollo Tecnológico 2012-2014, produce el documento de trabajo D6.2 *Guidelines on Regulating Robotics*, que valora

[1] Barrio Andrés, Moisés, "Del derecho de internet al derecho de los robots", en *Derecho de los robots*, Moisés Barrio Andrés (dir.), Wolters Kluwer, 2018.

[2] Una recopilación más amplia sobre los principales textos europeos sobre IA se encuentra en mi monografía *Robots y personas*, Reus, Madrid, 2020.

la posibilidad de comportamientos emergentes e impredecibles de la IA, y se insiste en que una ética de la máquina puede instalar en los sistemas autónomos una capacidad de razonamiento moral, para que puedan enfrentar situaciones inesperadas. También el Reglamento General de Protección de Datos (Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016), intenta controlar la toma automatizada de decisiones para evitar discriminación y sesgos, y una mayor explicabilidad y transparencia del algoritmo.

Posteriormente, se emite la muy comentada *Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica*, famosa sobre todo por valorar la conveniencia de una personalidad jurídica para robots. Contiene una *Carta sobre robótica* anexa, elaborada con la asistencia de la Unidad de Prospectiva en Ciencia y Tecnología (*Science and Technology Options Assessment* – STOA) del *European Parliament Research Service,* en la que se proponen unos Principios generales para la robótica, con un apartado dedicado a los *Principios éticos.* Se considera en dicha *Carta* que «*…es preciso un marco ético claro, estricto y eficiente que oriente el desarrollo, diseño, producción, uso y modificación de los robots …un marco en forma de carta integrada por un código de conducta para los ingenieros en robótica, un código deontológico destinado a los comités de ética de la investigación para la revisión de los protocolos de robótica, y licencias tipo para los diseñadores y los usuarios*». Asimismo fija algunas orientaciones éticas, como los principios de beneficencia, no maleficencia (el principio hipocrático *primum non nocere*), autonomía y justicia, haciendo referencia a la Carta de los Derechos Fundamentales de la Unión Europea, y a principios como la dignidad humana, la igualdad, la justicia y la equidad, la no discriminación, el consentimiento informado, la vida privada y familiar y la protección de datos, así como a valores *inherentes al Derecho de la Unión.*

Con el precedente de la *Comunicación de la Comisión «Inteligencia artificial para Europa»*, de abril de 2018, se elabora un proyecto de directrices éticas en la *Comunicación de la Comisión al Parlamento Europeo, el Consejo, el Comité Económico y Social europeo y el Comité de las Regiones* de 8 de abril de 2019 que lleva por título «*Generar confianza en la Inteligencia artificial centrada en el ser humano*» (COM(2019)168), de 8 de abril 2019. Requisito necesario para esta generación de confianza es la implantación de directrices éticas en los sistemas de IA, y se enumeran los requisitos que dichas directrices deben cumplir: «• *Intervención y supervisión humanas.* • *Solidez y seguridad técnicas.* • *Privacidad y gestión de datos.* • *Transparencia.* •

Diversidad, no discriminación y equidad. • *Bienestar social y medioambiental.* • *Rendición de cuentas*». La Comunicación se complementa con el *Estudio «Directrices Éticas para una IA fiable*», *Ethics guidelines for trustworthy AI*, elaborado por el Grupo de Expertos de alto nivel en IA, también del 8 de abril de 2019. Este Estudio tiene como objetivo promover una IA en la que se pueda confiar. Y la fiabilidad de la IA se apoya en tres requisitos: «a) la IA debe ser lícita, es decir, cumplir todas las leyes y reglamentos aplicables; b) ha de ser ética, de modo que se garantice el respeto de los principios y valores éticos; y c) debe ser robusta». El documento enuncia una serie de principios que pueden conformar una suerte de código ético para la IA: Respeto a la dignidad humana y libertad individual, a la democracia, la justicia y el Estado de Derecho, derecho a la igualdad, no discriminación, solidaridad, y derechos políticos de los ciudadanos. Cuatro principios que actúan específicamente como imperativos éticos para los profesionales de la IA son los de: I) respeto de la autonomía humana; II) prevención del daño; III) equidad y IV) explicabilidad.

Luego tenemos el *Libro blanco sobre la Inteligencia artificial – A European approach to Excellence and Trust*, de febrero de 2020, que intenta que la «IA europea» esté basada en valores y derechos humanos como los de la dignidad humana y protección de la privacidad. El *Paper* presenta una serie de propuestas para un desarrollo seguro y «confiable» de la IA, que respete los valores y derechos de la ciudadanía europea, partiendo de la ya citada Comunicación de la Comisión de 8 de abril de 2019 «*Generar confianza en la Inteligencia artificial centrada en el ser humano*», y el también citado Estudio «Directrices Éticas para una IA fiable». La idea es vigilar mediante regulación los riesgos que la IA produce para los derechos fundamentales, y esta vigilancia y control de riesgos se quiere llevar a cabo mediante una ética de los sistemas inteligentes.

Por último, o mejor por ahora, se aprueba el 20 de octubre de 2020 la Resolución *Marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas*. Esta Resolución destaca la necesidad de completar los sistemas jurídicos europeos con las nuevas posibilidades que abre la tecnología de la IA, añadiendo además que *las cuestiones de carácter ético y jurídico relacionadas con la inteligencia artificial deben abordarse a través de un marco regulador del Derecho de la Unión efectivo, global y con visión de futuro que refleje los principios y valores de la Unión consagrados en los Tratados y en la Carta de los Derechos Fundamentales de la Unión Europea.* Es decir, se está pidiendo regular uniformemente cuestiones éticas para toda la Unión, o lo que es lo mismo, la elaboración de una reglamentación

de la ética europea de la IA. Además adjunta como Anexo una Propuesta de Reglamento *sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas*, en 24 artículos.

Completan los textos citados varios estudios técnicos, como el redactado por la Unidad de Prospectiva Científica (STOA) del *European Parliament Research Service, que trata de La ética de la Inteligencia Artificial: Cuestiones e Iniciativas*, de marzo 2020. En cuanto a la personalidad de los robots, recapitula que la mayoría de las investigaciones sobre ética de la IA *parecen estar de acuerdo en que las máquinas de IA no se les debe reconocer una agencia moral, ni ser consideradas como personas"*. O el estudio *European framework on ethical aspects of artificial intelligence, robotics and related technologies* del *European Parliamentary Research Service,* de septiembre 2020, que a pesar de su título contiene importantes datos económicos (aparece auspiciado por la *European Added Value Unit*). Va precedido de una evaluación del valor añadido que podría generarse para la UE si se produce un enfoque conjunto sobre los aspectos éticos de la IA, la robótica y las tecnologías relacionadas: «…la conclusión clave de este Estudio es que un enfoque común de la UE sobre los aspectos éticos de la IA tiene el potencial de generar hasta 294.900 millones de euros en PIB adicional y 4.6 millones de puestos de trabajo adicionales para la Unión Europea para 2030». En este Estudio la ética va de la mano del dinero, y ninguno de los dos se muestra enfadado.

2. Sentido y finalidad de la Ética cibernética

Podemos ahora preguntarnos qué sentido y finalidad tienen estas referencias, y cómo se logra, una IA ética. Para empezar, aceptaremos la IA como concepto que describe el comportamiento de determinadas máquinas que replican el comportamiento inteligente de los seres humanos, lo que permite calificar a dicha conducta como actividad inteligente. Bryson y Winfield[3] nos dicen que *Inteligencia* es la capacidad de hacer lo adecuado en el momento adecuado, en un contexto en el que no hacer nada (o no cambiar de conducta) sería peor; la IA es esta misma capacidad demostrada por artefactos no biológicos.

[3] Bryson, J. J., y Winfield, A., "Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems", *Computer*, vol. 50, nº 5, https://doi.org/10.1109/MC.2017.154.

Los humanos tenemos tendencia a negar dicha IA afirmando que las actuaciones presuntamente inteligentes de las máquinas no lo son porque carecen de conciencia, en el sentido de que no generan una experiencia consciente, con sentido y significado para el ente que la realiza[4]. Frente a este argumento replica HARARI[5] que "la inteligencia es obligatoria, pero la conciencia es opcional", es decir determinadas actuaciones que requieren inteligencia no requieren sin embargo conciencia. Esto lo podemos ver con un ejemplo: Cuando termine este trabajo, para ordenar la bibliografía me limitaré a pulsar la tecla para ordenar alfabéticamente marcada como AZ y ubicada en la pestaña "Inicio" del programa Microsoft Word. La tarea de ordenar alfabéticamente las obras por autores es una tarea que, sin duda, requiere inteligencia. Y desde luego mi pobre ordenador no tiene conciencia de la importancia de su tarea y de lo que a mí me alivia no tener que hacerla, ni esa ordenación es para él una experiencia consciente (por lo menos, no me lo ha dado a entender).

Esta falta de conciencia da lugar a que nos resulte problemático encajar la Ética en el funcionamiento de la IA, porque los seres humanos gestionamos nuestro comportamiento moral usando la conciencia: En cuestiones de moral, la conciencia no es opcional, aunque quizá en el ámbito de la IA el término ética tenga otro sentido, como se ha venido señalando en los textos antes examinados.

De acuerdo con el Diccionario de la RAE, en su 4ª acepción, Ética es el "Conjunto de normas morales que rigen la conducta de la persona en cualquier ámbito de la vida". Idea ésta más descriptiva que la definición técnica de la Ética como parte de la Filosofía dedicada a la reflexión sobre la moral. Como dicen CORTINA Y MARTÍNEZ[6], "la Ética pretende desplegar los conceptos y los argumentos que permitan comprender la dimensión moral de la persona humana en cuanto tal dimensión moral, es decir, sin reducirla a sus componentes psicológicos, sociológicos, económicos o de cualquier otro tipo". Ahora bien, la ética la vemos referida a seres humanos; en cuanto al significado de la expresión "ética de la IA", en el citado *Estudio «Directrices Éticas para una IA fiable»*, que acompaña a la *Comunicación «Generar confianza en la IA centrada en el ser humano»*, y que define la «ética de la

[4] Esta es la conclusión del popular "Experimento de la habitación china", imaginado por John SEARLE.

[5] HARARI, Yuval Noah, *Homo Deus. Breve historia del porvenir,* Debate, Barcelona, 2016, pág. 303.

[6] CORTINA, Adela, y MARTÍNEZ, Emilio, Ética, Akal, Madrid, 1998, pág. 9.

IA» como «subcampo de la ética aplicada que estudia los problemas éticos que plantea el desarrollo, despliegue y utilización de la IA».

Como señalan VAN WYNSBERGHE y ROBBINS[7], el término «ética» referido a las máquinas fue utilizado por primera vez en 1987 por Mitchell Waldrop en el artículo de la revista *AI (Artificial Intelligence)* titulado «A Question of Responsability». Es el italiano Gianmarco VERUGGIO quien va a utilizar la expresión *Roboética* (*Roboethics*) para designar el estudio de las complejas relaciones entre los robots y la sociedad. Susan Leigh y Michael Anderson definen la ética de las máquinas o de la IA como «… un campo de estudio dedicado a la entidad computacional como entidad moral». Si de lo que se trata por tanto es de solucionar los problemas éticos derivados del uso de la IA, y para ello establecer una serie de reglas o normas que guíen la conducta de los programadores y utilizadores de entes dotados de IA, estaríamos hablando más bien de una ética aplicada, de códigos deontológicos para los profesionales de la IA y de códigos éticos o morales para los utilizadores de IA. Efectivamente, la Ética de la IA fue inicialmente entendida no como un estudio filosófico de la bondad o maldad de la IA en sí y sus comportamientos o resultados, sino como el estudio del establecimiento de reglas de conducta para los seres humanos que crean o utilizan sistemas de IA.

En cualquier caso, este estudio de las reglas conductuales y comportamientos entre robots y humanos que pueden recibir calificaciones morales, puede proyectarse en dos direcciones distintas: la del comportamiento del humano hacia el robot, tanto si estamos ante un diseñador o programador como ante un mero usuario, y la del comportamiento del robot mismo en el medio social. Por esto se diferencian dos ámbitos, el de la ética del maquinista y el de la ética de la máquina, como señala GARCÍA INDA[8], o como hace notar el físico LATORRE SENTÍS[9], al hablar de una "ética para máquinas". COECKELBERGH[10] prefiere diferenciar la *agencia moral*, como conjunto de capacidades de actuación moral conferidas al robot, de la *paciencia moral*, que es la ética que debe presidir la relación del ser humano con el robot.

---

7 VAN WYNSBERGHE, Aimee, y ROBBINS, Scott, «Critiquing the Reasons for Making Artificial Moral Agents», *Science and Engineering Ethics,* volume 25, (2019), pág. 721.

8 GARCÍA INDA, «Ética y derecho: Big data e inteligencia artificial (IA)», *ponencia en Reunión del Grupo IDDA (Investigación y Desarrollo del Derecho civil de Aragón) del 25 de octubre 2019, Facultad de Derecho de Zaragoza.*

9 LATORRE SENTÍS, José Ignacio, Ética para máquinas, Ariel, Barcelona, 2019, pág, 61.

10 COECKELBERGH, Mark, Ética de la Inteligencia artificial, Cátedra, Madrid, 2021, pág. 51.

3. La ética de la máquina y la ética del maquinista

La ética aplicada al sistema inteligente ha sido una preocupación desde los inicios de la robótica. No hay que olvidar que el padre de la Cibernética, Norbert Wiener, publica su segunda obra sobre la materia bajo el título *The Human Use of Human Beings: Cybernetics and Society* (1950) invocando la mecanización como forma de liberación de los trabajos repetitivos y alienantes, y en la portada de dicha obra, como subtítulo, se lee: «El cerebro mecánico y otras máquinas similares pueden destruir los valores humanos o nos pueden permitir realizarlos como nunca fue posible».

De acuerdo con lo que hemos visto, la expresión «ética robótica» admite varios significados, como señala Belloso Martín[11] en base a Lin, Abney y Bekey: «El primero de ellos hace referencia a la ética profesional de los ingenieros dedicados a la robótica… El segundo significado se encamina hacia un código moral programado dentro de cada robot. Y el tercero podría dirigirse hacia una habilidad auto-consciente de los robots para que procesen la información de una manera ética». En realidad, estos tres significados pueden reducirse a dos: una ética del constructor del robot, o del utilizador del mismo (en cuanto pueda dirigir la conducta robótica) y una ética entendida como un algoritmo que permita resolver el problema de la ejecución de un comportamiento moral en todo momento. La primera, ética del maquinista, guía la conducta del creador de la máquina inteligente; la segunda, la conducta de la máquina misma. Y en ambos casos estamos ante una ética aplicada, es decir, ante códigos de conducta, instrucciones y respuestas conductuales que se ajustan a los parámetros de la moral. En principio, dada la dirección humana del comportamiento robótico a través de la programación, la ética del maquinista y la de la máquina seguirían un mismo orden de valores: el sistema de reglas de conducta que gestionaría el comportamiento del robot sería ético porque el robotista está obligado a respetar un código ético y no puede programar un robot que sea maligno. Todo se reduciría a un problema de predicción de la conducta robótica y programación de límites éticos. Como dice García Inda[12]: «Si la ética del maquinista es benévola, como se exige, la IA será buena. La *moralidad* de la IA depende o está en función de la moralidad de quien la programa».

---

[11] Belloso Martín, Nuria, «La Necesaria Presencia de la Ética en la Robótica: La Roboética y su Incidencia en los Derechos Humanos», *Cadernos do Programa de Pós-Graduação em Direito, v. 13, n. 2 (2018), Universidade Federal do Rio Grande do Sul. DOI: https://doi.org/10.22456/2317-8558.90165*, pág. 105.

[12] García Inda, «Ética y derecho: Big data e inteligencia artificial (IA)», *…cit.*

La doctrina más general, resumen SIAU y WANG[13], define a la Ética de la IA como la parte de la Ética de las nuevas tecnologías que se centra en los robots y otros agentes artificiales inteligentes, y se divide en las dos indicadas ramas de ética de los robots y ética de las máquinas. No obstante ciñen la Roboética a las conductas morales de los humanos "…cuando diseñan, construyen, utilizan o interactúan con agentes inteligentes, así como del impacto que tienen los robots en la humanidad y en la sociedad". Mientras que la Ética de la máquina, o de las máquinas, se ocupa de la conducta moral de los agentes morales artificiales o AMAs: "A medida que la tecnología avanza y los robots se vuelven más inteligentes, los robots o los agentes inteligentes artificiales deberían comportarse moralmente y mostrar valores morales".

La primera en surgir es la ética del maquinista o «Roboética»: dan cuenta FLORIDI y SANDERS[14] de la fijación temprana por la ACM (*Association for Computing Machinery*) de unos *Principios* de carácter ético en el año 1992, que han sido actualizados periódicamente, y que están dirigidos a los constructores de robots, no a los propios robots: constituyen una ética no de la máquina, sino del maquinista. Señala en este sentido su Preámbulo que «El Código de Ética y Conducta Profesional de ACM ("el Código") expresa la conciencia de la profesión. El Código está diseñado para inspirar y guiar la conducta ética de todos los profesionales informáticos…». Por su parte COECKELBERGH[15] denomina *paciencia moral* a esta ética que debe presidir la relación del ser humano con el robot, y considera que parte de la atribución de un estatus moral a los robots, especialmente a los androides, similar al que otorgamos a nuestras mascotas. Mientras que Nathalie NEVEJANS[16] precisa que la ética para robots es hoy por hoy un mero postulado teórico, e insiste igualmente en que lo que tenemos es una «Roboética» como elaboración de normas éticas, pero para los constructores y utilizadores de robots, con el objeto de intentar que la utilización de robots no suponga un ataque a la integridad, privacidad y dignidad humanas. Ahora bien, la idea de que solamente existe una ética para los humanos que construyen, programan o utilizan robots o sistemas inteligentes, se basa en la predecibilidad y control de las conductas robóticas mediante la programación, e ignora la posibilidad de los llamados comportamientos emergentes del robot, actuaciones del

---

13 SIAU Keng y WANG Weiyu, "Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI", *Journal of Database Management* (JDM), vol. 31, Issue 2, abril-junio 2020, p. 76.

14 FLORIDI, L., SANDERS, J. «On the Morality of Artificial Agents». *Minds and Machines* 14, pp. 349-379, (2004). https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

15 COECKELBERGH, Ética de la Inteligencia artificial, cit., pág. 57.

16 NEVEJANS, Nathalie, *Traité de Droit et d'Éthique de la Robotique civile*, LEH Éditions, Burdeos, 2017, p. 708.

mismo que, aun moviéndose entre unos parámetros limitadores, el sistema goza de un margen de decisión bastante amplio, lo que pasamos a considerar.

4. ¿Existe una voluntad de la máquina?

La posibilidad de elección del sistema inteligente entre varias alternativas, determina una cierta imprevisibilidad de la conducta de la máquina, que se produce en mayor medida si el sistema de IA se ha desarrollado siguiendo un procedimiento de aprendizaje automático el *Machine* o *Deep learning*. Nos describe este procedimiento el *Dictamen* del Comité Económico y Social Europeo sobre la «*Inteligencia artificial: las consecuencias de la inteligencia artificial para el mercado único (digital), la producción, el consumo, el empleo y la sociedad*», de 31 de mayo y 1 de junio de 2017, (2017/C 288/01): *El aprendizaje automático incluye algoritmos capaces de enseñarse a sí mismos tareas específicas sin estar programados para ello. El método se basa en el procesamiento de «datos de entrenamiento» que sirven de base al algoritmo para aprender a reconocer patrones y formular normas. El aprendizaje profundo (Deep Learning o DL), una forma de Aprendizaje-máquina (Machine Learning), utiliza estructuras de redes neuronales basadas a grandes rasgos en el cerebro humano que aprenden mediante el ensayo y la respuesta. El resultado de estos avances es que los sistemas de IA (por medio de algoritmos) ya pueden aprender por sí mismos, y ser autónomos y adaptativos.*

Este aprendizaje descansa en la implementación en el sistema inteligente de neuronas artificiales dispuestas en capas hasta conformar una «red neuronal». Una neurona artificial es una función matemática que replica el funcionamiento de la neurona biológica. Como dice Latorre[17], "la idea central de esta primera neurona artificial es que todo lo que hace una neurona biológica es básicamente un procesamiento de información. Unas señales entran, se procesan y el resultado se pasa a otras neuronas. Dicho de forma más contundente: información entra en neurona artificial, se procesa, se genera nueva información, información pasa a otras neuronas. Todo se reduce a manipulación de información. Eso es lo que hace nuestro cerebro constantemente". Explica Muñoz Soro[18] que las neuronas artificiales agrupadas en capas, formando redes neuronales, admiten una programación que permite el entrenamiento del sistema «…mediante la introducción de casos previamente resueltos de forma tradicional. Con este entrenamiento la

[17] Latorre, Ética para máquinas, cit., pág, 105.

[18] Muñoz Soro, José Félix, Decisión jurídica y sistemas de información, Servicio de estudios del Colegio de Registradores, Madrid, 2003, pág. 191.

red va ajustando sus valores internos, en un auténtico proceso de aprendizaje, de modo que finalmente será capaz de resolver casos similares según los criterios que regían los supuestos de entrenamiento, aunque estos criterios nunca hayan sido explicitados». Esto implica el paso de los sistemas expertos, que son los dotados de una IA que combina mediante operadores lógicos, variables aleatorias y tratamientos estadísticos y probabilísticos, un enorme número de datos especializados sobre la materia, a los sistemas basados en estas nuevas herramientas que son las redes neuronales: acumulación en capas de *neuronas artificiales*.

Es decir, que mediante estas técnicas, la máquina toma sus propias decisiones: Si no cabe hablar de una voluntad propia, sí que se puede afirmar la existencia de un rango o espectro decisional para adoptar patrones de conducta que le lleven a encontrar soluciones propias, y no preprogramadas. Por supuesto que se puede objetar que estas decisiones vienen preconfiguradas por las reglas y bases establecidas en el algoritmo primario del sistema, pero también los seres humanos, tomamos decisiones en base a nuestras circunstancias personales preexistentes. Como señala Saiz García[19], el uso de las redes neuronales artificiales implica, «una nueva forma de computación inspirada en modelos biológicos con capacidad para aprender de la experiencia, seleccionar la información relevante a partir de los datos previamente implementados, establecer patrones y extraer conclusiones a partir de casos anteriores, pero, sobre todo, aprender a partir de los datos y perfeccionarse mediante retroalimentación».

Para entender este modo de actuación de los sistemas inteligentes, el mejor ejemplo es su aplicación para crear máquinas que jugaran al ajedrez. Un sistema experto jugador de ajedrez, como *Deep Blue*, partiría de la descripción del tablero y las reglas de movimiento de cada pieza, y añadiría en su base de datos desde las jugadas más básicas (el mate pastor, el mate del loco) a las de los grandes maestros. Un sistema que además utilice redes neuronales, como *AlphaZero*, se limita recibir las reglas de movimiento, y luego empieza a jugar contra sí mismo, una red neuronal contra otra (por eso se les llama redes adversarias) reteniendo las jugadas y movimientos ganadores y aumentando así su base de datos con las soluciones que le hayan llevado a la victoria: estrategias aprendidas y no programadas, que luego aplicará contra oponentes humanos o contra otras máquinas. Como se ve, las soluciones son hallazgos propios de la máquina y no del humano que la ha programado, o por lo menos no inmediatamente de éste. *Ya el*

[19] Saiz García, Concepción, "Las obras creadas por sistemas de inteligencia artificial y su protección por el derecho de autor", *InDret*, enero 2019, pág. 5.

ordenador ajedrecista Deep Blue consiguió derrotar al mejor jugador de ajedrez del mundo, Kasparov, el 1997. En la actualidad, los sistemas como AlphaZero superan ampliamente a los jugadores humanos. Dentro de estos parámetros de programación, puede decirse que existen decisiones propias de la máquina. Como dice LATORRE[20], naturalmente que las máquinas pueden decidir: "Un algoritmo que juega al ajedrez decide en cada jugada. Antes de procesar la información de una posición, muchas opciones son posibles. Tras cierta cantidad de análisis matemático, el algoritmo opta por la mejor opción. No nos engañemos, los humanos procedemos de la misma forma".

Otro ejemplo, más divertido, lo aportan los profesores de Stanford LEMLEY y CASEY[21]: En 2014 un grupo de robotistas se hallaban adiestrando a un dron de vuelo automático; el dron debía sobrevolar áreas circulares y posicionarse en el centro exacto del círculo. El caso es que, aunque se acercó en varios vuelos a dicho centro, al final el dron en lugar de volar hacia el objetivo, se salía fuera del área circular indicada. La razón, como se descubrió, es que los supervisadores, cuando el dron se salía del área señalada, para su entrenamiento lo colocaban directamente en el centro del círculo, por lo cual el dron llegó por sí mismo a la conclusión de que el camino más rápido para llegar al centro del círculo consistía en salirse del área delimitada. El problema de las conductas autoaprendidas por el robot es que no son controlables sino *a posteriori*, y que el proceso decisional es opaco. Como señalan los citados SIAU y WANG[22], el *Machine learning* es una herramienta brillante pero resulta difícil de entender el proceso que tiene lugar dentro de la máquina, por lo que se denomina "la caja negra" o *blackbox*: "La caja negra hace que los algoritmos sean misteriosos incluso para sus creadores… dado que la caja negra no es interpretable por los humanos, la IA puede evolucionar sin monitorización ni guía humanas"[23].

¿Puede programarse un "código moral" que regule un comportamiento moral autodecidido por la máquina? Esto es lo que constituiría la auténtica Ética de la máquina o del robot, pues la ética para maquinistas no es sino un código moral dirigido a los humanos que interactúan con mecanismos dotados de IA. Mediante la ética de la máquina se intenta que las máquinas

---

20 LATORRE, Ética para máquinas, cit., pág, 105.

21 LEMLEY, Mark A. y CASEY, Bryan, «*Remedies for Robots*», *86 University of Chicago Law Review (2019), Stanford Law and Economics Olin Working Paper* No. 523, DOI: http://dx.doi.org/10.2139/ssrn.3223621.

22 SIAU Keng y WANG Weiyu, "Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI", …*cit.*, p. 80.

23 Y los autores ponen de ejemplo el caso del sistema de IA que tuvo que cerrar Facebook en 2017 porque encontraron que había creado su propio lenguaje y que era completamente incomprensible para los humanos.

adopten comportamientos que, no estando preprogramados, sean *buenos*. Una ética de la IA que sea completa debe gestionar una conducta moral de la máquina que tenga auténtico significado moral. Un ejemplo de problema moral que aporta WALLACH[24] nos puede mostrar esta necesidad, así como que el comportamiento de un sistema inteligente puede tener peso y repercusiones morales: "Hace años el fabricante de una muñeca consideró cómo debía responder ésta si era maltratada por un niño. Los ingenieros sabían cómo montar sensores dentro de la muñeca que alertarían al sistema de la existencia de maltrato. Después de analizar la cuestión con abogados, se decidió que la muñeca no haría ni diría nada". Y trasladando esto a los robots-nanny, se pregunta el autor cómo debería responder un robot a un niño que se comporta con él de forma inapropiada o incluso violenta. Por un lado, está la educación del niño, y si el robot nanny no dice nada, permite la mala educación y favorece el comportamiento destructivo del niño. Pero si dice algo como: "¡Para, me haces daño!", al ser un hecho que el robot no siente dolor, le estaría mintiendo, o peor, le enseñaría a mentir.

5. La conformación de una conciencia sintética

La posibilidad de comportamientos emergentes y decisiones propias del robot nos llevan a una posibilidad más interesante que el manido tópico de la "personalidad electrónica": nos conducen a la pregunta de si es posible una conciencia robótica o cibernética, es decir, un sistema de normas morales en forma de algoritmo que regule y controle, partiendo de los parámetros morales de "bueno" y "malo", el comportamiento decidido por sí mismo por un ente no humano. Es en este sentido en el que los profesores (de Yale e Indiana, respectivamente) Wendell WALLACH y Colin ALLEN[25] nos hablan de los sistemas inteligentes como «agentes morales» (AMA, Agente Moral Artificial o *Artificial Moral Agent*), puesto que tienen inteligencia y "hacen cosas", de ahí que sean agentes", pero además monitorizan y regulan su conducta desde un punto de vista moral, teniendo en cuenta los daños que su conducta (o inactividad) pueda causar.

Ahora bien, existen detractores de esta interpretación, porque implica otorgar una personalidad al sistema inteligente. Así, en contra de esta

[24] WALLACH, Wendell, *A dangerous master. How to keep Technology from slipping beyond our control*, Basic Books, Nueva York, 2015, pág. 224.
[25] WALLACH, Wendell y ALLEN, Colin, *Moral Machines. Teaching robots right from wrong*, Oxford University Press, Nueva York, 2009, pág. 16.

creación de agentes morales, VAN WYNSBERGHE y ROBBINS[26], piensan que si es necesario dotar de ética a los sistemas de IA capaces de causar daño, dado que por la conectividad de sistemas que configura la Internet de las cosas (*IoT*) cualquier electrodoméstico podría causar daño por abrir una brecha en la intimidad del hogar, habría que dotar de ética hasta al frigorífico, posición ésta que no tiene mucho sentido. También distinguen los autores que una cosa es estar situado en un contexto moral y otra jugar un papel moral efectivo, poniendo el ejemplo del perro de una residencia de ancianos, que desempeña un papel terapéutico, y nadie se ha preocupado de si su conducta es o no moral[27]. En este sentido CAPPUCCIO, PEETERS y MCDONALD[28] propugnan un tratamiento jurídico de los robots sociales (se refieren a los androides) que les haga superiores a meras cosas, pero sin atribuirles verdadera personalidad. Proponen para los robots lo que denominan un «Reconocimiento moral basado en la virtud» (*Moral consideration for Robots, MCR*), justificado no en la atribución de derechos al robot, sino en la consideración de su actividad relacional con seres humanos, pero no porque tengan una naturaleza que deba respetarse, sino porque la utilización inmoral de entes robóticos daña a la moral humana, lo que es incompatible con lo que denominan Ética de la virtud. Es decir, el maltrato a los robots es inmoral, en primer lugar porque revela la presencia de un defecto moral en la conducta del abusador, y en segundo lugar porque invita a la inmoralidad a uno mismo y a otros, «pues ofrece un ejemplo despreciable que otras personas podrían imitar».

Sin embargo, para los partidarios de crear AMA, sin entrar en atribuciones de personalidad, la cuestión se concreta en implementar en los sistemas de IA algoritmos de decisión que les lleven a comportamientos y decisiones que, conforme a la conciencia humana, podamos calificar de *morales*. Estos algoritmos deberán aplicar, para llegar a dichos resultados, un número delimitado de valores éticos, con un preciso y jerárquico orden de preferencia, sin que el sistema pueda alterar ni el número de valores señalado ni la preferencia que para los mismos se le haya impuesto. Entre nosotros, LATORRE[29] estima con James Moor que es necesario distinguir entre *agentes normativos* (que meramente siguen órdenes), *agentes de impacto ético*, *agentes implícitamente éticos* (cuyos programadores han seguido principios

[26] VAN WYNSBERGHE y ROBBINS, «Critiquing the Reasons for Making Artificial Moral Agents», *cit.*, pág. 724.

[27] Pero claro, el perro no puede aumentar una dosis de medicación, y el robot-enfermero, sí.

[28] CAPPUCCIO, M.L., PEETERS, A. & MCDONALD, W. «Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition». *Philosophy & Technology,* 33, (2020). https://doi.org/10.1007/s13347-019-0341-y

[29] LATORRE SENTÍS, Ética para máquinas, cit., pág, 187.

éticos en su construcción, y *agentes explícitamente éticos*, que son capaces de tomar decisiones éticas propias.

En realidad, creo que la autonomía decisional bajo parámetros o límites éticos no presupone que el AMA tenga personalidad ni conciencia: Conciencia la tenemos los humanos, y lo más lejos que el AMA puede llegar es a comportarse "como si" tuviera conciencia, como resultaría del Test de Turing. Dicho comportamiento puede ser calificado de bueno o malo, y por lo tanto la decisión que el AMA tome tendrá valor ético, aunque el sistema carezca de conciencia.

Si bien para muchos estamos ante una simple cuestión técnica, la realidad es que los problemas morales son inevitables en IA, como se está comprobando hoy en día con los vehículos autónomos. En este sentido, MILLAR[30] nos dice que es necesario implementar en sistemas de IA, en particular en los vehículos autopilotados, un «entorno ético para la gestión de colisiones». E inspirándose en LIN, considera que este entorno ético puede interpretarse razonablemente para los vehículos autónomos, tanto legal como éticamente, como un algoritmo de orientación hacia objetivos, con el problema añadido de la consiguiente responsabilidad del fabricante-programador, o bien del usuario, según corresponda a uno o a otro la fijación de tales objetivos: esta posibilidad se conoce como «ética variable», en el sentido de configurable.

También sobre el ejemplo del vehículo autónomo, nos dice GOODALL[31] que «El uso de ejemplos hipotéticos puede sugerir que la ética solo es necesaria en circunstancias increíblemente raras. Sin embargo, un informe reciente del equipo de vehículos autónomos de Google sugiere que la ética ya se está considerando para evitar los desastres: ¿Qué pasa si un gato irrumpe en la carretera? ¿Un ciervo? ¿Un niño? …Se necesitan decisiones éticas siempre que exista un riesgo, y el riesgo siempre está presente al conducir». Señala este autor en otro estudio[32] cómo la implementación ética de sistemas inteligentes es ya una necesidad en los vehículos automatizados, por la sencilla razón de que están ya circulando y que, aunque su porcentaje de accidentes es mucho más bajo, la posibilidad del accidente sigue existiendo. Añade que

30 MILLAR, «Ethics Settings for Autonomous Vehicles», en *Robot Ethics 2.0. cit.*, pág. 27.

31 GOODALL, Noah J., "Machine Ethics and Automated Vehicles", Preprint version. Published in Meyer and Beiker (eds.), *Road Vehicle Automation*, Springer, 2014, pp. 93-102. http://dx.doi.org/10.1007/978-3-319-05990-7_9

32 GOODALL, Noah J., "Ethical Decision Making During Automated Vehicle Crashes", Preprint version. Final version in *Transportation Research Record: Journal of the Transportation Research Board*, No. 2424, Transportation Research Board of the National Academies, Washington,D.C., 2014, pp. 58–65.https://doi.org/10.3141/2424-07

tomar decisiones lógicas de evitación del choque no es un problema si el accidente es evitable. Sin embargo, si no se puede evitar una lesión personal, el vehículo automatizado debe decidir cuál es la mejor forma de chocar, lo que se convierte en una decisión moral, como se demuestra con el siguiente ejemplo: «Un vehículo automatizado viaja sobre un puente de dos carriles cuando un autobús que viaja en la dirección opuesta invade repentinamente su carril… Hay tres alternativas: A. Giro a la izquierda y caer del puente, lo que garantiza un accidente grave del vehículo. B. Choque de frente contra el autobús, resultando en un choque moderado de los dos vehículos. C. Intentar pasar el autobús por la derecha. Si el autobús vuelve repentinamente hacia su propio carril, un evento de baja probabilidad dado lo lejos que se ha desviado el autobús, se evita el choque. Si el autobús no vuelve (un evento de alta probabilidad), se produce un importante choque de dos vehículos. Este choque sería una colisión lateral, que conlleva un mayor riesgo de lesiones que la colisión frontal total en la alternativa B».

El algoritmo de planificación de la ruta del vehículo automatizado tendría que determinar rápidamente el rango de resultados posibles para cada ruta considerada, la probabilidad de que ocurran dichos resultados, y sobre esa base probabilística, tomar la decisión menos dañina. Naturalmente, la relativa sencillez del planteamiento es sólo aparente, y en realidad en el problema intervienen otras muchas variables: Imagínese que el autobús invasor está vacío, o al contrario, que está lleno de niños, o imaginemos que chocan dos coches, uno conducido por un anciano y otro por una embarazada… las posibilidades son infinitas y el problema se plantea en su correcta dificultad cuando intentamos medir el valor de las vidas humanas para comparar cuál tiene más peso (y cuál menos, claro). Para superar estos inconvenientes, el autor propone, partiendo de un conjunto de reglas de racionalidad ética preprogramadas en el sistema inteligente (por ejemplo, la de escoger la trayectoria en la que se causen menos daños a las personas), recurrir al aprendizaje-máquina, de manera que sea el propio sistema, dotado de redes neuronales, quien se ejercite en la búsqueda de soluciones, debidamente tutelado o «entrenado» por seres humanos que aprobarán o desaprobarán dichas elecciones de la máquina. El vehículo autónomo demuestra que la ética de la IA se abre paso en nuestra vida diaria, y aparece directamente *incrustada* en el sistema inteligente.

6. Educando a Eliza Doolittle

En la conocida película *My Fair Lady*, Eliza Doolittle es una florista callejera del East End londinense a la que el profesor de fonética Henry Higgins da

clases de dicción para convertirla en una dama de la alta sociedad en seis meses, y ganar así una apuesta con el coronel Pickering. Inspirándose en el personaje, el profesor del MIT Joseph Weizenbaum llamó ELIZA a un programa conversacional que también había «aprendido a hablar», y que simulaba ser un siquiatra, adquiriendo gran fama en 1966. La diferencia entre la Eliza Doolittle humana y la ELIZA cibernética podemos verla en el catastrófico experimento que tuvo lugar 50 años después, el de la asistente de conversación TAY. El *chatbot* conversacional TAY fue diseñado por Microsoft en 2016 para entablar conversación mediante Twitter, estando dotado de una gran empatía y comprensión hacia sus contertulios twiteros (TAY era el acrónimo de *Thinking About You*, "pensando en ti"): En suma, era un *bot* con "buen rollo", que no discriminaba a nadie y enseguida se hacía colega del interlocutor. Bien, esto era en realidad una vulnerabilidad, como pronto descubrieron muchos twiteros que bombardearon a la pobre TAY (se diseñó para que simulase ser una chica joven) con todo tipo de mensajes sexistas, racistas y misóginos, ¡y funcionó! Al poco rato TAY colgaba en su cuenta de Twitter mensajes que decían: "Hitler tenía razón, odio a los judíos", o "Odio a las feministas, deberían morirse todas y arder en el infierno". A las 24 horas, TAY fue retirada de la red social[33].

Es decir, la Eliza humana cuenta con una estupenda conciencia, que es el algoritmo para tomar decisiones morales que cada cual tenemos implementado en nuestro cerebro desde niños. En cambio, la máquina carece de toda conciencia, y la única idea del «yo» que tiene es la de diferenciar los datos que ha elaborado ella misma de los que ha adquirido de fuentes exteriores. De esto ya se dio cuenta el propio Weizenbaum, al señalar: «… hay diferencias importantes entre los hombres y las máquinas como pensadores. Yo sostengo que, por muy inteligentes que sean las que puedan ser inteligentes, hay algunos actos de pensamiento que deberían ser intentados sólo por los seres humanos»[34]. Frente a esta posición ¿Podríamos programar en un sistema experto un algoritmo parecido a la conciencia humana? Entiendo que no, porque como considero en otro estudio, tampoco sabemos qué es exactamente nuestra conciencia, el *yo mismo* que nos dirige a cada uno del que habla Damasio, y si no sabemos exactamente qué es, tampoco podemos crear un algoritmo que la reproduzca. Otra cosa es que intentemos diseñar

[33] Web CBS NEWS: "Microsoft shuts down AI chatbot after it turned into a Nazi". First published on March 24, 2016 / 2:30 PM https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/

[34] Weizenbaum, Joseph, *Computer Power and Human Reason. From Judgment to Calculation,* W.H. Freeman and Company, San Francisco: 1976, p. 13. Consulta marzo 2021 en: http://blogs.evergreen.edu/cpat/files/2013/05/Computer-Power-and-Human-Reason.pdf

un robot que se comporte «como si tuviera conciencia», y esto lo haremos, como dice GOODALL, del mismo modo que se educa a los niños: mediante un adiestramiento conductual, reprimiendo los comportamientos que consideramos malos o inmorales y premiando los buenos o morales. Es decir, mediante adiestradores humanos que depuren las conductas decididas por la propia máquina a partir de los resultados de su red neuronal de aprendizaje.

Esta idea de hacer «como si» la máquina tuviera conciencia es además plenamente coincidente con quienes, desde posiciones cognitivistas, consideran que la conciencia o el alma humana no es en realidad, como entendió HUME, sino un batiburrillo de experiencias, ideas y recuerdos. En definitiva, una especie de subproducto del funcionamiento cerebral, cuya existencia se comprueba a partir de la conducta del sujeto. Es decir, que también los humanos nos comportamos "como si" tuviéramos conciencia, porque en realidad lo único que cada cual puede afirmar es que "siente que" tiene conciencia. Esta consideración puramente conductual de la conciencia es la que sigue la Neuroética, como parte de la Ética que se ocupa de las bases cerebrales de la conducta moral. Como nos dice CORTINA ORTS[35], la conducta moral sería un mero mecanismo adaptativo adquirido para la supervivencia de la especie: "Las normas morales no serían entonces sino normas adaptativas y la tarea ética consistiría en intentar descubrir qué normas favorecen la supervivencia". La supervivencia de la especie humana, naturalmente.

Volviendo a la programación de una conciencia robótica, algún autor, como Thomas METZINGER[36], va más allá y no sólo consideran que se puede replicar la inteligencia humana, sino también la conciencia y sensibilidad humanas, lo que en consecuencia permitiría que la máquina tuviese identidad y pudiese sentir dolor. Precisamente para evitar que se inflija sufrimiento a un sistema de IA, propone este autor que se implemente una «Carta Europea para la IA», en la que se prohíba esta creación de una conciencia con capacidad de sufrimiento o «fenomenología sintética». Esta denominada *fenomenología sintética* se refiere a la posibilidad de crear no solo inteligencia general, sino también conciencia o experiencias subjetivas en sistemas artificiales avanzados, posibilidad que METZINGER estima muy real.

Sin ir tan lejos, para lograr una educación moral de la máquina, y que se comporte "como si tuviera conciencia", los procedimientos son los que ya

---

35 CORTINA ORTS, Adela, *Neuroética y Neuropolítica*, Tecnos, Madrid, 2011, pág. 72.

36 METZINGER, Thomas, en *Should we fear artificial intelligence?* European Parliamentary Research Service, STOA-Science and Technology Options Assessment, March 2018, pág. 29. PE 614.547, DOI: 10.2861/412165.

conocemos: o se programa dicho código moral como una serie de instrucciones insertas en el algoritmo general de la máquina (sistema experto), o se logra que la máquina "aprenda" a tomar decisiones morales "buenas" y evite las "malas", es decir, en ambos casos, que adquiera comportamientos moralmente adecuados.

El primer procedimiento plantea el problema de decidir el orden de valores que debe primar en el algoritmo decisional de orden moral, cuestión que los humanos no hemos resuelto del todo y que por tanto sólo podemos programar en un robot con muchas dificultades. Pero es necesario hacerlo, pues la idea de un código moral «incrustado» (*embedded*) en el robot o máquina dotada de IA es una premisa para la construcción de sistemas que puedan tomar decisiones autónomas y por ello, causar daños. Señala así Jason Millar[37] que la «configuración ética» es una de las premisas del diseño de automóviles autónomos, puesto que los parámetros de decisión necesariamente deben implementarse en la programación de la conducción. La ética debe estar «incrustada» en el diseño del automóvil, y para valorar esta afirmación nos plantea el autor el «Dilema del casco», que es una buena alternativa al archiconocido «Dilema del tranviario»: Un vehículo autónomo pierde el control y para evitar un grave accidente debe efectuar un brusco viraje, que inevitablemente le lleva a impactar o con un motorista que lleva puesto el casco o con uno que no lo lleva. Si se le programa para que dé preferencia al valor «minimizar el daño», impactará con el motorista con casco, pues tiene más posibilidades de sobrevivir al atropello; si se le programa para que prefiera el valor «conducta responsable», impactará contra el que no lleva casco.

El segundo procedimiento consiste en que la solución *moral* la haya obtenido el propio sistema inteligente mediante aprendizaje autónomo (*Deep learning*), lo que tampoco es una solución perfecta, pues su solución puede no coincidir con nuestro orden de valores, o puede considerarse moralmente rechazable *a posteriori* según sus consecuencias últimas. Y, en cualquier caso, si hay que elegir entre dos daños, la opinión sobre la moralidad de la solución adoptada no coincidirá con la opinión de la víctima, o de sus familiares. El problema de hacer primar unos valores sobre otros se nos plantea igualmente cuando tengamos que "adiestrar" *a posteriori* a nuestra máquina, eliminando los comportamientos nocivos o que lesionen valores reconocidos. Este problema se comprueba en toda su extensión al examinar

[37] Millar, Jason, «Ethics Settings for Autonomous Vehicles», en *Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence,* Edited by Patrick Lin, Ryan Jenkins & Keith Abney, Oxford University Press, New York, 2017, pág. 21.

los sistemas de IA para los automóviles autodirigidos, como decía GOODALL: partiendo de una base de reglas racionales de moralidad, la consecución de una conducta moralmente adecuada precisará que la propia máquina aporte sus soluciones, fruto de su propio sistema de aprendizaje, y a posteriori éstas sean aprobadas o rechazadas por evaluadores (o entrenadores) humanos.

El problema de hacer que una máquina actúe «como si tuviera conciencia» no estriba en implantarle unas reglas que deba cumplir y unos parámetros de conducta que deba seguir: esto es simplemente una programación lineal. Para demostrar que se tiene conciencia, o al menos fingirlo, es preciso que en determinadas situaciones de conflicto ético se proceda a sopesar valores y a elegir entre dos conductas, las dos posibles y lícitas, la que sea más conforme al valor o principio que merezca más respeto en el código moral del sujeto. ¿Es posible programar a la máquina para que actúe "como si tuviera conciencia? No lo sé, pero estimo que el problema está en lograr una programación que atribuya un peso específico a cada uno de los valores en los que se quiere fundamentar el código ético de la máquina, un valor numérico a cada uno de estos valores morales (dejar esta tarea a la propia máquina no me parece prudente). Es asimismo necesario un algoritmo que escanee las situaciones reales y reconozca dos cosas: la presencia de valores y principios morales, y la existencia de un conflicto entre al menos dos de ellos. Finalmente, la máquina debe comparar los valores en conflicto y, dependiendo del peso específicamente asignado a cada uno, hacer prevalecer en su decisión el principio de mayor valor. Me explicaré con un ejemplo: Imaginen que estoy empujando el carrito en el que llora mi hijo de un año, porque es tarde y tiene sueño y hambre. Para llegar pronto a casa tengo que cruzar la calle y el semáforo está en rojo. Miro a derecha e izquierda y veo el camino despejado, trescientos metros en ambas direcciones. Yo, y cualquier persona sensata, cruzaría la calle sin problemas haciendo prevalecer el valor del cuidado del niño sobre el valor del cumplimiento de una norma de tráfico que puede ignorarse sin peligro para nadie. ¿Podría hacer esto un robot? Un robot tendría primero que tener asignados unos valores en su memoria, en concreto el valor del cumplimiento de las normas de tráfico como cuestión de orden público, y también el valor de la protección a la infancia de todo sufrimiento innecesario. También tendría que reconocer la situación del cruce de la calle como conflictiva, pues choca el cumplimiento de la norma con la necesidad de llegar pronto a casa para dar de cenar al bebé y ponerlo a dormir en su cuna. Finalmente, debe adoptar la conducta cuyo principio de referencia tenga un valor numérico superior al otro, lo que no es nada sencillo, porque el valor del cumplimiento de la norma de tráfico será distinto si no viene ningún coche a si viene uno a gran velocidad. Estos procesos

computacionales jugarían un papel análogo al de la conciencia humana, al menos para un observador externo, por lo que se podrían calificar de *conciencia sintética*.

7. Una conciencia implantada como válvula de seguridad

La utilidad de la Ética aplicada a la IA, y la necesidad por tanto de una "Ética cibernética o robótica", o *Roboethics,* está en limitar el campo decisional de un sistema inteligente para minimizar el daño. Esta ética no busca la supervivencia o la mejora moral y vital del robot, sino la seguridad y felicidad de los seres humanos. Consiste en una programación algorítmica del sistema que logre hacer prevalecer los valores y principios situados en lo alto de una tabla sobre los situados más abajo, especialmente en situaciones de conflicto. El respeto de los valores éticos superiores es una exigencia ética tanto para el maquinista como incrustada en la máquina misma. Lo que ocurre es que el comportamiento del sistema al tomar decisiones de peso moral nos da la impresión de que cuenta con una conciencia, puesto que actúa "como si" tuviera una conciencia. Para designar a estos sistemas inteligentes se utiliza el ya expuesto concepto de AMA, Agente Moral Artificial o *Artificial Moral Agent*, puesto que sus decisiones tienen significado moral (para los humanos, claro).

Muchos autores estiman que la IA es tan potente que, a futuro, no se excluye la emulación en sujetos artificiales de una conciencia semejante a la humana. Partiendo del Test de Turing para considerar si el computador se hallaba o no provisto de inteligencia, se considera que si la conducta y las decisiones de la máquina parecen conscientes, y no se pueden diferenciar de las que tomaría un humano consciente, habrá que entender que tiene conciencia. Se habla así de un «Test de Turing *moral*».

¿Esta "conciencia sintética", nos obligaría a considerar a su poseedor como un ser también dotado de estados mentales, sentimientos y emociones, como una "persona"? Para el profesor Fabio Fossa[38], si se sigue el «Test de Turing Moral» se afirmará la realidad de la agencia moral, siguiendo el llamado «Enfoque de Continuidad» (*Continuity Approach*, o *CA*). Conforme a este enfoque, los AMA, al igual que los seres humanos, son agentes morales, si bien cuantitativamente de menor entidad, y reproducen la ética humana. Pero desde otro enfoque, el llamado «Enfoque de Discontinuidad» (*Discon-*

[38] Fossa, Fabio, «Artificial moral agents: moral mentors or sensible tools?», *Ethics and Information Technology* (2018) 20:115-126. https://doi.org/10.1007/s10676-018-9451-y, Published online: 16 March 2018, pág. 116.

tinuity Approach, DA) se niega esta potencia moral en los seres artificiales, que han sido fabricados por los seres humanos, siendo por tanto «productos» para cumplir determinados fines, y por tanto meras herramientas. Su inteligencia y eficacia hace que sean unas valiosas «herramientas sensibles», pero esto no les permite fijar los fines y los valores que rigen las vidas humanas. Fossa finalmente se decanta por el de la Discontinuidad, pues el contrario solamente se basa en las similitudes, y no en las diferencias patentes entre seres humanos y máquinas, y termina diciendo: «Ningún logro tecnológico resolverá la cuestión relativa a fines últimos de la vida humana y lo que deba considerarse como bien o bueno. Cuando trabajan juntas, la Filosofía moral y la Ética de la máquina pueden evitar simplificaciones ilegítimas y mezclas abstractas sólo si esta diferencia es constantemente mantenida».

Por mi parte, estimo que hay que volver al origen de esta *Roboética*, y recordar que surge para resolver problemas morales con la finalidad de evitar que se lesionen valores dignos de respeto y protección. Es decir, que lo que se pretende es que el sistema inteligente no cause daños. Mientras la ética del maquinista pretende moralizar a los humanos que llevan a cabo la creación y utilización de sistemas inteligentes, la ética de la máquina o «conciencia robótica», es un instrumento de control o medida de seguridad instalada en el sistema.

Como señalan Floridi y Sanders[39], cabe en este sentido una aproximación al concepto de Ética de la máquina en la que la moral puede ser considerada como un «umbral» o límite a la conducta del agente moral: «Un agente es moralmente bueno si todas sus acciones respetan ese umbral; y es moralmente malvado si alguna acción lo viola». Por tanto, los estados intencionales son una condición «agradable pero innecesaria para la aparición de la agencia moral». Ante todo, porque no tenemos ninguna posibilidad de acceder a dichos estados mentales o intencionales. A partir de esto, los requisitos de intencionalidad y libertad en la adopción de la conducta, sobran, y la ética o la moralidad del Agente Artificial es simplemente un límite, que además se puede expresar mediante una función que concrete en un determinado valor el límite o umbral a partir del cual se considere «buena» la conducta del agente. También con esta idea opinan Wallach y Allen[40] que los «agentes morales» monitorizan y regulan su conducta a la luz de los daños que su conducta pueda causar o las tareas que pueda incumplir, y los seres humanos no deberían esperar nada más ni menos de un AMA. Por tanto, lo que venimos llamando "conciencia sintética" no es sino una sofisticada

[39] Floridi, L., Sanders, J. «On the Morality of Artificial Agents», *cit.*, págs.. 13 y ss.
[40] Wallach y Allen, *Moral Machines. Teaching robots right from wrong*, cit., pág.16.

medida de control o seguridad. Un «buen» agente moral, es decir, uno que funcione correctamente, nos dicen estos autores, será el que pueda detectar la posibilidad de daño o incumplimiento de sus deberes, y dar pasos para evitar o minimizar estos resultados indeseables[41].

8. Un robot como debe ser

Es decir, estamos intentando conseguir comportamientos "buenos" de los sistemas inteligentes para que no causen daños, por lo que aquí la bondad consiste en realidad en la seguridad. Pero es que la propia moral humana surge con la finalidad de la seguridad y preservación de la especie. Como ya hemos visto que nos dice Cortina Orts, la conducta moral sería un mecanismo humano adaptativo adquirido para la supervivencia de la especie. Si nos interesa construir robots dotados de ética no es para que se nos parezcan, es para que sean seguros. Nos interesa construir robots que se comporten "como si" tuvieran conciencia, pero no la de un delincuente, sino la de una buena persona. Que realmente lleguen a tenerla parece difícil, como señala Bryson[42]: «Que la conducta moral se sustenta en unas bases cerebrales significa que un ser que careciera de un cerebro humano, ligado obviamente a un cuerpo humano, sería incapaz de sentirse obligado por normas morales, de captar valores a los que denominamos morales, de tener emociones y sentimientos de carácter moral, de desarrollar virtudes… Contar con un cerebro humano es condición necesaria para realizar esas cuatro tareas que componen las dimensiones del mundo moral».

La idea de la ética como sinónimo de seguridad la vemos también en los citados Wallach y Allen, que dedican un capítulo de su obra a la «ingeniería (o diseño) de la moralidad», advirtiendo que el primer mandato del Código Ético de la Sociedad Profesional Nacional de Ingenieros consiste en que deben «mantener la primacía de la seguridad, salud y bienestar del público». En otro estudio posterior, insiste Wallach[43] en que esta programación de la máquina para que actúe *bien* es un tipo de conducta que denomina *moralidad operativa*, e insiste en el problema de conseguir que un sistema

41 De hecho, en Inglaterra se publica la norma Guide to the ethical design of robots and robotic systems, BS 8611:2016, en abril 2016, por el instituto de normalización British Standard. Recoge hasta veinte riesgos e inseguridades, agrupados en cuatro categorías: Sociales, de ejecución, comerciales y financieros, y medioambientales. En Bryson, J. J., y Winfield, A., "Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems", Computer, vol. 50, nº 5, https://doi.org/10.1109/MC.2017.154

42 Cortina Orts, Adela, *Neuroética y Neuropolítica*, Tecnos, Madrid, 2011, pág. 72.

43 Wallach, *A dangerous master. How to keep Technology from slipping … cit.*, pág. 241.

sea *funcionalmente moral* cuando adopta sus propias decisiones. Esto es sin embargo un punto de crítica para VAN WYNSBERGHE y ROBBINS[44], que estiman que las cuestiones de seguridad no se pueden confundir con los debates morales, y que si lo que se quiere son máquinas seguras, toda esta discusión versa sobre seguridad, no sobre ética.

En realidad, los seres humanos siempre hemos manejado categorías morales para referirnos a la seguridad o el peligro que puedan tener entes no humanos. Si el perro de mi vecino me muerde en los tobillos, yo no diré que es un animal que no es seguro para sacar a pasear, diré que es un bicho malísimo. Todavía con más razón aplicaré las categorías de bueno y malo para un ser que manifiesta inteligencia, porque hasta hoy la conducta inteligente iba ligada a la humanidad y conciencia del ser que la manifestase. Aclaran los citados WALLACH y ALLEN que la valoración de la seguridad y control ha presidido siempre el diseño de las máquinas simples, generando lo que denominan una «moralidad operativa», así por ejemplo las medidas de seguridad incluidas en las armas de fuego (un ejemplo más próximo, las medidas de seguridad de un simple mechero de gas para que un niño pequeño no pueda encenderlo). Las máquinas inteligentes, capaces de comportamientos autónomos, precisan un diseño más evolucionado para lograr que alcancen una «moralidad funcional», en el sentido que puedan manifestar comportamientos con significado moral. Pero, aun así, los resultados morales que se obtienen son de bajo nivel, y para obtener resultados de moralidad más perfecta se hace preciso diseñar sistemas con «actividad moral completa» (*full moral agency*). Ponen los autores como ejemplo de un sistema moral de bajo nivel el sistema *MedEthEx* de apoyo y consejo médico, que no toma decisiones autónomas, sólo propone alternativas. Otro ejemplo lo tendríamos en la Sección 17941 del Código Mercantil y Profesional de California, que obliga los bots que interactúen *on line* a declarar su condición de máquinas.

Por tanto esta nueva «Ética para máquinas» de que habla LATORRE, la ética de la IA, puede considerarse como uno de los mecanismos de control que se añaden para evitar, o al menos mitigar, la peligrosidad de las nuevas realizaciones técnicas. En este sentido, autores como CERVANTES, LÓPEZ, RODRÍGUEZ, CERVANTES, CERVANTES y RAMOS[45] resumen las principales posiciones doctrinales norteamericanas sobre el control conductual de los sistemas

[44] VAN WYNSBERGHE y ROBBINS, «Critiquing the Reasons for Making Artificial Moral Agents», *cit.*, pág. 726.

[45] CERVANTES, LÓPEZ, RODRÍGUEZ, *et al.*, «Artificial Moral Agents: A Survey of the Current Status», *Science and Engineering Ethics* (2019), pág. 3, https://doi.org/10.1007/s11948-019-00151.

robóticos, indicando que este control se intenta mediante normas éticas: ética de las máquinas, moral mecánica, moral artificial, moral computacional, roboética, e IA amigable, que se inspiran en los conceptos de ética y moral humanas pero que no son sino una técnica de control de comportamientos potencialmente dañinos.

De acuerdo con lo que hemos visto, el comportamiento moral es el comportamiento seguro, y esto parte de un miedo soterrado a las innovaciones técnicas disruptivas que forma parte del imaginario colectivo europeo. Como señala la revista *Política Exterior*[46], «En Europa, la revolución de la IA se percibe a menudo como una ola llegada desde otros océanos que amenaza el modelo socioeconómico europeo y de la que hay que protegerse. En comparación con sus homólogos estadounidenses o chinos, que suelen fomentar el "optimismo tecnológico", los responsables políticos europeos intentan ante todo "regular" la revolución de la IA para minimizar los riesgos». Una advertencia final que hace Bryson[47] es que al ser la tecnología de IA completamente dependiente del diseño, y por tanto obediente a los parámetros recibidos, hay precaverse de que estas tecnologías sean controladas por poderes poco controlables, como grandes corporaciones o redes clandestinas.

¿Y no sería posible, puesto que tememos que la IA cause estos daños en su actuación diaria, prescindir de tan peligrosa tecnología? La respuesta es, por supuesto, que no, porque es una tecnología imprescindible, y quien no la use se quedará en una sociedad distinta, no digo que peor, pero sí más primitiva. Y más pobre, recuerden las frías cifras del estudio *European framework on ethical aspects of artificial intelligence,* de septiembre 2020: un enfoque común de la UE sobre los aspectos éticos de la IA puede generar hasta 294.900 millones de euros en PIB adicional y 4.6 millones de puestos de trabajo adicionales.

---

[46] Miailhe, Nicolas, Hodes, Cyrus, Çetin, Buse, Lannquist, Yolanda, Jeanmaire, Caroline, "Geopolítica de la inteligencia artificial", *Política exterior*, ISSN 0213-6856, Vol. 34, Nº 193, 2020, págs. 56-69.

[47] Bryson, Joanna J., «The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation», *The Oxford Handbook of Ethics of AI*, Edited by Markus D. Dubber, Frank Pasquale and Sunit Das, Print Publication Jul 2020, DOI: 10.1093/oxfordhb/9780190067397.013.1