ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Por la Académica de Número
Excma. Sra. Doña Adela Cortina Orts[*]

1. LA ANÉCDOTA COMO PUNTO DE PARTIDA: MICHIHITO MATSUDA

Esta intervención se inscribe en el marco de esa revolución 4.0, en la que vivimos y somos, y de la que se han ocupado con tanto acierto en esta Academia Juan Manuel González Páramo y Juan Miguel Villar Mir[1]. Por mi parte, me referiré muy especialmente a algunas de las cuestiones éticas (ventajas y problemas) que pone sobre el tapete la llamada "inteligencia artificial".

En sus trabajos de filosofía de la historia decía Kant que la invención del puñal precedió a la conciencia del imperativo categórico ("no matarás"), es decir, que los avances técnicos se anticiparon a las orientaciones morales sobre cómo hacer uso de ellos. Y sigue siendo cierto que el exponencial progreso de lo que hoy llamamos ya "tecnociencias" plantea una gran cantidad de preguntas éticas para las que es necesario ir encontrando respuestas. Precisamente porque lo moral no consiste en mapas de carreteras, ya cerrados, sino en una brújula que señala el norte, es posible y necesario encontrar mejores caminos ante los nuevos descubrimientos tecnocientíficos. De intentar encontrar *orientaciones éticas* acerca de cómo habérnoslas con la inteligencia artificial tratará esta intervención, que empezará con una anécdota para transitar después a la categoría o a las categorías.

---

[*] Sesión del día 7 de mayo de 2019.

[1] Este estudio se inserta en el Proyecto de Investigación Científica y Desarrollo Tecnológico FFI2016-76753-C2-1-P, financiado por el Ministerio de Economía y Competitividad (ahora Ministerio de Ciencia, Innovación y Universidades), y en las actividades del grupo de investigación de excelencia PROMETEO/2018/121 de la Generalidad Valenciana.

En abril de 2018 se celebraron elecciones municipales en un distrito de Tokio con más de 150.000 habitantes, llamado Tama New Town. Entre los candidatos se presentó un robot androide, con rasgos femeninos, Michihito Matsuda, quien (¿podemos decir "quién"?) quedó tercera en la segunda vuelta con 4.013 votos. Michihito había prometido acabar con la corrupción y ofrecer oportunidades justas y equilibradas para todos, y su propuesta había generado una elevada aceptación[2]. Según Michihito, el algoritmo, mediante el *Machine learning*, podría sustituir las debilidades emocionales de los seres humanos, causa de malas decisiones políticas, corrupción, nepotismo y conflictos, por un análisis objetivo de los datos generados acerca de las opiniones, expectativas, preferencias y costumbres de la ciudadanía[3].

Evidentemente, detrás de Michihito se encontraban seres humanos, en este caso Tetsuzo Matsumoto, Vicepresidente de Softbank, un proveedor de servicios móviles, y Norio Murakami, ex empleado de google Japón, quienes, según sus declaraciones, la presentaron a las elecciones con el objetivo de conseguir un gobierno justo, aplicando la inteligencia artificial.

Hasta aquí la anécdota, pero transitando a la categoría, ¿qué filosofía late en la base de todo ello? La convicción de algunos tecnocientíficos de que la IA permitirá recopilar datos sobre las necesidades relevantes de la ciudadanía, ayudará a planificar los recursos para satisfacerlas, calibrará las consecuencias de las medidas propuestas y asignará los recursos con justicia, basándose en datos objetivos. En el caso de Matsumoto, entiende que el sesgo emocional y motivacional de los seres humanos (el autointerés y la maximización del beneficio) le está arrastrando a la extinción. Una IA fuerte, exenta de rasgos emocionales, sería capaz de predecir hechos y consecuencias y aplicar políticas basadas en el bien común[4].

Sorprendentemente, este no es un relato aislado, sino que se han multiplicado los llamados "políticos virtuales". Por ej., SAM (Semantic Analysis Machine) se presenta en su cuenta oficial de Twitter desde noviembre de 2017 a las elecciones neozelandesas de 2020, alegando que fabrica decisiones basadas en hechos y opiniones y que no miente ni tergiversa información. Y también en el ámbito económico se produce un proceso de algoritmización en las tomas de decisiones. Existe un reconocido elenco de consejeros y directivos algorítmicos en instituciones, organizaciones y empresas, se coloca a un algoritmo frente a determinadas secciones o dentro de ellas[5]. Xerox, Google, Unilever, L'Oreal o

---

[2] Whiters, 2018.
[3] Matsumoto, 2018; Calvo, 2019.
[4] Matsumoto, 2018, 161.
[5] *Deep Knowledge Analitics*, un fondo de inversión del sector biotecnológico, en mayo de 2014 nombró a un algoritmo (*Vital*) como presidente de la junta directiva. Vital decide dónde debe invertir la empresa, junto a los demás miembros de la junta directiva, pero su voto es de calidad. Según Jessica Fontaine,

Amazon tienen un algoritmo dentro o al frente de la dirección de Recursos Humanos y también se multiplican en los medios de comunicación.

Sin duda este nuevo mundo plantea cuestiones éticas de gran calado, pero la primera de ellas, que será el punto de partida de esta intervención, consistirá en poner sobre el tapete la diferencia abismal que existe entre *hacer uso* de sistemas inteligentes (sean máquinas, algoritmos, robots) a la hora de tomar decisiones y *delegar* en esos sistemas inteligentes decisiones significativas para la vida de las personas y de la naturaleza.

Que Matsumoto se presente a las elecciones y, si gobierna, se sirva de Michihito como ayuda para tomar decisiones, no es lo mismo que poner el gobierno en manos de Michihito. Con los grandes avances en temas de IA, ¿se trata de que los seres humanos utilicen los sistemas inteligentes como instrumentos o de que estos sustituyan a los seres humanos? Y aquí surge la gran pregunta: ¿una ética de la inteligencia artificial es la que deben practicar los sistemas inteligentes desde sus propios valores, o es la que los seres humanos deberíamos adoptar para servirnos de los sistemas inteligentes?

Por poner un ejemplo, las tres leyes de la robótica que el profesor de Bioquímica y escritor de ciencia ficción Isaac Asimov introducía en los cerebros de los robots, y que son "formulaciones matemáticas impresas en los senderos positrónicos del cerebro" de los robots, aparecidas por primera vez en el relato "Círculo vicioso" de 1942, eran leyes que los robots debían cumplir para convivir con los humanoides[6]. Pero estas leyes ¿convierten a los robots en seres capaces de actuar por sus propias leyes, es decir, en seres autónomos, en el sentido de autonomía de que nos hablaba nuestro compañero Pedro Cerezo en su excelente intervención[7], o más bien son los diseñadores de los robots quienes inscriben esas leyes en los cerebros de seres que continuarán siendo heterónomos? ¿Insertar valores en los circuitos de los llamados coches autónomos para que "decidan" ante alternativas dolorosas (como por ejemplo, preservar la vida de los peatones o la de los pasajeros del coche) significa dotarles de una ética?[8].

Ciertamente, los 23 Principios de Asilomar, propuestos por el Future of Life Institute en la Conferencia Asilomar de 2017, a los que se considera como una ampliación de las Leyes de Asimov, pretenden asegurar que la IA sea diseña-

---

la portavoz de DKV, Vital puede tomar en consideración muchas más variables que un ser humano, y además, al carecer de emociones, no se enfada si se rechazan sus propuestas (Pardo, 2014).

[6] Es el escritor checo Karel Capek (1890-1938) quien utiliza por primera vez la palabra "robot" en su obra de teatro Robots Universales Rossum de 1920. Al parecer, el término procede de r'b del antiguo eslavo, que significa "esclavo" o bien del checo "robota", que significa "trabajo".

[7] Cerezo, 2019. Pedro Cerezo se ha ocupado del concepto de autonomía en diversos libros, artículos e intervenciones de la Real Academia de Ciencias Morales y Políticas.

[8] Una pregunta semejante se planteó a cuenta del nacimiento de la Neuroética: ¿se trata de una ética fundamental o de una ética aplicada? Ver Cortina, 2011.

da para el bien[9]. Consideran que los sistemas IA son instrumentos valiosos, que pueden empoderar a las gentes y de ahí que se formulen principios para conseguirlo. ¿En qué consiste entonces la ética de la inteligencia artificial?

A mi juicio, para responder a esta pregunta es necesario considerar al menos tres tipos de IA.

2. TRES TIPOS DE INTELIGENCIA ARTIFICIAL Y DE SU POSIBLE CORRESPONDIENTE ÉTICA

En principio, conviene recordar que es preciso hablar de "inteligencias", y no sólo de "inteligencia", en el caso de los humanos, en el de los animales, en el de los vegetales, e incluso en el de la tierra según la hipótesis Gaia. Podemos llamar entonces "inteligencia" en sentido amplio a la capacidad de perseguir metas, planificar, prever consecuencias de las acciones y emplear herramientas para alcanzar las metas. La inteligencia sería la capacidad de resolver problemas con instrumentos.

En cuanto a la *inteligencia artificial*, nace en 1955, en un congreso en Los Ángeles sobre máquinas que aprenden. John McCarthy introduce la expresión "inteligencia artificial" en 1956 y se refiere con ella a la creación de máquinas que pueden tenerse por inteligentes porque interactúan con los seres humanos hasta el punto de que una persona ya no sabe si está hablando con una máquina o con otra persona humana. Es lo que recibe el nombre de "test de Turing"[10]. La IA puede llegar a constituir un nuevo tipo de inteligencia[11]

Pero, por decirlo con mayor precisión, podemos recurrir a la caracterización del *High-Level Expert Group on Artificial Intelligence*, creado por la Comisión Europea, en sus Orientaciones Éticas para una IA confiable, publicadas en 2019, tras haber sacado a la luz y discutido con expertos dos borradores[12]. Según el texto de las *Guidelines*, los sistemas de IA son sistemas de software (y posiblemente también de hardware), diseñados por humanos que, dada una meta compleja, actúan en la dimensión física o digital percibiendo su entorno mediante la adquisición de datos, interpretando los datos recogidos, estructurados o no estructurados, razonando sobre el conocimiento o procesando la información derivada de estos datos y decidiendo las mejores acciones que hay que

---

[9] Las leyes de Asilomar se agrupan en 3 áreas: cuestiones de investigación, ética y valores, y cuestiones a largo plazo.

[10] Turing (1912-54) es el padre de la informática, gracias a su máquina (1936), que es una visión de IA. En 1950 defiende en *Mind* que los ordenadores pueden tener comportamientos inteligentes.

[11] López de Mántaras/Meseguer, 2017, 8

[12] *Ethics Guidelines for Trustworthy AI*, abril de 2019, p. 36.

realizar para alcanzar la meta. Los sistemas IA pueden utilizar reglas simbólicas o aprender un modelo numérico, y pueden también adaptar su conducta analizando cómo el entorno es afectado por las acciones previas.

Por otra parte, como disciplina científica, la IA incluye varios enfoques y técnicas, tales como *machine learning* (del que son ejemplos el *deep learning* y el *reinforcement learning*), el *machine reasoning* (que incluye planificar, programar, representaciones de conocimiento y razonamiento, búsqueda y optimización), y robótica (que incluye control, percepción, sensores y actuadores (actuators), y la integración de todas las demás técnicas en los sistemas ciberfísicos.

En este ámbito de la inteligencia artificial pueden distinguirse tres modalidades que, a mi juicio, plantean problemas éticos diferenciados[13]:

1) *La inteligencia superior o superinteligencia*. Con esta expresión nos referimos a un tipo de inteligencia que supera a la humana, de modo que las máquinas pueden sustituir al hombre. Esta modalidad de IA es la que da lugar a las propuestas transhumanistas y posthumanistas con la idea de la "singularidad".

John von Neumann fue uno de los primeros en vislumbrar la potencia de cálculo que un ordenador podía tener, y llega a afirmar que el progreso en la tecnología y los cambios en nuestra forma de vida "muestran signos de aproximarse a una especie de singularidad esencial en la historia de la especie". En 1983 Vinge menciona la posibilidad de una singularidad tecnológica y propone la superación de la mente humana por máquinas con inteligencia artificial. Pero la figura más conocida es Raymond Kurzweil, quien recoge la idea en *The Singularity Is Near: When Humans Transcends Biology*[14]. Según él, los humanos dejarán su soporte biológico y pasarán su inteligencia a las máquinas. Esto dará lugar a la *Singularity University*, fundada en 2008, cuyo lema es: preparando a la humanidad para un cambio acelerado de tecnología. Se supone que habrá un cambio de sustrato entre inteligencia humana e inteligencia artificial: la Singularidad necesita sobrepasar los límites impuestos por el tejido nervioso y el sustrato de la inteligencia artificial será de silicio. Desde esta perspectiva, los seres humanos somos un elemento más en la cadena de la evolución que culminará en esos seres singulares. No se trata en modo alguno del superhombre nietzscheano, para el que el cuerpo es esencial, sino de seres singulares en los que el cuerpo biológico será sustituido por la máquina. Será una especie nueva.

Ciertamente, existen amplias discrepancias en el ámbito de la IA sobre si estos pronósticos del "transhumanismo" y del "posthumanismo" van a cumplirse por tener base científica suficiente para ello. Algunos autores dan por sentado

---

[13] Llopis, 2019.
[14] Penguin, 2005; Lola Books, Berlin, 2012.

que se llegará a crear superinteligencias artificiales en este mismo siglo; en concreto, Kurzweil considera que en 2045 se conseguirá la singularidad tecnológica, gracias al incremento exponencial de las tecnologías de la información. Sin embargo, otros entienden que no existe base científica para esa suposición[15]. Pero la sola hipótesis ya abre un mundo de cuestiones éticas, que es preciso abordar para que el puñal no preceda a la brújula moral.

En primer lugar, los transhumanistas consideran que es un deber moral trabajar en la línea de intentar trascender la especie humana con todas sus imperfecciones para crear esos seres perfectos que compondrían la singularidad. Si el ser humano es intrínsecamente imperfecto, es un deber moral buscar su mejoramiento por medios técnicos[16]. Sin embargo, dejando de lado otras cuestiones, la pregunta se impone: ¿es realmente un deber moral construir seres superiores que van a plantear problemas como el de la convivencia de dos especies, una superior y otra inferior, que sería la nuestra? ¿No estaríamos abonando un mundo de amos y esclavos, en que los segundos no tendrían la menor posibilidad de revolución, sino que estarían a merced de las superinteligencias?

Por otra parte, ¿cuál será la ética de esas superinteligencias? Nick Bostrom, uno de los adalides del posthumanismo, aconseja integrar valores en esas inteligencias que, aprendiendo, se independizarán de los humanos[17]. Pero —a mi juicio— si esto fuera posible, y las máquinas aprendieran por su cuenta, poco podríamos hacer por conseguir que siguieran manteniendo como valores el respeto, la solidaridad, la justicia o la compasión. Serían los propios sistemas superinteligentes los que irían proponiendo sus valores y actuando o no de acuerdo con ellos. Ésta sí que sería una "ética de la inteligencia artificial", que no estaría en nuestras manos. ¿Es ahora un deber moral propiciarla?

Y sobre todo, en un mundo en que es una realidad sangrante el sufrimiento causado por las guerras, la pobreza, la aporofobia y la injusticia, ¿es un deber moral invertir una ingente cantidad de recursos en construir presuntos seres pluscuamperfectos, o es el modo en que empresas poderosas consiguen todavía más riqueza y poder? ¿No es una exigencia ética palmaria utilizar los grandes beneficios de la inteligencia artificial para resolver estos problemas acuciantes?

2) Un segundo tipo de inteligencia es la *inteligencia general*, aquella que puede resolver problemas generales. Ésta es la forma de inteligencia típicamente humana, y constituye el fundamento de la IA, en que trabajan las mentes más brillantes: el objetivo de la IA, como disciplina científica, es conseguir que una máquina tenga una inteligencia de tipo general, similar a la humana.

---

[15] López de Mántaras y Meseguer, 2017, 13.
[16] De este asunto, referido a la biomejora moral, me he ocupado en Cortina 2013 y 2017.
[17] Bostrom, 2014.

Fue John Searle quien distinguió ya en 1980 entre IA fuerte y débil[18]. La fuerte implicaría que un ordenador es una mente y es capaz de pensar igual que un ser humano, pero lo que intenta demostrar Searle es que la IA fuerte es imposible, porque la máquina carece de la *intencionalidad* por la que los humanos damos significado a lo que nos rodea: una máquina no conoce el significado de los símbolos que maneja. *Sin un cuerpo* las representaciones abstractas carecen de contenido semántico: no puede haber inteligencia general sin cuerpo[19].

Este punto es central: las máquinas carecen del conocimiento de *sentido común* que es posible por nuestras vivencias corporales. El cuerpo es esencial para dar significado a lo que nos rodea mediante la intencionalidad, para comprender e interpretar desde los contextos concretos, para contar con valores, emociones y sentimientos, para tomar decisiones desde ese êthos. La cuestión es entonces: ¿es posible dotar de sentido común a las máquinas, aunque no tengan un cuerpo como el humano? Realmente, la financiación que reciben quienes trabajan en ello es astronómica, pero por el momento no parece haberse logrado.

Sin embargo, en el caso de que fuera posible construir sistemas inteligentes con una inteligencia general como la humana, ¿tendríamos que aceptar que están dotadas de autonomía y, por lo tanto, son personas y que, en consecuencia, es preciso reconocerles dignidad y exigirles responsabilidad?, ¿tendrían derechos y deberes?, ¿deberíamos tratarlas con respeto y compasión?, ¿deberían ser ciudadanas del mundo político, elegibles como representantes en sociedades democráticas, sin estar manejadas por un ser humano?

Por el momento, parece sumamente improbable, y no sólo porque no se ha conseguido, sino también porque simularían intencionalidad, emociones, valores y sentido común, pero no dejaría de ser una simulación. Harían "como si" sintieran, pero para sentir se necesita un cuerpo[20].

3) Por último, *la inteligencia especial* es la que lleva a cabo trabajos específicos, es la propia de sistemas inteligentes capaces de realizar tareas concretas de forma muy superior a la inteligencia humana, porque pueden contar con una inmensa cantidad de datos y también con algoritmos sofisticados, que pueden llevar a resultados. Es lo que tenemos desde 1958 en diversos ámbitos.

El caso más conocido es el de la supercomputadora de IBM *Deep Blue*, que jugó al ajedrez con Gary Kasparov, campeón del mundo, en 1996 y 1997. En 1996 ganó Kasparov, pero en 1997, *Deep Blue* había aprendido de sus erro-

---

[18] Searle, 1980.

[19] A este respecto es célebre el experimento de la caja china del que habla Searle: una manipulación sintáctica sin comprensión semántica carece de sentido.

[20] Ferry, 2017, pp. 204 y 205; López de Mántaras y Meseguer, 2017.

res y derrotó a Kasparov. El revuelo fue enorme. Pero es que el sistema inteligente puede llevar a cabo tareas concretas contando con una infinidad de datos y con una capacidad de correlación muy superior a la de los seres humanos.

En este sentido, hay una gran cantidad de problemas que están siendo abordados con ayuda de buenos algoritmos, en el sector de la salud (analizar los síntomas de un paciente en muy distintas modalidades de la medicina, hacer un diagnóstico y proponer un tratamiento), en la predicción climatológica, en la productividad y eficiencia empresarial, en la comunicación, el ocio, la planificación del tiempo, el ahorro de tiempo, el abaratamiento de costes, en el asesoramiento a la hora de conceder un crédito, reconocer voces humanas y leer textos, aconsejar en el ámbito agrícola. El proceso consiste en construir un patrón, que permita adivinar el comportamiento futuro, porque se dice que somos humanos predecibles. La búsqueda sistemática de un patrón en un amplio registro histórico se llama *minería de datos* (*data mining*) y se utiliza de forma rutinaria tanto en investigación científica como en el mundo de los negocios[21].

Sin embargo, en todos estos casos el elemento directivo sigue siendo la persona humana que se vale de la potencia del sistema inteligente para calcular y tratar gran cantidad de datos, incluso para aprender de sus "experiencias".

Es en este tipo de IA en el que actualmente nos encontramos. No se trata, pues, por el momento de una ética de los sistemas inteligentes, sino de *cómo orientar el uso humano de estos sistemas de forma ética*.

3. ALGUNAS ORIENTACIONES ÉTICAS PARA EL USO DE SISTEMAS INTELIGENTES

Sin duda nuestro mundo es ya el de la digitalización y las inteligencias artificiales. El paso es irreversible y, por lo tanto, no cabe preguntar si debemos darlo, sino cómo hacerlo para conseguir el mayor bien posible. Las nuevas tecnologías son fuente de competitividad y de productividad y los países y organizaciones que se excluyan de ese mundo perjudicarán a sus propios miembros y a su entorno, porque perderán peso y relevancia en una carrera en que los demás seguirán progresando exponencialmente. Es uno de los dramas de la Unión Europea, que ha quedado dolorosamente rezagada frente a China y Estados Unidos, ha quedado condenada a la irrelevancia. Si en los debates anteriores a las elecciones del 28 de abril el tema de Europa no estuvo presente, tampoco lo estuvo el de la IA, que, sin embargo, es un gran reto para Europa y para España, si no quieren condenarse a la irrelevancia. ¿Cómo abordar con ética el nuevo mundo?

[21] Latorre, 91-95.

Afortunadamente, en el momento actual la realidad innegable de las éticas aplicadas ha dado cuerpo al sueño hegeliano de que la moral se encarne en las instituciones, porque un buen número de organismos está asumiendo su responsabilidad en esta ética aplicada a la IA y elabora informes muy valiosos. En 2017 vieron la luz los Principios Asilomar de la Inteligencia Artificial, a los que ya he aludido. Pero sobre todo en el contexto de la Unión Europea surgen propuestas de marcos éticos como el *Ethical Framework for a Good AI Society: Opportunities, Risks, Principles and Recommendations*, propuesto por el AI-4People en diciembre de 2018, las *Ethics Guidelines for Trustworthy AI del High-Level Expert Group on Artificial Intelligence* de la Comisión Europea de abril de 2019 (borradores en marzo y 18 de diciembre de 2018), o la *Declaración de Derechos Humanos para un Entorno Digital*, que la Universidad de Deusto presentó el 26 de noviembre de 2018.

Dado el escaso tiempo del que disponemos intentaré entresacar lo que, a mi juicio, es nuclear para una ética de la IA. En los últimos documentos mencionados la propuesta es clara: se trata de trazar el marco ético de una *IA confiable*, porque la confianza ha de ser la piedra angular de las sociedades[22]. En la línea seguida habitualmente por los documentos de la UE, se trata de unir progreso técnico y progreso ético, de ahí que una IA confiable en productos y servicios será el camino de la ciudadanía europea para lograr una ventaja competitiva. Si Europa quiere ser líder global, también en IA, debe maximizar los beneficios de los sistemas inteligentes, previniendo riesgos: un enfoque confiable posibilita "competitividad responsable", al ofrecer a los afectados una confianza, que llevará a Europa a ser líder global. La ventaja competitiva será la ética[23].

Las *Guidelines* se dirigen a todos los *stakeholders* de la IA que quieran adoptarlas voluntariamente para poner por obra su compromiso y que deben tener en cuenta la diversidad de campos a los que se aplican: (tratamientos médicos, relación entre empresa y consumidor, entre empresa y empresa, entre empleadores y empleados, en las relaciones público-ciudadano).

---

[22] De hecho se dice textualmente que "la nueva revolución tecnológica, referida a tecnologías digitales inteligentes (como la IA, los algoritmos machine learning, el deep learning y las redes coneccionistas, la mecatrónica y la robótica), ha llevado a la necesidad de reflexionar sobre el marco ético que queremos dar al diseño, producción, uso y gobierno de la IA, la robótica y los llamados sistemas "autónomos".

[23] La meta de estas *Guidelines* es promover una *IA confiable*, con tres componentes: 1) legal, cumpliendo las legislaciones, 2) ética, ateniéndose a los principios y valores, y 3) robusta desde el punto de vista ético y social, porque aun con buenas intenciones puede causar daños. Las *Guidelines* diseñan un marco con estos tres componentes, pero se ocupan sobre todo de los dos últimos. El Capítulo I desarrolla los principios éticos de respeto a la autonomía humana, prevención del daño, *fairness* y explicabilidad. El capítulo II desarrolla 7 exigencias para una IA confiable: 1) agencia humana y supervisión (*oversight*), 2) robustez técnica y seguridad, 3) privacidad y gobierno de los datos, 4) transparencia, 5) diversidad, no discriminación y *fairness*, 6) bienestar medioambiental y social, y 7) rendición de cuentas. Habla de los métodos técnicos y no técnicos para implementar esas exigencias.

Y en cuanto a la pregunta que formulábamos desde el comienzo de esta intervención, la propuesta se reconoce explícitamente como *humano-céntrica*. Se afirma sin ambages que los sistemas inteligentes son instrumentos para mejorar la vida humana y la naturaleza, y no fines en sí mismos.

De ahí que no se trate de una competencia entre inteligencias —humana y artificial—, sino que la segunda tiene que estar supeditada a la primera, sin posibilidad de sustitución. Sin duda porque así lo exige el presupuesto ético básico, que es el reconocimiento de la autonomía de las personas humanas. Los principios éticos de explicabilidad, beneficiar, no dañar y justicia tendrán como base el reconocimiento de la autonomía y la dignidad.

4. UN MARCO ÉTICO PARA LA INTELIGENCIA ARTIFICIAL

Un buen marco es el que ofrece el AI4People del Atomium European Institute, que cuenta con cuatro principios clásicos, aplicados a entornos digitales, a los que añadiría un quinto: la explicabilidad y la rendición de cuentas. Los principios clásicos serían el de beneficencia, que exigiría ahora poner los progresos al servicio de todos los seres humanos y la sostenibilidad del planeta; el de no-maleficencia, que ordenaría evitar los daños posibles, protegiendo a las personas en cuestiones de privacidad, mal uso de los datos, en la posible sumisión a decisiones tomadas por máquinas y no supervisadas por seres humanos; pero también el principio de autonomía de las personas, que puede fortalecerse con el uso de sistemas inteligentes, y en cuyas manos deben ponerse tanto el control como las decisiones significativas; y, por supuesto, el principio de justicia, que exige distribuir equitativamente los beneficios. A ellos se añadiría un principio de explicabilidad y accountability, porque los afectados por el mundo digital tienen que poder comprenderlo.

Dada la centralidad de los principios de autonomía y explicabilidad, desearía, por terminar, hacer unos puntualizaciones sobre ellos.

4.1. Autonomía y dignidad

Aunque a menudo se hable de "coches autónomos" y de "sistemas autónomos", lo cierto es que la palabra "autonomía" sólo puede aplicarse a los seres humanos. La autonomía no consiste sólo en tomar decisiones y actuar con *independencia* respecto a otros, sino en la *capacidad de autolegislarse* y de *autodeterminarse*, es decir, en la capacidad de poder determinarse a sí mismo a seguir las leyes o eludirlas, la capacidad de darse metas y seguirlas, y de optar

no solo por normas idiosincráticas, sino también por leyes universales[24]. Estas capacidades están ligadas al reconocimiento de la *dignidad* de las personas, que es el núcleo de las orientaciones éticas. De ahí que sea un deber preservar y potenciar la autonomía y la agencia de las personas, también con el uso de sistemas inteligentes, llegando a la "inteligencia aumentada", pero utilizando esos sistemas como instrumentos. Son las personas las que tienen un valor intrínseco y las personas son seres humanos.

Por el contrario, los mal llamados "sistema autónomos" no lo son realmente: son artefactos, son autómatas, a pesar del aprendizaje profundo (*deep learning*). Los sistemas inteligentes pueden resolver problemas y actuar independientemente de los seres humanos, pero no son autónomos. No pueden decidir qué se debe hacer, qué metas hay que perseguir. De aquí se sigue que son los seres humanos los que tienen dignidad y merecen respeto, pero también que son responsables, porque la responsabilidad exige autonomía, entendida como la capacidad de autodeterminación.

Ciertamente, del reconocimiento de la autonomía de las personas se siguen conclusiones como las siguientes.

No pueden ponerse en manos de máquinas inteligentes decisiones que afectan a la vida de las personas, sin supervisión humana, simplemente aplicando un algoritmo, que es una fórmula matemática, que a menudo ni siquiera sus creadores son capaces de explicar, y que suele ser diseñado por encargo por organizaciones distintas a las que lo aplican[25]. Siempre tiene que ser un ser humano quien tome la decisión última y deba dar razón de ella, en caso necesario.

No es extraño que el Reglamento General de Protección de Datos de la UE (que entró en vigor en mayo de 2018) establezca que los ciudadanos europeos no deben ser sometidos a decisiones "basadas únicamente en el proceso automático de datos", no deben ser sometidos "a las prácticas de contratación digital sin intervención humana". En este contexto, no es de extrañar que los trabajadores rechacen que sea una máquina quien les mande, prefieren otra persona humana.

Por otra parte, la responsabilidad moral no puede atribuirse a la "tecnología autónoma", sino que el control humano es esencial para hablar de responsabilidad moral. Los humanos, y no los computadores y algoritmos, han

---

[24] Cerezo, 2019; Conill, 2019.

[25] La palabra "algoritmo" viene del latín tardío "algobarismus" y ésta, del árabe clásico "hisabu Igubar", que significa cálculo mediante cifras arábigas (RAE, 209). Un algoritmo es 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. La aplicación a la IA lleva a considerarlo como "un código software que procesa un conjunto limitado de instrucciones" (Monasterio, 2017, 186).

de permanecer en el control y ser moralmente responsables. En qué modo han de servir al bien deben decidirlo los humanos, no un algoritmo.

No deja de ser curioso, sin embargo, que se haya hablado de "personas electrónicas". La Comisión de Asuntos Jurídicos del Parlamento Europeo propuso a la Comisión Europea en un informe de 2016 crear una personalidad jurídica específica para los robots, "de modo que al menos los robots autónomos más complejos puedan considerarse personas electrónicas con derechos y obligaciones específicos, incluida la obligación de reparar los daños que puedan causar; la personalidad electrónica se aplicaría a los supuestos en que los robots puedan tomar decisiones autónomas inteligentes o interactuar con terceros de forma independiente".

Por último, el respeto a la dignidad humana exige que un ser humano sepa siempre si está hablando con otro ser humano o con una máquina, no debe recurrirse al engaño.

4.2. Explicabilidad y rendición de cuentas

Totalmente ligado al respeto a la autonomía se encuentra el principio de explicabilidad o de trazabilidad, según el cual, los afectados tenemos derecho a controlar el uso de nuestros datos y a conocer los algoritmos que los manejan. Porque los seres humanos tienen sesgos, como se ha mostrado hasta la saciedad, pero también los sistemas autómatas los tienen y son más invisibles que en el caso de los humanos. Entre otras razones, porque los diseñadores introducen los sesgos en los sistemas inteligentes, consciente o inconscientemente. Es revelador uno de los ejemplos que presenta la analista financiera, experta en matemáticas, Cathy O'Neil en su interesante libro *Armas de destrucción matemática*. El relato es el siguiente.

En 2007 el alcalde de Washington D. C., Adrian Fenty, quería corregir la situación de las escuelas deficientes de la ciudad. La teoría generalmente aceptada era que los alumnos no aprendían lo suficiente porque los profesores no trabajaban bien. La rectora en centros educativos de primaria y secundaria de Washington, Michelle Rhee desarrolló una herramienta de evaluación del personal docente, llamada IMPACT, y aplicándola despidió a los profesores que no alcanzaron un determinado nivel. Con la aplicación del algoritmo se trataba de evitar sesgos humanos en la apreciación de los profesores (amistades, subjetivismos) y prestar atención a valoraciones objetivas: las puntuaciones de los alumnos en matemáticas y en lectura. Entre los profesores despedidos figuraba Sarah Wysocki, una maestra extraordinariamente apreciada por director, padres y alumnos. La puntuación en competencias lingüísticas y matemáticas, que suponía la mitad de la valoración global, fue muy baja. Dolida por la puntuación preguntó Sarah cómo se había llegado a ella y no encontró

respuesta en la consultora Mathematica que elaboró el algoritmo, porque los análisis se subcontrataban a programadores y estadísticos, que dejan que hablen las máquinas. El modelo es una caja negra (11-20). El final del relato es agridulce Sarah fue contratada por una escuela privada, que se sintió feliz de poder contar con ella, porque no se había sometido a máquinas para contratar.

Ciertamente, si tomamos en serio el principio de autonomía y el hecho de que los seres humanos son interlocutores válidos cuando se trata de asuntos que les afectan, los afectados por el mundo digital tienen que poder comprenderlo; tienen que conocer la trazabilidad de los algoritmos que afectan a sus vidas: quién los construye, con qué sesgos, con qué objetivos. Teniendo en cuenta que en un mundo global digitalizado los afectados somos a menudo todos los seres humanos, el imperativo de la explicabilidad es verdaderamente exigente.

Pero además la supervisión humana se hace necesaria por razones de eficiencia, como es palmario en el caso de Sarah.

4.3. No dañar

En cuanto al principio ético clásico de "no dañar" recuerda que no se debe todo lo que se puede. Y, ciertamente, es muy difícil seguirlo teniendo en cuenta que en un mundo en competencia otras empresas sí que van a comercializar el producto, a poner en marcha la innovación, y no sólo otras empresas, sino países que son más laxos que otros y no se preocupan en exceso de no dañar.

En cualquier caso, la obligación de respetar a los seres humanos ordena evitar los daños posibles, protegiendo a las personas mediante "derechos digitales" (4.ª generación): derecho a la privacidad en entornos digitales, protección de la integridad personal (intimidad), derecho a la propia imagen y honra, a estar a salvo de contenidos nocivos (discursos de odio, ciberacoso), y al mal uso de los datos obtenidos, que no pueden utilizarse sin consentimiento de la persona y para fines distintos a los que justifican su obtención.

4.4. Promover un mundo justo

Por último, pero no en último lugar, sino tal vez en el primero, el principio de justicia exige distribuir equitativamente los beneficios de las nuevas tecnologías, porque todos son afectados, promocionar un mundo inclusivo, en que la brecha digital no divida a la humanidad con una nueva fórmula. El desigual acceso a los bienes tecnocientíficos socava la cohesión.

Obviamente, un capítulo sustancial, en el que no podemos entrar, es el de la transformación del mundo laboral. Teniendo en cuenta que el ADN de la UE son los derechos fundamentales, entre los que cuentan los sociales (igualdad de oportunidades, acceso al mercado de trabajo, protección social e inclusión), se hace necesario mejorar las competencias digitales de la ciudadanía y organizar el mundo del trabajo de tal modo que no queden excluidos. Es preciso propiciar una transición a la sociedad digital que proteja los derechos sociales de las personas, atendiendo a propuestas como una renta básica universal, la predistribución, o que los robots paguen impuestos, porque sustituyen a las personas.

Es urgente pensar en estas medidas si realmente deseamos que la promoción de la inteligencia artificial merezca confianza, promueva un mundo confiable.

REFERENCIAS BIBLIOGRÁFICAS

AI4People, *Ethical Framework for a Good AI Society: Opportunities, Risks, Principles and Recommendations*, Minds and Machines (2018): 28.689-707

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

Bostrom, N. (2014): *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press (trad. esp. *Superinteligencia, caminos, peligros, estrategias*, Teell, 2016).

Calvo, P. (2019): "Gobierno algorítmico: el neuroaprendizaje moral de las máquinas en la política y la economía", en prensa.

Cerezo, P. (2019): "A vueltas con la autonomía: tensiones, aporías y perplejidades", sesión del 2 de abril de 2019, RACMYP.

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

Cortina, A. (2011): *Neuroética y neuropolítica*, Tecnos, Madrid.

—— (2013): "Neuromejora moral: ¿un camino prometedor ante el fracaso de la educación?", publicado en *Anales de la Real Academia de Ciencias Morales y Políticas*, Madrid, año LXV, n.º 90, 313-331.

—— (2017): *Aporofobia, el rechazo al pobre*, Paidós, Barcelona.

Ferry, L. (2017): *La revolución transhumanista. Cómo la tecnomedicina y la liberización del mundo van a transformar nuestras vidas*, Alianza (original: Plon, 2016).

González Páramo, J. M. (2016): *Reinventar la banca: de la gran recesión a la gran disrupción digital*, Real Academia de Ciencias Morales y Políticas, Madrid.

—— (2017): "Cuarta revolución, empleo y Estado del Bienestar", RACMYP, 5 de diciembre.

High-Level Expert Group on Artificial Intelligence set up by the European Comission (2019): *Ethics Guidelines for Trustworthy AI*, abril de 2019 (borradores en marzo y 18 de diciembre de 2018).

Keane, J. (2019): "La nueva era de la revolución de las máquinas", *Letras libres*, abril, año XVIII, 24-32.

Kurzweil, R. (2005): *The Singularity Is Near: When Humans Transcends Biology,* Penguin (hay trad. esp. en Lola Books, Berlin, 2012).

Latorre, J. I. (2019): Ética para máquinas, Ariel, Barcelona.

Llopis, R. (2019): "La inteligencia artificial como factor de innovación en la empresa", conferencia pronunciada en el Seminario de la Fundación ÉTNOR el 14 de marzo de 2019.

López de Mántaras, R. y Meseguer, P. (2017): *Inteligencia artificial*, Los Libros de la Catarata/CSIC, Madrid.

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

——(2018b): 4 de abril de 2018, 4:12, Twitter.

Monasterio, A. (2017): "Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos", *Dilemata*, 185-217.

O'Neil, C. (2018): *Armas de destrucción matemática*, Capitán Swing, Madrid.

Pardo, P. (2014): "Un algoritmo sentado en el consejo de un fondo chino", *El Mundo*, 9 de junio.

Parlamento Europeo, Comisión de Asuntos Jurídicos, Proyecto de informe con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103): 31.5.2016.

http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+-TA+P8-TA-2019-0081+0+DOC+XML+V0//ES

Searle, J. R. (1980): "Minds, Brains and Programs", Behavioral and Brain Sciences 3 (3), 417-457.

Villar Mir, J. M. (2019): "Los retos de la era digital", discurso pronunciado en la RACMYP el 26 de febrero de 2019.

Whiters, P. (2018): "Robot to run for mayor in Japan promising 'fairness and balance", Express, https: www, express.con.uk/news/world/947448/robots-japan-tokyo-mayor-artificial-intelligence-ai-news