ARTÍCULOS

Revista de **Filosofía**
ISSN: 0034-8244

https://dx.doi.org/10.5209/resf.87612

EDICIONES COMPLUTENSE

¿Es la seguridad moralmente relevante para la inteligencia artificial confiable? El valor de la dignidad humana en la sociedad tecnologizada[1]

Antonio Luis Terrones Rodríguez[2]

Recibido: 14 de marzo 2023 / Aceptado: 10 de julio de 2023

Resumen. El discurso sobre la seguridad que viene planteándose en el ámbito de la Inteligencia Artificial (IA) se caracteriza por un predominio de las perspectivas técnica y normativa. Los efectos ambivalentes de esta tecnología y su consideración como un sistema sociotécnico plantean la necesidad de complementar este discurso integrando una perspectiva moral. Así pues, el objetivo principal de este trabajo consiste en argumentar que la seguridad moral es un elemento indispensable para alcanzar una IA confiable, en virtud de las diversas situaciones donde la dignidad humana puede verse comprometida.
Palabras clave: ética; inteligencia artificial; confianza; seguridad; dignidad.

[en] Is safety morally relevant to trustworthy artificial intelligence: the value of human dignity in the technologized society

Abstract. The discourse on security that has been raised in the field of Artificial Intelligence (AI) is characterized by a predominance of technical and regulatory perspectives. The ambivalent effects of this technology and its consideration as a socio-technical system raise the need to complement this discourse by integrating a moral perspective. Thus, the main objective of this work is to argue that moral security is an essential element to achieve a trustworthy AI, due to the various situations where human dignity can be compromised.
Keywords: ethics; artificial intelligence; trust; security; dignity.

Cómo citar: Terrones Rodríguez, A.L. (2024): "¿Es la seguridad moralmente relevante para la inteligencia artificial confiable? El valor de la dignidad humana en la sociedad tecnologizada", en *Revista de Filosofía* 49 (2): 583-595.

[1] La presente investigación se integra en los resultados del proyecto "Ética cordial y democracia inclusiva en una sociedad tecnologizada" (ETICORDIAL), PID2022-139000OB-C22, financiado por MCIU/AEI/10.13039/501100011033/FEDER, UE

[2] Universidad de Valencia
antonioluis.terrones@gmail.com

1. Introducción

La presencia de la IA es cada vez más notoria en diversos espacios de la vida humana, debido a que constituye un apreciado recurso para suministrar importantes beneficios, aunque, también acarrea considerables riesgos que requieren ser pensados. International Data Corporation estima un incremento del gasto en IA que superará los 97.000 millones de dólares en 2023, dando lugar a aumento del 28,4% en el periodo 2018-2023 (Joshi, 2019), una información muy significativa. Este crecimiento responde a los altos niveles de complejidad que la automatización ha alcanzado y a una mayor disponibilidad de datos. Ahora bien, más allá de estas cifras, el desafío se encuentra en el planteamiento de cuestiones sustanciales para reflexionar sobre los profundos impactos que presenta la actividad de los intelectos sintéticos en la sociedad tecnologizada.

Los efectos ambivalentes del despliegue de la IA, así como la complejidad y opacidad de los sistemas, han motivado a la Comisión Europea (CE) para impulsar la constitución del Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial, con el propósito de sentar las bases de un ecosistema fundamentado en la confianza. En el documento *Directrices éticas para una IA fiable* (2019), este grupo de expertos señala un conjunto de directrices y principios éticos que procuran un desarrollo de la IA alineado con las necesidades y expectativas de la ciudadanía. En otro texto, también de la CE y titulado *Libro Blanco sobre inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza*, es posible apreciar un claro interés en la generación de un escenario que ponga en valor la confianza. Estas iniciativas europeas constituyen expresiones de interés para dibujar un horizonte ético en el terreno de la IA. Por consiguiente, resulta imprescindible explorar en este contexto qué elementos pueden fortalecer la confianza y, en consecuencia, precisan ser analizados. Este es el caso de la seguridad moral, que representa un planteamiento complementario a las formas tradicionales en las que se ha venido configurando la seguridad, principalmente normativas y técnicas.

El artículo 17 de la *Constitución Española* (CE) de 1978 señala que "toda persona tiene derecho a la libertad y a la seguridad. Nadie puede ser privado de su libertad, sino con la observancia de lo establecido en este artículo y en los casos y en la forma previstos en la ley" (CE, 1978). De un modo similar, en el artículo 6 de la *Carta de Derechos Fundamentales de la Unión Europea* se detalla que "toda persona tiene derecho a la libertad y a la seguridad" (UE, 2000). Adicionalmente, el artículo 3 de la *Declaración Universal de los Derechos Humano*s expresa que "todo individuo tiene derecho a la vida, a la libertad y a la seguridad de su persona" (ONU, 1948). Estas prescripciones ponen de relieve la importancia de la seguridad en el espacio normativo, identificándola como asunto esencial para el bienestar ciudadano. En cuanto al plano técnico, en *Directrices éticas para una IA fiable* son destacados un conjunto de requisitos claves en los que es posible apreciar una referencia a la seguridad en el apartado "Solidez técnica y seguridad". En esta sección la solidez técnica para la prevención del daño, la resistencia a los ataques, los salvaguardias para un plan de repliegue en caso de que surjan problemas, la precisión, la fiabilidad y la reproducibilidad constituyen, a juicio de los expertos, elementos indispensables para garantizar la seguridad en los sistemas artificiales (2019: 20-21). Como puede constatarse, la seguridad ha sido y es objeto de discusión normativa y técnica. Así pues, en virtud de seguir enriqueciendo este debate, es necesario llevar a cabo un

ejercicio de contextualización en el seno de la sociedad tecnologizada y dirigir la atención a los impactos de la IA en aras de la confianza.

A diferencia de continuar la senda de los estudios que analizan los detonantes para el surgimiento de la confianza en la interacción humano-tecnología (Papenkordt y Thommes, 2022), en virtud de la concepción socioténica de la IA asumida, se examinará el valor de la seguridad moral como un determinante esencial para la generación de un entorno con predominio de la confianza. A este respecto, el objetivo principal de este trabajo consiste en demostrar que para sentar las bases de una IA confiable debe garantizarse, previamente, un entorno seguro, en términos morales, para los sujetos afectados por un determinado sistema. En primer lugar, se subrayará el carácter relacional de la experiencia humana de la confianza, puntualizando su condición derivada en el contexto de las tecnologías. En segundo lugar, se llevará a cabo una aproximación a la seguridad moral, advirtiendo que la ideología del solucionismo tecnológico supone un obstáculo para su consideración. Finalmente, se pondrá en valor la dignidad, dado que es un elemento esencial de la condición humana que solicita ser reconocido, estimado y cuidado en vista de los diversos impactos de la IA.

2. Confianza y experiencia relacional

La justificación de la seguridad moral como un elemento determinante para el surgimiento de la confianza precisa un análisis previo sobre algunos elementos relacionales que resultan de interés. La consolidación de la presencia de los intelectos en numerosos espacios y actividades conlleva un incremento de la interdependencia respecto a esta tecnología. Por esta razón, la confianza surge como un elemento indispensable en las relaciones de los sistemas sociotécnicos. Estas relaciones involucran a dos partes específicas: una parte que confía (fideicomitente); y otra parte en la que se debe confiar (fideicomisario). Generalmente, estas partes se identifican con agentes humanos, aunque, esta identidad resulta más compleja en la sociedad tecnologizada, en virtud de la omnipresencia de sistemas sociotécnicos, como es el caso de la IA. Si bien existe una falta de diferenciación clara entre los factores que contribuyen a la confianza, la presencia de riesgo e incertidumbre es un elemento determinante por el que un fideicomitente, en vista de su vulnerabilidad, confiaría en un fideicomisario (Boon & Holmes, 1991; Das & Teng, 2004). En tal sentido, Isabel Thielmann y Benjamin E. Hilbig (2015) señalan que las actitudes hacia las perspectivas de riesgo es uno de los tres determinantes centrales para el surgimiento de la confianza, junto a las expectativas de confiabilidad y la sensibilidad a la traición. En todo caso, valorando la diversidad de enfoques existentes en torno a la conceptualización de la confianza, esta podría definirse como "la voluntad de una parte de ser vulnerable a las acciones de otra parte con base en la expectativa de que la otra realizará una acción particular importante para el fideicomitente, independientemente de la capacidad de monitorear o controlar esa otra parte" (Mayer et al., 1995: 712).

Conviene subrayar que el ser humano es, en términos ontológicos, un ser vulnerable, susceptible de sufrimiento y fragilidad, como han apuntado algunas figuras destacadas (MacIntyre, 2001; Nussbaum, 2006; Turner, 2006; Ricoeur, 2008; Butler, 2008; 2021). Además, desde la óptica de la perspectiva relacional asumida en

este trabajo, es preciso reconocer que la vulnerabilidad aflora tras la experimentación de daños en los propios intereses, causados por actores o actividades que comportan un riesgo o amenaza (Goodin, 1985). Así pues, podría afirmarse que la vulnerabilidad se encuentra estrechamente vinculada a la confianza sobre la base de una expectativa de seguridad para la eliminación o mitigación de los riesgos asociados a determinadas relaciones de un contexto.

La discusión llevada a cabo por S. D. Noam Cook (2010) en torno al concepto de confianza presentado por Joseph C. Pitt (2010), pone de manifiesto la complejidad de esta idea. Pitt señala que la confianza se erige a partir de una promesa establecida mediante una comunicación lingüística. En su explicación, este filósofo norteamericano afirma que una promesa implica cognición en el terreno de las relaciones sociales, de tal forma que la confianza es un asunto estrictamente humano que excluye a los animales. Por este motivo, la confianza es fundamentalmente interpersonal, ya que se sostiene en promesas, como afirma Pitt (2010: 448). Adicionalmente, cuestiona si es posible confiar en la tecnología, afirmando que, en vista de que esta es únicamente un medio empleado por los humanos, no es posible. En conclusión, Pitt entiende que la confianza es asunto exclusivo de los humanos y que el desarrollo de un proceso comunicativo y cognitivo es una condición *sine que non* para su aparición.

Cook aborda esta cuestión optando por un sentido diferente, sin centrar su atención en el significado del término, sino en la experiencia que está presente en aquellas relaciones que implican expectativas mutuas (2010: 455). Recurre a varios ejemplos para evidenciar que uno de los pilares que sostienen las relaciones son las expectativas que las personas tienen entre sí. El cumplimiento de estas expectativas es un requisito fundamental para que la experiencia humana pueda celebrarse en confianza. A diferencia de Pitt, que identifica el lenguaje como un aspecto indispensable de la confianza, Cook pone el énfasis en la práctica. Pese a reconocer el peso del componente lingüístico, el motivo que lo lleva a distanciarse de Pitt es la apreciación de la promesa como un elemento resultante de la confianza y no al revés. De acuerdo con esta secuencia, Cook construye una perspectiva de la confianza a partir de la práctica en el seno de la experiencia y no exclusivamente desde el lenguaje y la cognición. Ambas posiciones insisten en que la tecnología es un asunto humano, en atención a que la relación establecida se encuentra condicionada por múltiples factores ubicados en circunstancias y contextos particulares, así como en un entramado de relaciones.

Reconocer que la tecnología es un asunto humano implica concebir la IA como un sistema sociotécnico, ya que sus crecientes desafíos exigen admitir que integra componentes técnicos y elementos sociales (Sartori y Theodorou, 2022). Recogiendo el testigo de Pitt y Cook, no es consecuente afirmar que un determinado artefacto, como tal, es depositario de confianza. En todo caso, al estar integrado en una red de relaciones en las que diseñadores, fabricantes, distribuidores, gestores, controladores, etc., juegan un papel fundamental, lo más lógico es dirigir la confianza hacia el sistema en su conjunto (Comisión Europea, 2019: 6). Esta consideración supone una premisa muy importante para este trabajo, a saber, que para pensar los condicionantes de la confianza es necesario atender a la experiencia relacional. A modo de ejemplo, el siguiente episodio pone de manifiesto cómo la confianza en la IA depende de elementos sistémicos. Supongan que un sector de la ciudadanía con escasos recursos solicita una ayuda a la Administración pública y ve desestimada esta solicitud, como resultado de una mala gestión. Este fue el caso de las personas

afectadas por la Gobernación de Indiana a finales de 2006.. Esta Administración pública firmó un contrato con un consorcio de empresas, entre las figuraba IBM, para la automatización y privatización de los procesos de elegibilidad del servicio de asistencia social. Virginia Eubanks (2018) relató cómo los resultados no fueron los esperados y la rigidez del sistema informático no diferenció entre errores honestos, errores burocráticos e intentos de fraude, deduciendo que en todos los casos estaba cometiéndose un posible delito o falta. Este episodio indica que la desconfianza experimentada por la ciudadanía solicitante de ayuda no es proyectada a la IA, en sí misma, sino a la Administración pública, gestionada por agentes humanos a través de un entramado de relaciones. Por lo tanto, es importante subrayar que la tecnología es objeto de una confianza derivada de las relaciones interpersonales de los agentes que dan forma a un sistema, puesto que no es posible afirmar que un intelecto sintético tiene la capacidad de responder a las consideraciones morales de un modo similar al ser humano.

Philip J. Nickel *et al*. (2010) han investigado la particularidad de la confianza y su impacto en la tecnología, señalando que este elemento está presente, aunque en un sentido derivado. La concepción sociotécnica de la IA insiste en que esta tecnología se halla integrada en una red de promesas en relación al funcionamiento y los resultados, dando lugar a que los objetos depositarios de confianza no sean los artefactos, sino los agentes humanos involucrados en su despliegue. Retomando el ejemplo anterior, la ciudadanía que precisa ayuda no desconfía del sistema artificial encargado de la gestión, sino de la Gobernación de Indiana, como una institución política que no ha atendido su llamado y, por consiguiente, satisfecho sus necesidades. Así pues, es comprensible que las personas afectadas perciban este suceso como una falta de garantías a la protección de sus derechos y su reacción consista en la desafección política y el malestar hacia las instituciones responsables.

El caso de la Gobernación de Indiana pone de relieve que la reflexión sobre la confianza en el terreno de la IA exige atender a la experiencia del entramado relacional que configura la actividad tecnológica. Asimismo, la perspectiva sociotécnica nos enseña que el incremento de la complejidad de la automatización da como resultado sistemas híbridos, que en parte son técnicos y en parte sociales. En consecuencia, la seguridad adquiere su estatus moral en un horizonte de relaciones en el que la experiencia del reconocimiento, la estimación y el cuidado de elementos humanos como la dignidad constituye una acción vital que motiva la aparición de la confianza.

3. La dimensión moral de la seguridad

Como fue señalado en la introducción, el concepto de seguridad habitualmente manejado en el entorno de los sistemas artificiales se circunscribe a coordenadas estrictamente técnicas y normativas. En vista de las limitaciones de esta circunscripción para manejar los desafíos de la IA, es conveniente proporcionar e integrar una dimensión alternativa de la seguridad que está presente en la tecnología, debido a su naturaleza política (Winner, 1980). Esta dimensión sugiere que el discurso de la confianza que ha venido planteándose en los últimos años estimula una reflexión más ambiciosa sobre la seguridad que debe comenzar a formularse desde un plano moral, en un momento en el que la IA despliega todo su potencial en numerosos espacios de la vida humana. ¿Es posible hablar de una IA confiable sin

insistir en el valor de la seguridad moral? ¿Son seguras las decisiones de un sistema caracterizado por la opacidad? ¿Hasta qué punto transmite seguridad y, por tanto, es apropiado confiar en un sistema artificial que puede pasar por alto una determinada condición sexual o étnica que merece especial reconocimiento y cuidado? Estas son algunas cuestiones problemáticas que apuntan a la seguridad como un elemento esencial de la moralidad que demanda ser analizado, con el objeto de esclarecer su posible vínculo con la confianza.

El concepto de IA confiable impulsado por el Grupo de Expertos de Alto Nivel propone un desarrollo humanocéntrico a través de la alineación con las necesidades y expectativas de la ciudadanía en aras de un beneficio social y respeto a los derechos fundamentales. Para hacer posible la realización de esta visión humanocéntrica es primordial la incorporación de elementos morales, con el interés de enriquecer las estrategias técnicas y normativas. En razón de esta idea y la naturaleza relacional de la confianza, se requiere un acercamiento a la seguridad sobre la base del terreno de la moral. Esta nueva dirección responde a la necesidad de reconocer, estimar y cuidar aspectos humanos que pueden verse comprometidos en toda experiencia, en este caso tecnológica. Así pues, con la finalidad de seguir profundizando en el valor de la dimensión moral de la seguridad, a continuación, serán analizados algunos elementos importantes.

Jessica Wolfendale (2017) distingue dos modalidades de seguridad moral. En primer lugar, considera que la seguridad moral es subjetiva cuando existe un reconocimiento de la posición, las necesidades y las capacidades de una persona, considerado como una condición primordial para asegurar el cuidado de sus intereses y bienestar, que posibilita el surgimiento de un sentimiento moral. Mientras que, en segundo lugar, desde el lado objetivo, la seguridad moral podría ser entendida como un reconocimiento llevado a cabo en una comunidad política donde la moralidad obtiene su estatus a partir de la estimación y el cuidado que personas e instituciones practican. En el caso particular de esta perspectiva objetiva, puede ampliarse el horizonte de estimación y cuidado mediante la inclusión de sistemas sociotécnicos como la IA. En ambos casos, es importante vigilar que la seguridad moral discurre por la senda de la experiencia de las relaciones humanas. Por consiguiente, la articulación de la propuesta europea de una IA confiable debería plantearse a través de un proyecto que integre y ponga en valor la seguridad desde un punto de vista moral.

Wolfendale recoge el testigo de Axel Honneth e identifica tres categorías morales que describen formas de daño que no es posible atribuir exclusivamente al terreno físico: insulto, humillación y falta de respeto. Estas categorías morales pueden constituir una injusticia, riesgo o amenaza para las personas, fruto de una lesión que impacta negativamente en la comprensión positiva que un sujeto tiene de sí mismo y que ha sido asumida de un modo relacional e intersubjetivo (Honneth, 1995: 131). Además, es preciso admitir que todas las personas tienen un autoconcepto estrechamente vinculado a la autoestima, por esta razón, un mal experimentado por una persona puede entrañar un daño moral y, por tanto, un riesgo para su seguridad. El concepto de daño moral no es nuevo, pues el análisis de esta experiencia se remonta al trabajo de Jonathan Shay (1995). Una experiencia que combina la estructura de la identidad que un sujeto tiene de sí mismo, la dimensión emocional y la evaluación cognitiva (Kvitsiani et al., 2023). A este respecto, conviene recordar que el ser humano es un ser vulnerable para el que la seguridad es un factor determinante

de su bienestar. Realizada esta aclaración, es importante puntualizar la influencia de un sesgo tecnológico que puede dificultar el reconocimiento de estas categorías morales que escapan a los límites de lo estrictamente físico. La sobreestimación de la IA constituye una respuesta que condiciona y modula la relación cognitiva, epistemológica y moral con el mundo.

La sobreestimación de los sistemas artificiales responde a una visión teocrática que estimula el surgimiento de un sentimiento similar al de la fe religiosa en la Divina Providencia. En este caso, la fe se manifiesta en una ilusión de control computacional extendido y en la gestación de un concepto de superioridad técnica que desemboca en una relación de amparo epistémico e incluso moral (Bogost, 2015). La identificación de este obstáculo teocrático para la apreciación de la dimensión moral de la seguridad no implica el rechazo de la tecnología, sino la invitación a una crítica a cierta mentalidad que impregna la sociedad tecnologizada y promueve la simplificación de la realidad humana, reduciéndola a aspectos exclusivamente cuantitativos y computables.

En *La locura del solucionismo tecnológico*, Evgeny Morovoz articula una crítica sobre la ideología del solucionismo tecnológico, un ideario que, a su parecer, legitima y ratifica la obsesión por la optimización y el perfeccionamiento llevado a cabo a través de vías tecnológicas (2015:24). Para el pensador bielorruso, esta ideología ensalza ciegamente la tecnología mediante la voluntad constante de mejora y el diagnóstico simplificado y desafortunado de los problemas. El solucionismo suministra una visión muy acotada y superficial de problemas que demandan una comprensión serena, detallada y, por ende, compleja. En tal sentido,

> la sobreestimación de los sistemas artificiales puede suponer un obstáculo para la seguridad moral, en vista de la dificultad para reconocer y, posteriormente, estimar y cuidar aspectos vinculados a la dignidad humana que no se circunscriben necesariamente al terreno físico ni son gestionados exclusivamente con una visión técnica de la realidad que acarrea un empobrecimiento (Pigem, 2018: 143) Ciertamente, el despliegue de la IA en numerosos espacios constituye un problema moralmente relevante. Por esta razón, la insistencia en la dimensión moral de la seguridad representa un marcador para la evaluación de los impactos de esta tecnología, así como para la mitigación de sus riesgos y amenazas. La expansión del solucionismo tecnológico, como resultado del espíritu mercantil de Silicon Valley (Gumbrecht, 2020), se ha fundamentado en un discurso público sustancialmente vacuo respecto a la evaluación moral y ha subestimado el valor de la deliberación como una herramienta indispensable para el progreso.

En conclusión, ante el tsunami de las tecnologías disruptivas, entre las que la IA asume el protagonismo, es apremiante abrazar la oportunidad de experimentar el asombro y sentir respeto por aquello que nos constituye y nos hace ser lo que somos como humanos, un aspecto esencial de nuestra condición que puede verse comprometido en un tiempo en el que todo es reducido a una base material desontologizada y convertida en presa de la datificación.

4. La dignidad, un referente indispensable

La historia nos enseña cómo numerosas muestras de experiencias de injusticia, olvido y silencio se fueron convirtiendo en reivindicaciones morales que, en muchos casos, se tradujeron en luchas por intereses. Al fin y al cabo, estas reivindicaciones aspiran al reconocimiento de diversas naturalezas, señalando un referente indispensable e irrenunciable (Marina y de la Válgona, 2000). Este es el caso de la dignidad humana, un aspecto inseparable de la dimensión ética y del que conviene recordar su valor para estrechar una sana y responsable relación con el mundo (Pigem, 2018: 158), especialmente, una relación que es mediada e instrumentalizada por la tecnología más avanzada, la IA. Para este propósito resulta fundamental poner en tela de juicio la visión materialista que infunde un pensamiento contraproducente para nuestra vida, a saber, que la realidad es ajena a nosotros mismos, objetiva e independiente y, por tanto, susceptible de ser dominada. Esta visión constituye el principio originario que motiva la ideología solucionista e impregna el desarrollo de los intelectos sintéticos a través de seductoras estrategias que muestran únicamente los brillos de la tecnología. Pero como un ser paradójico que combina seguridad y ampliación de posibilidades, el ser humano está llamado a asumir responsabilidad de un modo activo en la custodia de un aspecto inseparable de su condición, que exige ser reconocido, estimado y cuidado (Jonas, 2004).

Para demostrar la importancia de la profundización en la dimensión moral de la seguridad, es importante ubicar el discurso en torno a un concepto fundamental, considerado a su vez un asunto vital. El concepto de dignidad goza de un amplio reconocimiento moral y jurídico, sin embargo, aún es objeto de discusión en torno a la construcción de su sentido universal, su finalidad normativa y práctica, así como su amplitud, entre otras consideraciones (Trueba Atienza y Pérez Cortés, 2018: 7-9). Ciertamente, la cantidad de fuentes y nociones de las que bebe este concepto son tan variadas que es conveniente concretar unas coordenadas que resulten lo suficientemente provechosas para pensar el impacto de la IA.

El término *dignitas*, como una condición o característica inherente de los seres humanos, comenzó a cobrar sentido en determinados textos latinos que destacan por la influencia del estoicismo. No obstante, es importante advertir que, como señala Diego Gracia, el vínculo que a menudo suele establecerse entre el estoicismo y ciertos textos latinos, especialmente algunos pasajes de Cicerón, responde a un sesgo hermenéutico que arroja una mirada fundamentalmente moderna (2008: 20). Posteriormente la conceptualización de la dignidad fue dibujada en la Edad Media sobre la base de una perspectiva creacionista; en el Renacimiento, vinculada a la libertad en la obra de Giovanni Pico della Mirandola (2004); y en la Modernidad, durante el siglo XVIII, en especial, a partir de *Fundamentación de la metafísica de las costumbres*, una de las obras más destacables de Immanuel Kant. Recogiendo el testigo de la madurez alcanzada por la idea de dignidad, en tanto que condición intrínseca del ser humano consolidada en el contexto cultural de la Ilustración, el de Königsberg inserta una naturaleza metafísica en el concepto, tejiendo un hilo conductor entre la racionalidad, la libertad y la moralidad (Durán Casas, 2018). De este modo, Kant dibuja una conceptualización dotada de contenido moral y ontológico. Cuando el filósofo de Königsberg se refiere a la humanidad como un fin en sí mismo, a la par, señala un principio práctico supremo y un imperativo categórico que debe impregnar todo acto de la voluntad humana:

La humanidad misma es una dignidad; porque el hombre no puede ser utilizado únicamente como medio por ningún hombre (ni por otros, ni siquiera por sí mismo), sino siempre a la vez como fin, y en esto consiste precisamente su dignidad (la personalidad), en virtud de la cual se eleva sobre todos los demás seres del mundo que no son hombres y sí que pueden utilizarse, por consiguiente, se eleva sobre todas las cosas. Así pues, de igual modo no puede autoenajenarse por ningún precio (lo cual se opondría al deber de la autoestima), tampoco puede obrar en contra de la autoestima de los demás como hombres, que es igualmente necesaria; es decir, que está obligado a reconocer prácticamente la dignidad de la humanidad en todos los demás hombres con lo cual reside en él un debe que se refiere al respeto que se ha de profesar necesariamente a cualquier otro hombre (Kant, 2008: 335-336).

Esta invención ética de Kant constituye un signo de progreso moral a lo largo de historia y sirve para constatar el dinamismo humanizador de la especie humana. Asimismo, ofrece un criterio de justificación para consagrar la dignidad humana como un referente indispensable para el motor de la innovación en la sociedad tecnologizada. Varios sucesos en los que el detonante ha sido el impacto de la actividad de la IA evidencian como la dignidad humana puede verse comprometida en función de un dogmatismo computacional que complica la seguridad, en términos morales.

El trabajo de Stephen Cave y Kanta Dihal (2020) dirige sus esfuerzos a la blanquitud predominante en las diversas manifestaciones de IA recogidas en robots humanoides, chatbots, asistentes virtuales, imágenes de archivo y representaciones en el cine y televisión. Cave y Dihal subrayan la racialización que experimentan las máquinas, como resultado de los atributos que les pueden ser otorgados a los artefactos y que, posteriormente, dan lugar a su identificación con determinadas categorías racionales humanas. A la habitual antropomorfización que ya experimentan los sistemas artificiales, este trabajo puntualiza que habría que agregar un componente racial expresado en la blanquitud. El robot humanoide Sofía, el avatar femenino Evie, el largometraje Ex Machina y el gran número de imágenes de archivo suministradas por el motor de búsquedas de Google, vienen a confirmar lo señalado por Cave y Dihal. Esta racialización tiene un alcance considerable en el que es adecuado insistir. Según los investigadores, en primer lugar, amplifica una serie de sesgos que existen previamente y podrían contribuir negativamente a una injusticia social; en según lugar, no contribuiría únicamente a la injusticia, sino que la exacerbaría, trasladando la jerarquía de poder ya existente en la sociedad a la esfera de la automatización de las decisiones, motivando preocupantes consecuencias; y, en tercer lugar, podría distorsionar la percepción de los verdaderos riesgos y beneficios de la IA.

Han pasado casi cuarenta años de la publicación del *Manifiesto Cyborg* de Donna Haraway (2020), un texto en el que la profesora estadounidense insistió en la necesidad de reorientar el pensamiento feminista para alcanzar mayores cuotas de compromiso con la ciencia y la tecnología, promoviendo, de ese modo, nuevos recursos para la emancipación y la crítica feminista. En la actualidad, algunas actividades de la IA impactan negativamente sobre el colectivo de las mujeres, propiciando diversas situaciones y estructuras caracterizadas por la injusticia. Entre los nuevos e imprevistos desafíos que han traído consigo los sistemas artificiales es posible destacar sesgos resultantes de la recopilación de datos demográficos que contienen ciertos aspectos de la experiencia humana que constituyen una falta de reconocimiento hacia otros datos que son ignorados o infrarrepresentados, como

consecuencia de una falta de sensibilidad (Latorre Ruiz y Pérez Sedeño, 2023: 61). El procesamiento de lenguaje natural (NLP, siglas en inglés *de natural lenguaje procesing*) es una de las áreas de la IA más afectadas por el empleo de datos sesgados. Muchas de las actividades llevadas a cabo en las sociedades tecnologizadas dependen en gran medida de sistemas y aplicaciones que utilizan este tipo de procesamiento. Una de estas actividades es la traducción automática, donde es posible observar la influencia negativa de los sesgos. Como señala Stefanie Ullmann, a pesar del reflejo adecuado de la demografía que proyectan algunos datos, la mayoría representan de una manera incorrecta las distribuciones reales en la sociedad, una estrategia que puede ocasionar graves problemas, principalmente desventajas y desigualdades en contextos específicos que afectan a las personas, principalmente mujeres (2022: 127). Asimismo, también es posible apreciar el impacto negativo de la IA en el colectivo de las mujeres negras del Sur global, despojadas de los derechos de privacidad y protección de datos, a raíz de herramientas basadas en IA para el reclutamiento y la explotación laboral con bajos salarios y escasa cualificación (Schelenz, 2022: 230).

Otro caso que merece una mención especial es COMPAS (siglas en inglés de *Correctional Offender Management Profiling for Alternative Sanctions*), un algoritmo utilizado durante los últimos años en algunos tribunales de Nueva York, Wisconsin, California y Florida, entre otras administraciones de Estados Unidos, para ofrecer asesoramiento sobre el ingreso en prisión o la puesta en libertad de los ciudadanos y ciudadanas acusadas en un juicio. La Administración judicial de estos estados contrató los servicios de la empresa Northpointe, propietaria de COMPAS, con el propósito de encontrar un soporte para mitigar la intuición y el sesgo humano mediante la asignación de un valor numérico vinculado a la probabilidad de riesgo. Movidos por la teocracia computacional, los contratantes de este servicio creen que la IA posee siempre un valor epistémico superior a la subjetividad humana, sesgada por valores e intereses morales, a diferencia de la tecnología que suministra objetividad mecánica (Daston y Galison, 1992). En este caso es posible apreciar como la sobreestimación de los sistemas artificiales ha originado un problema de control y motivado la polémica en torno a un posible sesgo racista. Tras la investigación llevada a cabo por la agencia de noticias ProPublica, las decisiones judiciales que contaron con el soporte de COMPAS hicieron saltar todas las alarmas, pues fue posible observar como a las personas de tez oscura les fue asignada una puntuación de riesgo superior respecto a las personas de piel blanca, presentando el doble de probabilidad e incurriendo, de este modo, en un sesgo racista (Angwin *et al*., 2016).

Estos casos señalan cómo la teocracia computacional y el solucionismo tecnológico logran motivar una falta de reconocimiento y descuido de elementos humanos que demandan especial consideración, de modo que la IA puede poner en riesgo la dignidad de determinadas personas o colectivos. Estas experiencias dolorosas dificultan el cultivo de la confianza, dado que no facilitan la generación de un ambiente seguro para los afectados por la tecnología. En ese sentido, es fácil apreciar la existencia de un estrecho vínculo entre la seguridad moral y la confianza en el despliegue de los intelectos sintéticos a través de sus diversos impactos. Por tanto, es razonable asumir que la experiencia tecnológica puede ser dolorosa y que es preciso tomar en serio este dolor para fundamentar la invención ética (Marina y de la Válgona, 2000: 29).

5. Conclusión

El acelerado despliegue de la IA impulsado por el mercado tecnológico sigue su curso imparable. El solucionismo tecnológico, la teocracia computacional y la datificación se expanden y ocupan numerosos procesos y espacios de la vida humana. Ante esta situación, toca armarse de valor y afrontar el presente con responsabilidad para dibujar un futuro cuidadoso. La crítica proyectada en este trabajo no representa una enmienda a la totalidad para rechazar al completo el desarrollo tecnológico. Por el contrario, supone una invitación para apropiarse teórica y prácticamente de los propósitos de la IA y aspirar a un escenario alternativo, que no instituya un empobrecimiento de la experiencia humana, donde elementos constitutivos fundamentales sean menospreciados u olvidados.

En las páginas que anteceden se ha subrayado la importancia que encarna la defensa de la dimensión moral de la seguridad para hacer posible una IA confiable. Sin seguridad no es posible la confianza, esta ha sido una premisa fundamental asumida en las presentes líneas. Sin embargo, es preciso insistir en que, al margen de las nociones técnicas y normativas de la seguridad, existen caminos diferentes, en unos casos y complementarios en otros, para atender el llamado de las iniciativas morales que simbolizan diversas formas de lucha por la dignidad y que, en algunos casos, pueden ser doloras.

6. Referencias bibliográficas

Angwin, J., Larson, J., Mattu, S. y Kirchner, L. (2016): "Machine bias", *ProPublica*. Disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bogost, I. (2015): "The Cathedral of Computation", *The Atlantic*. Disponible en: https://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/

Boon, S. D. & Holmes, J. G. (1991): "The dynamics of interpersonal trust: Resolving uncertainty in the face of risk", en R. A. Hinde & J. Groebel (eds.), Cooperation and prosocial behavior, Cambridge, Cambridge University Press, pp. 190-211.

Butler, J. (2008): *Vida precaria. El poder del duelo y la violencia*, Buenos Aires, Paidós.

Butler, J. (2021): *La fuerza de la no violencia*, Barcelona, Paidós.

Cave, S. & Dihal, K. (2020): "The Whiteness of AI", *Philosophy & Technology*, 33, pp. 685-703.

Comisión Europea (2019): *Directrices éticas para una IA fiable*. Disponible en: https://op.europa.eu/es/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Comisión Europea (2020): *Libro Blanco sobre inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza.* Disponible en: https://op.europa.eu/es/publication-detail/-/publication/ac957f13-53c6-11ea-aece-01aa75ed71a1

Constitución Española. Boletín Oficial del Estado, 29 de diciembre de 1978, núm. 311, 29313 a 29424.

Cook, S. D. N. (2010): "Making the Technological Trustworthy", *Knowledge, Technology & Policy,* 23, pp. 455–459.

Das, T. K. & Teng, B. –S. (2004): "The risk-based view of trust: A conceptual framework", *Journal of Business and Psychology*, 19, pp. 85-116.

Daston, L. & Galison, P. (1992), "The Image of Objectivity", *Representations*, 40, pp. 81-128.

Durán Casas, V. (2018): "La dignidad humana en Kant: una tarea, no un privilegio", en C. Trueba Atienza y S. Pérez Cortés (eds.), *Dignidad: perspectivas y aportaciones de la filosofía moral y la filosofía política*, Barcelona, Anthropos Editorial, pp. 175-201.

Goodin, R. E. (1985): *Protecting the Vulnerable: A Reanalysis of Our Social Responsibilities*, Chicago, University of Chicago Press.

Gracia, D. (2008): "¿Es la dignidad un concepto inútil?", en T. Ausín y R. Aramayo (eds.), *Interdependencia. Del bienestar a la dignidad*, Madrid, Plaza y Valdés, pp. 17-35.

Gumbrecht, H. U. (2020): *El espíritu del mundo en Silicon Valley*, Barecelona, Deusto.

Haraway, D. (2020): *Manifiesto Ciborg*, Madrid, Kaótica Libros.

Jonas, H. (2004): *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*, Barcelona, Herder Editorial.

Joshi, N. (2019), "How we can build Trustworthy AI", *Forbes.* Disponible en:
https://www.forbes.com/sites/cognitiveworld/2019/07/30/how-we-can-build-trustworthy-ai/

Kant, I. (2008): *La Metafísica de las Costumbres*. Madrid, Tecnos.

Kvitsiani, M., Mestvirishvili, M., Martskvishvili, K. Odilavadze, M. (2023): "Dynamic model of moral injury", *European Journal of Trauma & Dissociation*, 7(1), pp. 1-8.

Latorre Ruiz, E. y Pérez Sedeño, E. (2023): "Gender Bias in Artificial Intelligence", en J. Vallverdú, J. (eds.), *Gender in AI and Robotics. Intelligent Systems Reference Library*, Cham: Springer, Cham, pp. 61-75.

MacIntyre, A. (2001): *Animales racionales y dependientes: por qué los seres humanos necesitamos las virtudes*, Barcelona, Paidós.

Marina, J. A. y de la Válgona, M. (2000): *La lucha por la dignidad. Teoría de la felicidad política*, Barcelona, Editorial Anagrama.

Mayer, R. C., Davis, J. H. & Schoorman, F. D. (1995): "An Integrative Model of Organizational Trust", *The Academy of Management Review*, 20(3), pp. 709-734.

Morozov, E. (2016): *La locura del solucionismo tecnológico*, Buenos Aires, Katz Editores.

Nickel, P. J., Franssen, M. & Kroes, P. (2010): "Can We Make Sense of the Notion of Trustworthy Technology?, Knowledge, Technology & Policy, 23, pp. 429-444.

Nussbaum, M. (2006): *Las fronteras de la justicia: consideraciones sobre la exclusión*, Barcelona, Paidós.

ONU-Organización de Naciones Unidas (1948), *Declaración Universal de Derechos Humanos*. Disponible en: https://www.un.org/es/about-us/universal-declaration-of-human-rights

Papenkordt, J. & Thommes, K. (2023): "Determinants of Trust in Smart Technologies". En C. Röcker & S. Büttner (eds.), *Human-Technology Interaction.* Cham, Springer; pp. 335-359.

Pico della Mirandola, G. (2004): *Discurso sobre la dignidad del hombre*, México D. F., Universidad Nacional Autónoma de México.

Pigem, J. (2018), Ángeles o robots. La interioridad humana en la sociedad hipertecnológica, Barcelona, Fragmenta Editorial.

Pitt, J. C. (2010): "It's Not About Technology", *Knowledge, Technology & Policy,* 23, pp. 445–454.

Ricoeur, P. (2008): "Autonomía y vulnerabilidad", en P. Ricoeur (ed.), *Lo justo 2: Estudios, lecturas y ejercicios de* ética aplicada, Madrid, Trotta, pp. 70-86.

Sartori, L. & Theodorou, A. (2022): "A sociotechnical perspective for the future of AI: narratives, inequalities, and human control", Ethics and Information Technology, 24(4), pp. 1-11.

Schelenz, L. (2022): "Artificial Intelligence Between Oppression and Resistance: Black Feminist Perspectives on: Emerging Technologies", en A. Hanemaayer (ed.), *Artificial Intelligence and Its Discontents. Social and Cultural Studies of Robots and AI*, Cham, Palgrave Macmillan, pp. 225-249.

Shay, J. (1995): *Achilles in Vietnam: Combat Trauma and the Undoing of Character*, United States, Scribner.

Thielmann, I. & Hilbig, B. E. (2015): "Trust: An Integrative Review From a Person-Situation Perspective. *Review of General Psychology*, 19(3), pp. 249-277.

Trueba Atienza, C. y Pérez Cortés, S. (2018): "Introdución", en C. Trueba Atienza y Sergio Pérez Cortés (eds.), *Dignidad: perspectivas y aportaciones de la filosofía moral y la filosofía política*, Barcelona, Anthropos Editorial, pp. 7-12.

Tuerner, B. S. (2006): *Vulnerability and Human Rights. Pennsylvania*, Pennsylvania State University Press.

Ullmann, S. (2022): "Gender Bias in Machine Translation Systems", en A. Hanemaayer (ed.), *Artificial Intelligence and Its Discontents. Social and Cultural Studies of Robots and AI*, Cham, Palgrave Macmillan, pp. 123-144.

UE-Unión Europea (2012), *Carta de los Derechos Fundamentales de la Unión Europea*. Disponible en: https://www.europarl.europa.eu/charter/pdf/text_es.pdf

Winner, L. (1980): "Do artifacts have politics?", *Daedalus*, 109, pp. 121-136.

Wolfendale, J. (2017): "Moral Security", *The Journal of Political Philosophy*, 25(2)**,** pp. 238-255.