LA PARADOJA DE LA TRANSPARENCIA EN LA IA: OPACIDAD Y EXPLICABILIDAD. ATRIBUCIÓN DE RESPONSABILIDAD

THE PARADOX OF TRANSPARENCY IN AI: OPACITY AND EXPLAINABILITY. ALLOCATION OF RESPONSABILITY

F. Javier Blázquez Ruiz
Universidad Pública de Navarra, Pamplona, España
javier.blazquez@unavarra.es

Recibido: septiembre de 2022
Aceptado: noviembre de 2022

Palabras clave: algoritmos, transparencia, explicabilidad, trazabilidad, cajas negras, responsabilidad.
Keywords: algorithms, transparency, explainability, traceability, black boxes, responsability.

Resumen. Con la irrupción de las nuevas tecnologías, la Inteligencia Artificial (IA) está convirtiéndose en una especie de nueva religión monoteísta, y los algoritmos, como si fueran oráculos, se tornan cada vez más omniscientes. No importa su grado de opacidad o la falta de trazabilidad que ostenten. Cada día que pasa el culto a los algoritmos deviene cada vez más reverencial. Parecen sagrados. Entre tanto, las cajas negras impiden que se cumplan los principios de transparencia y explicabilidad. A su vez la atribución de responsabilidad tiende a diluirse.

Abstract. With the emergence of new tecnnologies, Artificial Intelligence (IA) is becoming a kind of new monotheistic religion, and algorithms, as if they were oracles, they are more and more omniscient. It does not matter their degree of opacity or the lack of traceability they suffer. With each passing day the cult of algorithms becomes more reverential. They seem sacred. Meanwhile, black boxes prevent the exercise of the principles of transparency and explainability. In turn, the attribution of responsibility tends to be diluted.

1. Introducción: la paradoja de la transparencia

1. Tal vez sea pronto para confirmar las palabras del escritor de origen austríaco S. Zweig en su célebre ensayo *El mundo de ayer. Memorias de un europeo* (2012) cuando precisaba "Obedeciendo a una ley irrevocable, la historia niega a los contemporáneos la posibilidad de conocer en sus inicios los grandes movimientos que determinan su época". En este mismo sentido Thomas Khun advertía en *La estructura de las re-*

voluciones científicas (2005) sobre la dificultad de apreciar la trascendencia que puede adquirir la irrupción de un nuevo paradigma en un momento determinado en el que emerge y compite con otros arquetipos que se mantienen vigentes, hasta que se consolida de forma irreversible.

Sin embargo, es muy posible que estemos asistiendo actualmente a un acontecimiento singular en medio de un periodo de transición tecnológica. De hecho, apenas han transcurrido seis decenios desde la aparición de la IA con sus múltiples aplicaciones (López de Mántaras, R., 2019, 14). No obstante, lejos de tratarse de un proyecto innovador de contornos indefinidos o materia asociada a la ciencia-ficción, la IA forma ya parte activa de la realidad que nos rodea. Y podría decirse que ha venido para quedarse.

Es evidente que las posibilidades que ofrece su desarrollo y el manejo de macro datos (Big data) son ingentes. De hecho, la IA y los Big Data abarcan sectores tan diversos como las finanzas, la actividad sanitaria, el campo de las comunicaciones, el transporte o la vigilancia policial y la seguridad ciudadana, entre otros.

Y es posible que con el paso del tiempo los sistemas de IA lleguen a provocar cambios amplios y profundos en la estructura de la sociedad, tal vez equivalentes a los que originaron en su momento la máquina de vapor, la invención de la telefonía o la expansión de la aviación, entre otros, con la consiguiente transformación económica y social que constituyeron en su momento.

En realidad, cada día que pasa vivimos rodeados de algoritmos aunque no seamos conscientes de la incidencia que provocan en nuestra vida. El universo de la técnica ha empezado a permear y configurar el contexto social hasta el punto de que ámbitos tan diferenciados en un pasado reciente como tecnociencia y sociedad, que antes convivían e interactuaban solo de manera coyuntural, ahora se encuentran imbricados estructuralmente, y resulta cada vez más difícil separar uno del otro (Alonso, J., 2018, 2.)

2. Una sucinta mirada retrospectiva nos permite recordar que en siglos precedentes, la irrupción del conocimiento científico en el preludio de la Modernidad, tras el eclipse secular al que fue sometido el cultivo de la racionalidad durante varias centurias, resultó tan destacado como irreversible. Su advenimiento, progresivo, fue "ante todo, el resultado de su emancipación de los lazos en los que la teología la mantuvo cautiva durante la Edad Media" (Kelsen, H., 2006, 15). A partir de entonces el auge del Renacimiento impulsó el afán de búsqueda e innovación en campos tan diversos que logró erigir al ser humano en protagonista de su destino.

Y sin embargo, la evolución posterior de la tecnociencia ha seguido un camino ajeno a las expectativas previstas. Lo mismo cabe decir del poder creciente que ha alcanzado, que parece no ofrecer límites. Tal y como recordaba M. Heidegger en su *Carta sobre el humanismo* con precisión "el sentido original de la técnica no era el dominio, sino una forma de conocimiento que fabricaba útiles al servicio de metas auténticas, verdaderas". No obstante, con el paso del tiempo "la técnica ha perdido ese impulso originario al convertirse en un instrumento de dominación" así como de control efectivo sobre la vida cotidiana de los ciudadanos (Heidegger, M., 2013).

En ese sentido se orientan igualmente las críticas provenientes de los miembros de la Escuela de Frankfurt, que nos re-

cuerdan de forma explícita que la razón ilustrada aspiraba a emancipar al ser humano de su minoría de edad y pretendía instaurar un orden social y político en el que fuera posible aplicar y desarrollar los ideales ilustrados de libertad, igualdad y fraternidad. Pero estas aspiraciones no se han colmado. Por el contrario, han fracasado porque la civilización occidental, que tanto ha valorado desde Platón el cultivo de la racionalidad humana –como critica una y otra vez María Zambrano (1992) a lo largo de sus obras- ha olvidado la originaria unión del ser humano con la naturaleza y se ha alejado de ésta para doblegarla, dominarla y explotarla.

A partir de esa dinámica, la civilización occidental en lugar de hacer uso de una racionalidad crítica en sentido amplio, que permitiera establecer ideales y fines para los seres humanos –éticos, estéticos, jurídicos- ha cultivado únicamente la racionalidad instrumental con el fin de tratar y someter a la naturaleza. Pero ese intento de dominio, a través de la técnica, que había surgido para sobrevivir y para satisfacer sus necesidades vitales se ha visto acompañado por la imposición de unas determinadas leyes que se han impuesto a sí mismo y a los otros, convirtiéndose en un pesado lastre. El problema es que, como como advertía Ortega y Gasset en el primer tercio de siglo XX (2019, V) tanto el sentido como la causa de la técnica están fuera de ella. Les son ajenos. Se encuentran "en el empleo que da el hombre a sus energías vacantes, liberadas por aquella. La misión inicial de la técnica es esa: dar franquía (liberar) al hombre para poder vacar a ser sí mismo".

Posteriormente Max Horkheimer, ahondaba en la misma dirección y precisaba que la razón instrumental es tan solo una dimensión restringida de la racionalidad humana. Ha devenido en un tipo de razón que, convirtiendo al ser humano en amo y dominador de la naturaleza**,** le llena de innumerables medios técnicos y materiales, pero al mismo tiempo le deshumaniza y empobrece progresivamente. Por todo ello el dominio de la razón instrumental, del pensamiento calculador y pragmático, utilitario, ha debilitado el pensamiento reflexivo que nos permite conformar y desarrollar una identidad personal, que a su vez nos facilita arraigo con la naturaleza y procura el vínculo social. Como resultado de ese proceso unidimensional, se ha generado una sociedad contraria a la que pretendía alcanzar: la sociedad industrializada a partir de un "pensamiento administrado", homogéneo, uniforme, que cada día utiliza un lenguaje más empobrecido y mecanizado (Horkheimer, M, 2010).

Otro antecedente digno de destacar, en términos históricos, respecto del riesgo que conlleva el uso sobredimensionado de la racionalidad técnica, podríamos encontrarlo en la obra narrativa de Mary Shelley: ***Frankenstein*****,** que además de estar impregnada por el auge de la ciencia moderna, entonces emergente, puede considerarse también como una novela representativa de una época presidida por la exaltación continua de la naturaleza, mientras nos alerta de los peligros que conlleva el desarrollo de la tecnociencia y sus múltiples aplicaciones.

Tal y como evoca el subtítulo de la obra: *El nuevo Prometeo*, la autora se alinea con otros pensadores que se planteaban preguntas y formulaban advertencias sobre el futuro de la humanidad y el curso que puede seguir la implementación de la tecnología. No es de extrañar que propuestas actuales de carácter innovador que generan debates de cierta trascendencia

como sucede con el *Transhumanismo* y la *Singularidad tecnológica* (Diéguez, A., 2017) derivada de la IA encuentren precedentes tanto en la historia de las religiones occidentales como en el pensamiento filosófico, especialmente si nos fijamos en la tradición judeocristiana y en el platonismo. En realidad, existe una tradición tanto en la cultura occidental como en otras culturas no occidentales en virtud de la cual, pueden encontrarse iniciativas y propuestas innovadoras para crear seres vivos a partir de materia inerte. Así puede rastrearse en las historias de la creación provenientes de las tradiciones sumeria, china, judía, cristiana y musulmana. De hecho los antiguos griegos ya albergaban la idea de "crear humanos artificiales, en particular mujeres artificiales" (Coekelberg, M., 2020, 27)**.**

Sin embargo, si nos detenemos brevemente en esta novela gótica, *Frankenstein*, publicada en 1818, la autora –hija de la filósofa Mary Wollstonecraft– se plantea la creación de vida inteligente a partir de materia inanimada como resultado de un complejo proyecto biomédico. El personaje principal, el joven estudiante de medicina Víctor Frankenstein, investiga apasionadamente durante varios meses, sin contacto con el exterior, y consigue crear, con el concurso de la incipiente electricidad, un ser vivo de aspecto humano a partir de diversas partes de cuerpos provenientes de cadáveres.

Pero después, una vez concluido su proyecto, el artífice que ha insuflado vida al engendro, se desentiende por completo del logro alcanzado, y a partir de entonces, tras ignorar el curso que puede seguir su existencia, pierde el control sobre su criatura con las consecuencias subsiguientes. Incluídas diversas muertes. No obstante, conviene precisar que el argumento central de esta novela no trata de erigirse en un alegato crítico contra las posibilidades que ofrecía en aquel momento la ciencia y su correspondiente aplicación. En modo alguno. La obra refleja como si se tratara de un espejo el ambiente de inquietud científica que rezuma la época, pero el mensaje principal que se desprende del texto incide en que los científicos no pueden permanecer ajenos y "necesitan asumir la responsabilidad de sus creaciones. El monstruo huye, pero lo hace porque su creador lo rechaza. Es importante tener esta lección en mente para la ética de la IA" (Coekelbergh, M., 2020, 28).

De ahí que podamos encontrar en esta obra clásica una versión moderna, y por ende anticipada en el tiempo, del problema que conlleva para el conjunto de la sociedad la pérdida del control sobre las innovaciones tecnológicas, escenificada en diversas películas de ciencia ficción que han alcanzado un éxito destacado en las pantallas a lo largo de las últimas décadas. Nos advierte de la necesidad de tener presentes de forma inexcusable las implicaciones éticas, jurídicas y sociales de la investigación científica. Esa fue precisamente la razón principal de que el premio Nobel de medicina J. Watson –descubridor con Crick de la doble hélice– impulsara la creación de ELSI (Ethical, legal and social implications) mientras se desarrollaba el megaproyecto Genoma Humano desarrollado tanto por inicitaiva pública como privada (Blázquez Ruiz, F. J., 1999).

2. Explicabilidad y trazabilidad

1.No cabe duda de que, como apuntábamos supra, las posibilidades que ofrece actualmente la IA son ingentes hasta el

punto de que en los últimos años el orbe de la Inteligencia Artificial se ha convertido en una prioridad estratégica para los países más avanzados que tratan de convertirse en líderes mundiales en las nuevas tecnologías. Sin embargo, a la hora de evaluar el auge de la IA y examinar las consecuencias de sus aplicaciones, conviene tener presente no solo las posibilidades y retos que ofrece, sino también los riesgos y el eventual impacto negativo que puede provocar entre poblaciones vulnerables cuyos derechos son susceptibles de ser conculcados más fácilmente.

Porque no todo es luz en el complejo universo de los algoritmos. También hay que tener presente la existencia de un lado oscuro, menos visible, tal y como sucede con la luna respecto de la cual solo vemos su cara iluminada, pero no tanto la sombra que proyecta. De hecho, existe otra vertiente de la IA apenas perceptible, más opaca, que no se muestra, pero que se encuentra impregnada de brumas provenientes de los intereses y expectativas que defienden las empresas multinacionales.

A este respecto, es fácil constatar en la vida cotidiana cómo las entidades bancarias, los buscadores de internet, las diversas compañías telefónicas, etc. se nutren copiosamente de nuestros datos que obtienen a través de diversos medios, valiéndose de las pantallas e incluso de eufemismos como "copias de seguridad" a la hora de succionar nuestros teléfonos móviles. Sin embargo, nosotros no podemos recabar ni obtener información de las empresas que nos prestan servicios ni tampoco someter los algoritmos a ese mismo escrutinio.

Se produce así lo que podría denominarse "paradoja de la transparencia" en virtud de la cual el acceso, procesamiento de datos masivos y su respectiva combinación, permiten acceder -de forma invasiva- a través de las aplicaciones a información personal de carácter privado. Pero ese proceso continuo de apropiación se produce solo en una dirección. Siempre es unidimensional y por tanto esa relación que se establece es manifiestamente asimétrica.

En muchos casos tan siquiera somos conscientes de que existe un algoritmo que está detrás de cada consulta que realizamos. Tampoco pensamos en las consecuencias que pueden provocar posteriormente sobre nuestras vidas, que se encuentran cada vez más expuestas a una especie de panóptico permanente del que resulta difícil liberarse por los vínculos continuos que van generándose (Foucault, M., 2022). En este contexto, a veces da la impresión que en medio de esta dinámica de sobreexposición, la novela de ficción distópica, *1984*, escrita por J. Orwell fuera un vaticinio premonitorio del grado de control y supervisión permanente al que nos vemos sometidos desde la irrupción de estas tecnologías, a pesar de que esta obra fuese publicada en 1949, pocos años después de finalizar la Segunda Guerra Mundial.

En realidad, los algoritmos, que han sido diseñados inicialmente por programadores contratados *ad hoc* van generando otros nuevos algoritmos cuyo grado de complejidad resulta cada vez mayor, hasta el extremo de que su comprensión deviene inaprehensible. No ha de extrañar por tanto que el principio de transparencia o *explanability,* se haya convertido en uno de los requisitos cada vez más demandados y que debería cumplir, obligatoriamente, cualquier empresa u organización vinculada a la IA. Máxime si los algoritmos que se introducen se tornan a veces inextricables incluso para el propio creador.

2.Y es que, no podemos obviar que el procesamiento de datos está preñado de una manifiesta complejidad. Su manejo pasa por muchas manos e intervienen personas diferentes, de ahí que, después, su huella no sea fácil de seguir. El primer escollo que puede aparecer es el riesgo de parcialidad a la hora de determinar los datos que son recabados o las preguntas que se formulan para poder obtener esos datos. Ya que esta dinámica inicial dista de ser en todos los casos imparcial. Además, cabe la posibilidad de que los algoritmos diseñados sean tan sesgados como los científicos o técnicos que los han programado. Por tanto, podrán reflejar, de manera consciente o no, los prejuicios e intereses de todos los que han intervenido para configurar el modelo. Y este riesgo de imprimir sesgos puede darse en diferentes momentos del proceso (Alonso, J., 2018, 3).

A consecuencia de lo cual, pueden quedar ocultos y permanecer sedimentados por el manto de la opacidad. En tal caso la obscuridad que preside los tratamientos de datos y la subsiguiente combinatoria correlacional permite a sus creadores mantener en secreto el diseño del algoritmo así como los pasos que se han dado, pudiendo aducir, eventualmente, el derecho a la propiedad intelectual. El problema adicional que puede derivarse es que, la aplicación de los sistemas de inteligencia artificial, pueden generar después, o incluso perpetuar, prejuicios o estereotipos establecidos, desfavoreciendo más aún a grupos étnicos o sociales que se han visto históricamente marginados.

Esa es una de las razones por las que se apela cada vez más al principio de explicabilidad, cuya puesta en práctica conlleva hacer comprensibles no solo el funcionamiento de los algoritmos que han sido conformados, sino también los procedimientos seguidos en su elaboración así como su relación con los resultados obtenidos, facilitando de ese modo la trazabilidad de todo el proceso (Megías Quirós, J. J., 2022, 148). Solo así, los ciudadanos que puedan verse afectados negativamente por una decisión basada en la aplicación de la IA podrán conocer en un lenguaje que resulte comprensible, los motivos por los que un algoritmo determinado ha llevado a tomar esa decisión y no otra (López Baroni, M. J., 2019, 5-28). Y a continuación, podrán adoptar medidas jurídicas y recurrir si lo estiman oportuno, llegado el momento.

En este sentido, es fácil constatar cómo con el paso del tiempo, estos problemas relacionados directamente con la opacidad están provocando una pérdida progresiva de confianza entre los ciudadanos respecto de las innovaciones aportadas por la IA y en sus programadores**.** De ahí que, o se corrige esa tendencia, o en otro caso la falta de transparencia persistente va a provocar un déficit creciente de confianza en el manejo de la tecnología y en sus expectativas futuras, amén de innumerables demandas en el terreno legal.

En realidad, la actividad de explicar decisiones forma parte de lo que los humanos hacen habitualmente cuando se comunican y toman decisiones en los más diversos ámbitos, ya sea en el campo político, laboral o familiar. Pero además, el hecho de explicar deviene también en todo Estado de derecho una exigencia cívica desde el plano moral y jurídico. De ahí que el requisito de la explicabilidad se erija en una condición necesaria respecto del comportamiento y la toma de decisiones que sea responsable (Coekelbergh, M., 2020, 104).

Conviene precisar a este respecto que la demanda de transparencia y de expli-

cabilidad no comportan necesariamente "revelar el código del software" ni tampoco exigir un análisis pormenorizado y exhaustivo de todo el proceso seguido. Y es que no se trata de incurrir en posiciones extremas. Ni por defecto ni por exceso, porque en la práctica, ambos planteamientos serían inviables. Es cuestión de explicar en términos razonables las decisiones tomadas por el algoritmo, lo cual no equivale necesariamente a mostrar el funcionamiento detallado y pormenorizado del sistema. Se trata más bien de rendir cuentas, y por tanto de de no ocultar su trazabilidad (Cotino, L., 2017), de exponer los pasos seguidos para llegar al resultado con transparencia. Y para ello es preciso conocer cómo logra la IA sus respectivas recomendaciones facilitando los motivos y criterios a partir de los cuáles alcanza finalmente esa decisión (Coekelbergh, M., 2020, 103).

En última instancia, podría decirse que los principios de transparencia y explicabilidad, se erigen en trasunto de la exigencia de recibir información precisa sobre el procedimiento, uso, fines y resultados del sistema de IA, con el fin de poder ejercer un control eficaz, a partir de evidencias y garantías concretas, para evitar los eventuales sesgos que pudieran derivarse de su aplicación. Cuestión en modo alguno baladí que adquiere cada vez mayor trascendencia social (Megías Quirós, J. J., 2022, 148).

De ese modo, el cumplimiento de estos principios permitirá evitar la presencia de sesgos discriminatorios que pueden irrumpir relacionados con la edad, usos lingüísticos, rasgos étnicos, discapacidad, etc. en los que puedan incurrir los algoritmos de los sistemas de IA, ya sea por reproducción de los existentes en la realidad social circundante o bien por la creación de otros de naturaleza diversa, que se introducen de forma consciente o inconscientemente en los programas (Megías Quirós, J. J., 2022, 147).

3. Opacidad y cajas negras

1. Conviene precisar que cuando hablamos de opacidad respecto a los algoritmos, estamos haciendo referencia a la falta de transparencia motivada por la existencia de una especie de caja negra que carece de capacidad explicativa y dificulta su correspondiente inteligibilidad (Felzmann, H., et al., 2020). De hecho, desconocemos por completo su funcionamiento, pero la experiencia permite constatar que cuando el algoritmo adopta una resolución, ésta se convierte en una especie de veredicto –concesión o denegación de préstamos, hipotecas, puestos de trabajo, ayudas sociales etc.,– ante el cual no es posible objetar ni recurrir, a pesar de que esa decisión haya podido alimentarse de datos parciales, eventualmente incorrectos o insuficientes, o que han podido ser malinterpretados, sin revelar nada a cambio. Por ejemplo, cuando un cliente solicita un préstamo a una entidad bancaria, los responsables de la gestión bancaria deberían explicar por qué motivos rechazan concederlo.

Y sin embargo, el peligro que se cierne a largo plazo, de seguir con esta dinámica, no radica solamente en la eventual manipulación y dominación de las élites tecnocráticas, que podrían aspirar a crear una sociedad dicotómica, escindida. El riesgo adicional y, quizás más profundo que late tras ese modo de proceder, es promover una sociedad altamente tecnológica, cada vez más tecnificada e instrumental, en la que "incluso las élites ignoran lo que es-

tán haciendo, y en las que nadie puede responder por lo que pasa" (Coekelbergh, M, 2020, 102).

No podemos obviar que el algoritmo que se alberga bajo el manto de esa caja negra y que se nutre copiosamente de nuestros datos más privados, de carácter personal, viene a ser como un "bastión inexpugnable protegido por las leyes de propiedad intelectual" (Alonso, J., 2018, 3). En esa relación asimétrica que se establece, la desigualdad entre las partes es manifiesta, y a medida que el perímetro de la privacidad e intimidad se ve cada vez más desprotegido y vulnerable, más angosto, la opacidad empresarial parece verse más reforzada.

De ahí la pertinencia, como propone de forma explícita López de Mántaras, (2019, 12) de dotar de significación y capacidad explicativa a los sistemas de aprendizaje profundo incorporando módulos que permitan explicar con claridad el proceso en virtud del cual los algoritmos han llegado a los resultados y conclusiones alcanzados. Ya que "la capacidad de explicación es una característica propia e irrenunciable "en cualquier sistema inteligente".

2. Por otra parte, es evidente que más allá de las propuestas y reivindicaciones legítimas que podamos plantear como ciudadanos, en realidad ante el desarrollo y aplicación de las tecnologías, es evidente que en las sociedades abiertas y globalizadas en las que el riesgo es inherente, en puridad, el riesgo cero, no existe propiamente (Beck, U., 2006). Ahora bien, aunque se trata de una cuestión no exenta de complejidad en el ámbito de las tecnologías emergentes, eso no obsta para que al mismo tiempo, en tanto que sociedad, debiéramos tratar de determinar y regular el umbral de riesgo que estamos dispuestos a asumir (Lecuona, 2020, 150). Porque, en ese contexto, ni la pasividad ni la dilación a la hora de crear y aplicar normas se convierten en los mejores aliados (Gil, E., 2011).

Y es que, más allá de las dificultades que comporta el aprendizaje automático profundo, podría decirse que existe, adicionalmente, un problema muy extendido respecto al conocimiento de la IA, en la medida en que, como advertíamos supra, son muchos los que participan y utilizan las aplicaciones de IA. Unos y otros desconocen realmente lo que hace la IA, con los efectos –imprevistos- subsiguientes. Sin embargo, esta actitud, no deja de ser un problema de responsabilidad, y, por tanto, deviene un problema que podríamos denominar éticamente serio, trascendente en primer lugar, con las implicaciones jurídicas subsiguientes.

Porque a partir del desarrollo de la IA y sus aplicaciones no examinamos el grado de corrección o de incorrección del comportamiento humano en la vida cotidiana y sus consecuencias, tal y como acontecía hasta hace décadas, antes de la irrupción de las nuevas tecnologías. Ahora el planteamiento es muy distinto. "Somos hijos de nuestras decisiones" insistía Miguel de Cervantes en las páginas de *Don Quijote de la Mancha,* pero con la irrupción de los algoritmos debemos analizar las decisiones que adoptan las máquinas autómatas en las que se han convertido los ordenadores programados por seres humanos (Alonso, J., 2018, 2). Y hemos de valorarlas con el fin de evitar el riesgo que conlleva esa opacidad así como el peligro de confiar demasiado en las innovaciones y promesas de la tecnología, teniendo en cuenta que los sesgos interfieren antes o después. De hecho, no dejan de impregnar la dinámica de la sociedad (Coekelbergh, M., 2020, 111).

4. Delegación y atribución de responsabilidad

1.A pesar de encontrarnos inmersos en una nueva era en la que la hegemonía progresiva de las tecnologías disruptivas es incontrovertible, sin embargo, tal y como advertía uno de los fundadores de la histórica Escuela de Frankfurt, Th. Adorno en *Dialéctica negativa* (1992) parece como si la única preocupación de la sociedad tecnológica fuese innovar y producir bienes, mercancías, servicios, mientras que las consideraciones éticas son, una y otra vez, ignoradas. No contaban entonces para nada. Tampoco ahora, podría decirse igualmente.

Tal y como acabamos de referir en los apartados anteriores, una parte de las decisiones que nos afectan cada día, surgen de una u otra forma, a partir de modelos matemáticos que son programados y entrenados para tomar resoluciones de forma autónoma. El problema que se deriva después es que cuando tratamos de especificar un "sujeto responsable" de la acción ejecutada nos vemos ante una máquina despersonalizada a la que no resulta fácil interpelar (Colmenarejo, 2018, 126).

Ante esa tesitura la pregunta no admite demora ¿cómo identificar al sujeto que debe asumir responsabilidad? ¿Cómo podemos atribuir y distribuir las responsabilidades? Nos referimos a la hora de determinar quién ha de hacerse cargo de la realidad y por ende de las consecuencias que se provocan con su decisión, independientemente de que se trate de la fase del proceso de generación, gestión o utilización posterior del conocimiento que se ha alcanzado en el tratamiento respectivo de los datos (Colmenarejo, 2018, 126).

Es evidente que las máquinas que intervienen pueden ser agentes instrumentales, no son pacientes, pero tampoco son precisamente agentes morales dado que carecen de intencionalidad, y están privadas de emociones, de valores. A este respecto, conviene recordar el legado de la cultura occidental aportado por Aristóteles (2014) cuando advertía de forma explícita en la *Ética a Nicómaco* que sólo los seres humanos son capaces de realizar acciones de forma voluntaria y ser conscientes de ellas. Máxima que, a pesar del tiempo transcurrido y los avances aportados por la tecnología digital, sigue todavía vigente. Por tanto es preciso asumir que hay que "hacer responsables a los humanos de lo que hace la máquina" Porque a pesar de la interacción que se produce entre las innovaciones tecnológicas y los seres humanos, una cosa es que los humanos deleguen la capacidad de actuar en las máquinas que han sido programadas, pero otra cosa distinta es que eludan la responsabilidad (Coekelberg, M., 2020, 97).

Por otra parte, el problema viene después a la hora de determinar cómo debería distribuirse la responsabilidad entre las diferentes partes involucradas en el proceso de toma de decisiones desde el momento en el que se comienza a seleccionar datos, se eligen unos y se discriminan otros, se crea un algoritmo, se busca un patrón, se toma la decisión, etc. (Alonso, J. 2018, 2). Si descendemos al terreno concreto, podemos pensar v.g. en un proyecto científico que ha sido elaborado en una universidad por un grupo de investigadores, que se prueba en el laboratorio de ese centro académico, después se aplica en el sector sanitario, y finalmente se extiende masivamente en un contexto militar ¿Quién es el responsable? ¿A quien se puede pedir rendimiento de cuentas?

Ya que, conviene precisar que los algoritmos de la IA son elaborados por numerosos especialistas que participan en el proceso. Después van incorporándose otras manos que se ven involucradas en su gestión y aplicación, lo cual dificulta la cuestión de atribución de la responsabilidad (Coekelberg, M., 2020, 98).

Por otra parte, no podemos obviar que, *sensu stricto*, la IA actual no es consciente de lo que hace, propiamente. De hecho, no tiene conciencia y por tanto no sabe lo que puede provocar. Es evidente que la IA está programada para recabar y registrar datos, combinarlos, procesarlos, pero en realidad "no sabe lo que hace" a diferencia de los seres humanos que tienen conciencia e intencionalidad, que razonan y reflexionan sobre las acciones que realizan y las consecuencias que se derivan de sus respectivas decisiones (Cortina, A., 2019, 385).

Por tanto, son los participantes en los diversos procesos quienes saben realmente cómo procede la IA, así como lo que ellos, diseñadores, programadores y aplicadores, llevan a cabo. Solo ellos pueden responder por sus acciones y explicar lo que han hecho. No obstante, a pesar de esta primera aproximación, si tratamos de ahondar y miramos con más perspectiva podría decirse que tampoco esta descripción se corresponde con la realidad que acontece con el manejo de la IA. Probablemente los programadores principales conozcan el código y sepan cómo funciona todo el proceso, o al menos tengan en mente la parte inicial que ellos han programado, pero eso no implica que otros programadores que adaptan después y desarrollan los algoritmos, o que utilizan los algoritmos para aplicaciones determinadas, comprendan y puedan explicar qué está haciendo la IA. (Coekelberg, M., 2020, 101)

Podría decirse entonces que los programadores saben qué quieren hacer con la IA, o mejor dicho, saben lo que quieren que la IA haga por ellos. Para lo cual crean algoritmos, asignan unos objetivos específicos y a continuación delegan en la IA la tarea subsiguiente que ésta ha de acometer (Gil, E., (2011). Pero en realidad, si hablamos en términos precisos no siempre saben concretamente qué tipo de actividad está haciendo la IA en cualquier momento del proceso, y por tanto cabe afirmar que "no siempre pueden explicar lo que hizo o cómo tomó su decisión" (Coekelberg, M., 2020, 100).

Por ejemplo, si nos referimos a las IA de aprendizaje automático, y más en concreto a las que se sirven del aprendizaje profundo de las redes neuronales, el principio de explicabilidad que posibilita el proceso de toma de decisiones no es posible. Eso significa que la transparencia desaparece y, por ende, no cabe aportar una explicación (Pasquale, F., 2015). En otras palabras, podría decirse que es posible saber cómo funciona el sistema, pero no es posible explicar una decisión en particular (Coekelberg, M., 2020, 100-101).

2. Por todo ello, cabe concluir que las tecnologías basadas en procesos de automatización plantean problemas específicos de atribución de responsabilidad, y en particular determinados tipos de IA ofrecen dificultades especiales debido al carácter hermético e inaccesible que conlleva la existencia de las cajas negras (Pasquale, F., 2015).

Es preciso también advertir que cuando se habla de la existencia de "sistemas autónomos" habría que utilizar otros términos. Tendríamos que decir que en realidad se trata de artefactos o "autómatas", que no es lo mismo. De hecho, los sistemas inteli-

gentes pueden ser operativos a la hora de resolver problemas determinados y actuar independientemente de los seres humanos, pero en realidad no son autónomos. No son capaces de establecer objetivos o metas concretas para alcanzar (López de Mántaras, R., 2021). Y es que los autómatas no pueden decidir por sí mismos qué deben hacer. De ahí que carezca de sentido dejar en manos de máquinas inteligentes, por innovadoras y sofisticadas que puedan ser sus múltiples aplicaciones, las decisiones que afectan a la vida de las personas, sin pasar por el tamiz previo de la supervisión humana. Ya sean referidas al ámbito de la seguridad ciudadana, salud, préstamos bancarios o en la relación de a IA con los ciudadanos en el seno de la administración pública.

La motivación es clara e inquívoca. Solo los individuos deben tomar la decisión última y dar razón de ella. Explicarla en caso necesario y tratar de justificarla. Solo ellos pueden ser responsables también en términos jurídicos, porque la responsabilidad exige contar con autonomía y estar dotado de capacidad de autodeterminación (Cortina, A., 2019, 389).

Otra realidad distinta podríamos encontrarnos si en un futuro, más o menos próximo, se creasen sistemas inteligentes con una inteligencia general como la que caracteriza a los humanos (López de Mántaras, R., 2019, p. 4). En ese caso, y solo en ese supuesto, habría que formularse preguntas explícitas tales como: ¿podríamos atribuirles autonomía y reconocerles responsabilidad por las decisiones que tomen? ¿Tendrían entonces derechos y deberes? ¿Deberían asumir las consecuencias que se deriven? De momento, podemos afirmar taxativamente que este supuesto tiene que ver poco con la realidad. Tal vez hablaríamos en términos de simulación: actuarían como si tuvieran intencionalidad, como si contasen con emociones, y también sensaciones y sentido común, pero en cualquier caso "no dejaría de ser una simulación" (Cortina, A., 2019, 385).

Bibliografía

Adorno, Th., (1992) *Dialéctica negativa,* Madrid, Ed. Taurus.

Alonso, J., (2018) "Ética y algoritmos: una combinación necesaria. La falsa imparcialidad de las máquinas", 16 de mayo 2018. https://telos.fundaciontelefonica.com/etica-algoritmos-una-combinacion-necesaria/ p. 2

Aristóteles, (2014) Ética a *Nicómaco,* Madrid, Alianza Editorial.

Beck, U., (2006) *La sociedad del riesgo: hacia una nueva modernidad,* Paidós Ibérica.

Blázquez Ruiz, F. J. (1999) *Derechos Humanos y Proyecto Genoma,* Granada, Ed. Comares.

Cervantes, M. de, (2020) *Don Quijote de la Mancha,* Madrid, Editorial Austral.

Coekelbergh, M., (2020) *IA Ethics*/Ética de la inteligencia artificial, Madrid, Ed. Cátedra**.**

Colmenarejo, Fernández, R., (2018) "Ética aplicada a la gestión de datos masivos", *Anales de la Catedra Francisco Suárez*, 52, pp. 113-129.

Cortina, A., (2019) "Ética de la inteligencia artificial" *Anales de la Real Academia de Ciencias Morales y Políticas*, LXXXI, n.96, pp. 379-394.

Cotino Hueso, L. (2017) "Big data e inteligencia artificial. Una aproximación a su tratamiento jurídico desde los derechos fundamentales" *Dilemata*, n. 24, pp. 131-150

Diéguez, A., (2017) *Transhumanismo. La búsqueda tecnológica del mejoramiento humano*, Barcelona, Ed. Herder.

Felzmann, H., Fosch-Villaronga, E., Lutz, Ch., Tamo-Larrieux, A., (2020) "Towards transparency by design for Artificial Intelligence", *Science and Engineering Ethics*, VL, 6, pp. 3333-3361.

Foucault, M., (2022) *Microfísica del poder,* Ed. Siglo XXI, Buenos Aires.

Gil, E., (2011) *Big data, privacidad y protección de datos*, Madrid, Agencia Española de protección de datos, BOE.

Horkheimer, H., (2010) *Crítica de la razón instrumental***,** Madrid, Ed. Trotta.

Kelsen, H., (2006) ¿Una nueva ciencia de la política?, Buenos Aires, Katz, 2006, p. 15.

Khun, Th., (2005) *La estructura de las revoluciones científicas,* FCE.

Lecuona, I. de, (2020) "Aspectos éticos, legales y sociales del uso de la inteligencia artificial y el big data en salud en un contexto de pandemia", *Revista internacional de pensamiento político***,** vol.15, 2020, pp. 139-16l.

López Baroni, M. J., (2019), "Las narrativas de la inteligencia artificial", *Revista de Bioética y Derecho***,** pp. 5-28.

López de Mántaras, R., (2019) "El futuro de la IA: hacia inteligencias artificiales realmente inteligentes" BBVA, 2019, https://www.bbvaopenmind.com.

López de Mántaras, R., (2021) "La inteligencia artificial nunca será como la humana" https://www.lavanguardia.com/ciencia/20210329/6607152/inteligencia-artificial-nunca-sera-humana.html

Llano, F. , (2018) *Homo excelsior. Los límites ético jurídicos del Transhumanismo,* Valencia, Tirant lo Blanch.

Megías Quirós, J. J., (2022) "Derechos humanos e inteligencia artificial" *DIKAIOSYNE*, n. 37, Enero, 2022, pp. 139-163.

Ortega y Gasset, J., (2019) *Meditaciones sobre la técnica,* Madrid, Alianza.

Orwell, J., (2007) *1984,* Madrid. Editorial Espasa.

Pasquale, F., (2015) *The black box society: the secret algorithms that control money and information,* Boston, Harvard University Press.

Shelley, M., (2020) *Frankenstein,* Madrid, Ed. Susaeta.

Sloterdijk, P. (2022) *Hacer hablar al cielo***,** Madrid, Ed. Siruela.

Zambrano, M., (1996) *Filosofía y poesía,* Madrid, Fondo de Cultura Económica.

Zweig, S., (2012) *El mundo de ayer. Memorias de un europeo,* Madrid, Acantilado.