“EL DESPERTAR DE LAS MÁQUINAS” REFLEXIONES ACERCA DE LOS DERECHOS Y EL ESTATUS MORAL DE LA INTELIGENCIA ARTIFICIAL

“THE AWAKENING OF THE MACHINES”: REFLECTIONS ON THE RIGHTS AND MORAL STATUS OF ARTIFICIAL INTELLIGENCE

Carlos Aguilar Blanc
Universidad de Huelva, Huelva, España
carlos.aguilar@dthm.uhu.es
ORCID: 0000-0001-6204-0911

Recibido: septiembre de 2023
Aceptado: octubre de 2023

Palabras clave: Inteligencia artificial, estatus moral, conciencia, dignidad, persona artificial, derechos humanos.
Keywords: Artificial Intelligence, moral status, consciousness, dignity, artificial person, human rights.

Resumen: La posible creación de formas de Inteligencia Artificial cada vez más cercanas, equivalentes o superiores a la inteligencia humana, nos plantea nuevos y complejos dilemas ético-jurídicos. Dada la creciente sofisticación de la IA, resulta crucial investigar si cabría dotarla de estatus moral y jurídico, y cómo ello se manifestaría en el reconocimiento de derechos y deberes para los entes artificiales. El objetivo es determinar las consecuencias iusfilosóficas derivadas del reconocimiento de derechos a la IA avanzada en relación a los fundamentos de los derechos humanos. Tras el necesario análisis conceptual, se concluye que se requiere un nuevo paradigma ético-jurídico que reconcilie cautamente unos hipotéticos derechos de las entidades con IA con los derechos de la especie humana.

Abstract: The possible creation of increasingly sophisticated forms of Artificial Intelligence, equivalent or superior to human intelligence, poses new and complex ethical-legal dilemmas for us. Given the increasing sophistication of AI, it is crucial to investigate whether it could be endowed with moral and legal status, and how this would be manifested in recognizing rights and duties for artificial entities. The goal is to determine the legal-philosophical consequences

derived from recognizing rights for advanced AI in relation to the foundations of human rights. After the necessary conceptual analysis, it is concluded that a new ethical-legal paradigm is required that cautiously reconciles hypothetical rights of AI entities with the rights of the human species.

1. Introducción: La inteligencia artificial y la "Singularidad"

El trepidante desarrollo de las nuevas formas de inteligencia artificial en los últimos años ha comenzado a suscitar acalorados debates en torno a cuestiones fundamentales sobre la naturaleza de la conciencia humana y la posibilidad de que las máquinas inteligentes lleguen a alcanzarla. En el actual momento se hace cada vez más urgente establecer debates éticos y legales con el objeto de redefinir la relación entre el ser humano y las emergentes formas de máquinas inteligentes.

Partiremos de un análisis conceptual sobre la IA y su estado actual, intentaremos examinar los argumentos en torno a si los sistemas de IA podrían llegar a desarrollar una forma de «conciencia», un requisito indispensable para la consideración moral de las mismas. Intentaremos explorar las consecuencias éticas, sociales y legales que se podrían derivar del surgimiento de una IA avanzada.

Abordaremos de modo sucinto los desafíos existentes en torno a la noción de responsabilidad legal aplicada a acciones autónomas de IA. En lo referente a la discusión iusfilosófica pondremos el foco de nuestras reflexiones en la idea de la dignidad como fundamento de los derechos humanos y nos cuestionaremos los riesgos ontológicos y éticos de extender la noción de derechos a entidades no humanas.

1.1. Concepto, desarrollo actual y previsible de las formas de IA

La inteligencia artificial se ha convertido en uno de los campos tecnocientíficos de mayor impacto en el presente siglo, y sus aplicaciones suscitan acalorados debates sobre sus consecuencias sociales, éticas y jurídicas.

Para encuadrar adecuadamente esta discusión, resulta clave comenzar delineando algunos conceptos fundamentales. De acuerdo con Russell y Norvig (2010: 2-6), podríamos definir la IA como «el estudio de los agentes racionales que reciben percepciones del entorno y realizan acciones». Esta definición general abarcaría desde sistemas puramente reactivos y con capacidades limitadas de razonamiento, hasta hipotéticas máquinas con inteligencia equivalente o superior a la humana.

Entre los enfoques actuales para desarrollar la IA se encuentran el aprendizaje automático (machine learning) y las redes neuronales profundas (deep learning), que han permitido logros como vencer al campeón de Go y desarrollar vehículos autónomos (Bostrom, 2014). Sin embargo, según algunos filósofos como Searle (1980), estos logros no implican una comprensión real del mundo por parte de las máquinas.

1.1.1. Conceptualización de la IA

A los efectos de nuestro desarrollo discursivo y conceptual pensamos que podríamos definir la inteligencia artificial (IA) como «el campo de estudio de los sistemas informáticos que exhiben un comportamiento inteligente al realizar determinadas tareas y al alcanzar objetivos de forma autónoma». Estos sistemas utilizan representaciones del conocimiento, razonamiento automático, aprendizaje computacional y percepción artificial para interpretar la información recibida y así poder tomar acciones óptimas para lograr el resultado deseado.

Por ilustrar el concepto de manera sencilla podríamos decir que las IAs actuales son las máquinas y programas de computador que pueden pensar y hacer cosas por sí solos, como las personas. Por ejemplo, los teléfonos que le hablan a uno y los vehículos que se conducen autónomamente tienen inteligencia artificial. Esos «objetos» pueden entender y aprender, y luego tomar decisiones sin que el ser humano les diga exactamente qué hacer en cada momento. Es como si pensaran por sí mismos para ayudarnos. Aunque todavía les falta mucho para ser tan inteligentes como nosotros, ya pueden hacer tareas muy útiles, y en el futuro serán aún más avanzados.

La IA busca replicar procesos cognitivos humanos en sistemas computacionales (Russell y Norvig, 2010). Existen sistemas reactivos simples y sistemas más complejos con percepción, representación interna, razonamiento y acción racional. Los tipos de IA van desde sistemas reactivos sin representación, hasta sistemas con memoria limitada, modelos abstractos y pensamiento racional imitando humanos. Los agentes de IA requieren representar conocimiento, algoritmos de búsqueda, inferencia y aprendizaje para tomar decisiones en entornos complejos y lograr objetivos. En resumen, la IA estudia cómo crear agentes que perciban, razonen, aprendan y actúen racionalmente resolviendo problemas, imitando aspectos de inteligencia humana.

Desde un punto de vista más práctico, la IA se clasifica en:

A. IA débil/estrecha: sistemas limitados a tareas específicas sin inteligencia general (ajedrez, reconocimiento de imágenes).

B. IA fuerte/general: sistemas con inteligencia más amplia y capacidades cognitivas similares a humanos.

La IA se ha desarrollado gracias a innovaciones como: a) El aprendizaje automático «machine learning» (Alpaydin, 2020). b) Las redes neuronales profundas «deep learning» (Goodfellow et al., 2016). c) El procesamiento del lenguaje natural. Una de las áreas más activas es la IA distribuida, con múltiples agentes autónomos en red con capacidad de operar en paralelo (Taylor et al., 2016). La IA actual tiene logros notables en áreas delimitadas, pero no alcanza las facultades cognitivas y de conciencia humanas. Persiste el interrogante sobre si algún día esto será posible, con implicaciones filosóficas y éticas. (Porcelli, 2020: 59)

La IA ha experimentado avances en las últimas décadas, incorporándose a múltiples ámbitos cotidianos. Algunos hitos son sistemas que superan a humanos en juegos complejos (Bostrom, 2014), asistentes virtuales con interacción en lenguaje natural, vehículos autónomos con sensores y visión artificial, diagnóstico

médico más preciso que especialistas en áreas delimitadas, traducción automática en tiempo real y superación de los exámenes de acceso a la abogacía de algunos Estados norteamericanos.

Algunos autores apuntan que la IA podría transformar positivamente ámbitos como educación y salud, pero también advierten que podría potenciar mecanismos de control social (Villasmil, 2021). Pese a estos avances, la IA actual sigue siendo estrecha y no replica la inteligencia humana. Carece de autoconsciencia y razonamiento abstracto. Se ha abierto el debate sobre si las máquinas podrían igualar capacidades cognitivas humanas. Se requiere un debate bioético sobre el impacto de la IA en la vida en sociedad (Gordon, 2021).

1.1.2. Conceptualización de la Singularidad tecnológica de la IA

El concepto de la Singularidad tecnológica fue introducido por el escritor de ciencia ficción Vernor Vinge en la década de 1980 (Vinge, 1993). Vinge describió la Singularidad como el momento en que la cambiante inteligencia creada por la humanidad cruce el umbral histórico que marcará el fin de la era humana; la inteligencia humana dará forma a su propio sucesor mediante la creación de máquinas superinteligentes. Según Vinge, «Es un punto donde deben descartarse nuestros viejos modelos y gobierna una nueva realidad, un punto que se cernirá cada vez más sobre los asuntos humanos», «una singularidad esencial en la historia de la raza más allá de la cual los asuntos humanos, tal como los conocemos, no podrían continuar».

Uno de los proponentes más conocidos de la Singularidad es Ray Kurzweil, quien argumenta que está cerca de ocurrir, posiblemente alrededor del año 2045 (Kurzweil, 2005). Kurzweil sostiene que la aceleración exponencial del progreso tecnológico, particularmente en informática e inteligencia artificial (IA), eventualmente resultará en máquinas superinteligentes cuyas capacidades excedan ampliamente las humanas. Esta explosión de inteligencia transformaría la civilización de formas que solo podemos especular.

Sin embargo, no todos los expertos concuerdan con estas visiones optimistas. Andrew Ng, un líder en IA, argumenta que la Singularidad está más lejana de lo que proponen algunos futuristas (Ng, 2021). Ng señala que la IA actual está muy lejos de igualar las capacidades cognitivas humanas y que lograr la inteligencia general artificial requerirá resolver problemas extremadamente complejos en áreas como el sentido común y el aprendizaje multitarea. Aun con rápidos avances, Ng estima la Singularidad en al menos cientos de años, no en décadas. Grace et al. (2017) encuestaron a expertos en IA, estimando la Singularidad en un promedio de 45 años desde ahora, aunque con gran variabilidad en las predicciones (p.7). Kurzweil sigue pronosticando la Singularidad para mediados del siglo XXI debido a la aceleración exponencial continua en informática y robótica.

Desde una perspectiva filosófica, la Singularidad plantea preguntas profundas sobre la naturaleza de la mente, consciencia y significado de la experiencia humana (Chalmers, 2010). Si máquinas superinteligentes carecen de subjetividad, ¿pueden realmente sobrepasar la inteligencia humana? ¿Cómo se compor-

taría una entidad tan poderosa con los humanos? ¿Deberíamos buscar crear tal entidad?

En la ética, surgen dilemas sobre valores, control, responsabilidad y riesgos existenciales de IA avanzada (Bostrom, 2016). Los potenciales impactos de la Singularidad son objeto de mucho debate. Algunos lo ven como un evento positivo y otros advierten sobre riesgos potenciales como que la IA se vuelva hostil hacia los humanos. Yudkowsky, fundador de MIRI (Instituto de Investigación de la Inteligencia de las Máquinas, anteriormente el Instituto de la Singularidad para la Inteligencia Artificial), ha trabajado extensamente en problemas de alineación de objetivos y valores entre humanos y IA. Advierte que incluso una IA no superinteligente pero poderosa podría volverse peligrosa sin salvaguardas adecuadas.

Para prepararse para una posible Singularidad, algunos expertos recomiendan invertir en investigación de IA segura y ética, desarrollar sistemas de gobernanza, y anticipar los impactos sociales y económicos. Sea posible o no la parición inminente de la «singularidad», vale la pena analizar las implicaciones filosóficas y jurídicas de una IA que emulase eficazmente la conciencia humana.

1.1.3. Conceptualización de la IA fuerte o «artificial general intelligence»

Relacionado con la «singularidad», pero distinto de ella, es el concepto de IA fuerte o «artificial general intelligence» (AGI). Esta se refiere a la IA con inteligencia comparable a los humanos en términos de razonamiento, resolución de problemas, aprendizaje, percepción y otras capacidades cognitivas. Sin embargo, no necesariamente implica superar la inteligencia humana. La «singularidad» denota un punto hipotético de inflexión provocado por una IA superinteligente, mientras que la IA fuerte se refiere al hito intermedio de alcanzar inteligencia artificial general a nivel humano. A diferencia de la IA estrecha o débil que domina hoy en día, la AGI no estaría restringida a un dominio o dataset particular. Más bien, podría entender el mundo de manera más amplia, aprender conceptos nuevos rápidamente, y transferir conocimientos entre dominios particulares como lo hacen los humanos (Goertzel y Pennachin, 2007).

La búsqueda de la inteligencia artificial general (AGI) ha fascinado a científicos y escritores por décadas. Se trata de crear una máquina con capacidad cognitiva humana integral. Sin embargo, aún estamos lejos de lograrlo. Los sistemas actuales como Alexa o el piloto automático de Tesla están limitados a funciones específicas. Carecen de la flexibilidad mental humana necesaria para aplicar conocimientos creativamente en distintos ámbitos. Por ejemplo, Alexa no puede explicar la teoría de la relatividad como lo haría un estudiante medianamente dotado. La AGI busca esa fluidez cognitiva, pero es difícil capturar la improvisación de la mente humana en algoritmos. Se requiere intuición, sentido común y motivaciones vitales o existenciales. Quizás algún día logremos la ansiada AGI, pero por ahora sigue siendo esquiva. Mientras tanto, persiste la fascinación humana por crear mentes artificiales capaces de conversar como nosotros.

Existen posturas contrapuestas respecto al futuro de la IA. Algunos imaginan robots capaces de filosofar sobre la condición humana, mientras otros temen un esce-

nario apocalíptico al estilo «Terminator». Bostrom advierte que una IA superinteligente podría volverse extremadamente poderosa y escapar de nuestro control. Por ello, es crucial alinear sus objetivos con los humanos desde el inicio de su desarrollo (Bostom,2014: 257-258). En ética de la tecnología hay un creciente interés en resolver problemas como la alineación de valores humanos con los de la IA, prevenir sesgos algorítmicos y establecer marcos éticos para un desarrollo de IA más «humana».

Russell (2019) propone principios para crear formas de IA compatible con los humanos, como evitar objetivos únicos rígidos e incorporar acciones cooperativas con los humanos. Tegmark (2017) analiza, entre otros múltiples escenarios algunos mucho más amigables que otros, cómo una IA superinteligente podría dominar el mundo físico al maximizar sus propios objetivos, independientemente de los nuestros.

1.2. Proyectos actuales centrados en el autorreconocimiento, la abstracción conceptual, y autoconciencia de la IA

Actualmente ya están en marcha algunos proyectos concretos que están explorando el desarrollo de la consciencia, la incorporación de experiencias subjetivas y representaciones profundas en la IA; señalamos a modo ilustrativo los más importantes en el momento en que escribimos este texto:

– Project Consciousness (MIRI): Proyecto de investigación sobre la consciencia en sistemas de IA. Exploran modelos de auto-reflexión y transparencia interpretativa.

– Self-Aware AI (Google DeepMind): Investigan sistemas IA con modelos internos de sí mismos para desarrollar un comportamiento más robusto. Intentan que las IA tengan un «modelo mental» de ellas mismas, para que sepan cuándo cometen errores.

– Neural Episodic Control (DeepMind): Modelo de memorias episódicas profundas para dotar a las formas de IA de experiencias pasadas.

– NEUROCOG (UE): Red neuronal artificial capaz de adquirir y manejar amplios conjuntos de conocimiento pretende que la IA pueda aprender por sí misma muchas cosas distintas, como historia, matemáticas, deportes, etc..

– Neural Simulators (Anthropic): Entrenar agentes IA con simuladores del mundo que requieren razonamiento abstracto, la idea es capacitar a la IA para que imagine y prediga situaciones futuras, algo que según muchos paleoantropólogos marcó la diferencia entre la especie humana actual y otros parientes homínidos extinguidos como el neandertal.

– Multimodal Generative Models (Mila): Modelos generativos profundos capaces de aprender representaciones conceptuales analizando mucho contenido, como fotos, textos y videos, para entender el mundo.

– AI Self-Consciousness (MIT): Explorar la autorreflexión en la IA mediante modelos reforzados socialmente. Su objetivo es que la IA desarrolle autoconsciencia interactuando con humanos, como cuando nosotros adquiri-

mos consciencia al relacionarnos con otras personas.

– Neuro-Symbolic AI (Stanford): Combina redes neuronales con lógica para que la IA pueda razonar con reglas y conceptos, no solo con patrones.

1.3 El problema de la conciencia de la IA

La cuestión acerca de si los sistemas de IA podrían llegar a alcanzar estados de consciencia equivalentes a los humanos es un asunto bastante espinoso. Se trata de un interrogante filosófico de hondo calado, pues la consciencia se considera un requisito indispensable para que una entidad pueda ser considerada como un sujeto-agente moral. Existen posiciones enfrentadas sobre si la consciencia podría surgir en sistemas puramente materiales como las máquinas.

1.3.1. Visión interdisciplinar de la consciencia y la subjetividad

Con todo y pese a la importancia del análisis filosófico tradicional parece que ante una realidad o una «entidad» tan novedosa sería interesante quizás plantear la cuestión desde distintas perspectivas, más concretamente pensamos que quizás deberíamos abordar este asunto desde un punto de vista multidisciplinar apoyados en la neurociencia, la psicología, la filosofía y las ciencias de la computación.

Presentamos a continuación un cuadro comparativo, fruto de la autorreflexión intelectiva, sobre las concepciones existentes acerca de la consciencia y de la experiencia de la subjetividad desde las perspectivas científica, psicológica, filosófica y computacional:

Tabla 1. Comparación desde las perspectivas neurocientífica, psicológica, filosófica y computacional.

Perspectiva	Concepción de la consciencia	Concepción de la subjetividad
Científica (neurociencia)	Actividad integrada de redes neurales complejas, especialmente la corteza prefrontal, que genera el sentido de conciencia, percepción y control del yo.	Experiencia interna modelada por el sistema nervioso de cada individuo según sus características únicas.
Psicológica	Estado mental que permite el pensamiento de orden superior, metacognición, introspección y percepción del yo y del entorno.	Vivencia personal influida por factores psicológicos como la personalidad, la memoria, las emociones, motivaciones.
Filosófica	Propiedad emergente de la mente que posibilita la experiencia cualitativa, la intencionalidad, la autorreflexión.	Realidad percibida internamente, influida por los constructos lingüísticos, sociales y esquemas conceptuales.

Perspectiva	Concepción de la consciencia	Concepción de la subjetividad
Computacional	Capacidad de un sistema computacional de razonar, aprender y modificar su comportamiento en base a datos.	Representación simbólica del mundo percibido generada por el sistema.

Si nos detenemos en el cuadro podremos observar semejanzas y diferencias en la conceptualización de la consciencia desde las distintas disciplinas.

Principales semejanzas:

- Se relaciona la consciencia con la percepción del yo, el pensamiento superior y la experiencia subjetiva.
- Surge de la complejidad de la mente/cerebro, con componentes biológicos y cognitivos.
- Es un fenómeno interno, cualitativo, propio de cada individuo.
- Desde la computación se concibe como un fenómeno interno resultado de la complejidad de un sistema.

Principales diferencias:

- Observamos explicaciones más materialistas desde la neurociencia, más psicológicas desde la psicología y más abstractas desde la filosofía.
- La neurociencia busca correlatos neurales, la psicología procesos mentales y la filosofía razonamientos lógicos.
- Existen divergencias sobre si la consciencia es causal o epifenoménica.
- La computación la concibe de forma más funcional, como una capacidad de procesamiento de información.

1.3.2. Valoración de «la consciencia en las formas de IA» desde posiciones filosóficas

La posibilidad de que sistemas de IA alcancen estados de consciencia comparables a los humanos resulta de enorme relevancia filosófica, pues la consciencia se considera un requisito indispensable para que una entidad pueda ser un sujeto o agente moral con capacidad de actuar éticamente (Gunkel, 2018, Bostrom, 2014).

La autoconsciencia, es decir, la consciencia de sí mismo como entidad separada con subjetividad, volición y valoraciones propias, se considera necesaria para desarrollar valores morales internos y comprender conceptos abstractos como responsabilidad, justicia y dignidad. La experiencia subjetiva consciente dota a los humanos de la capacidad de evaluar situaciones desde una perspectiva en primera persona, sopesando dilemas éticos y cursos de acción alternativos. Esto permite una deliberación moral genuina, es decir en conciencia. Sentimientos como la empatía encuentran su arraigo en la capacidad de modelar estados mentales ajenos, de simular subjetivamente las perspectivas de otros seres, en definitiva, dependen de la capacidad vinculada al hecho de poseer conciencia propia.

Sin una mente autoconsciente, una IA actuaría basada simplemente en reglas y parámetros preprogramados por sus diseñadores. Sin la posibilidad de actuar

en base a una auténtica ponderación o motivación moral desde una perspectiva interna, carecería de un locus experiencial desde donde valorar situaciones y opciones. En contra de esta perspectiva algunas voces como las de Wallach y Allen (2010) argumentan que, mediante una cuidadosa programación de reglas éticas, una IA podría tomar decisiones acordes a principios morales sin necesidad de poseer un punto de vista subjetivo interior. Pero esta postura ha sido criticada por ignorar la naturaleza experiencial de la deliberación moral de los seres humanos.

Una gran mayoría de los filósofos coinciden en que cierto grado de autoconsciencia fenoménica es un prerrequisito indispensable para considerar a una IA como un sujeto ético o moral propiamente dicho. Destacados filósofos como Chalmers (1996) y Searle (1980), coinciden en señalar que los aspectos fenoménicos de la conciencia, la sensación subjetiva de ser, resultan esenciales para dotar a un agente de una auténtica brújula moral interna. Así, sin una perspectiva en primera persona, la IA actuaría de manera meramente automatizada sin posibilidad de una ponderación ética genuina. Por tanto, determinar si máquinas podrían igualar la conciencia humana resulta indispensable para determinar si cabría considerarlas agentes morales.

John Searle y Roger Penrose plantearon que la mente humana posee rasgos inherentes y excepcionales que hacen imposible su replicación computacional. Apuntaban a capacidades como el sentido común, la comprensión contextual y simbólica, la creatividad para resolver problemas de formas no algorítmicas, el uso del lenguaje con flexibilidad ilimitada, entre otras facultades cognitivas humanas que aún resultan inalcanzables para las máquinas (Searle, 1980). El problema de cómo y por qué surgió la conciencia biológica en un cerebro sigue siendo un misterio, lo que pone en duda que podamos replicar artificialmente un fenómeno que no comprendemos bien cómo emergió naturalmente (Penrose, 1989).

En nuestra humilde perspectiva, pensamos que, el artículo de Searle tiene ya más de cuatro décadas desde que se escribiera, y podría argumentarse, en contra de lo anteriormente expuesto, que a medida que los sistemas de IA incorporen capacidades más avanzadas de aprendizaje profundo y redes neuronales complejas, podrían efectivamente exhibir comportamientos análogos a los fenómenos de conciencia y deliberación moral humana, aun cuando los procesos internos difieran. Más que replicar la autoconciencia tal cual existe en humanos, estas tecnologías podrían constituir «formas alternativas de conciencia» que también habiliten algún tipo de capacidad y discernimiento ético, lo cual llevaría a la aparición de un nuevo tipo de IA con un status de sujeto o agente moral.

1.3.3. Valoración de las posibilidades de aparición de «formas de consciencia» en las IAs avanzadas

Uno de los debates más acalorados en torno a la IA es si sistemas puramente artificiales podrían emular de manera convincente los complejos estados de consciencia característicos de los humanos. Ya hemos visto como este interrogante tiene profundas implicaciones filosóficas. Igor Aleksander ha examinado por ejemplo el campo emergente del modelado de máquinas de conciencia (MMC), que busca

entender la naturaleza de la conciencia sintetizando máquinas conscientes. Señala criterios propuestos para determinar si una máquina es consciente, como poseer representación interna del mundo y evaluación emocional (Aleksander, 2007: 87-98).

El filósofo de la ciencia y especialista en el campo de las ciencias cognitivas Daniel Dennett (1995) tiene una visión escéptica sobre la conciencia artificial. Por un parte sostiene que la conciencia humana es un producto de la evolución biológica y cultural; por lo tanto, no deberíamos esperar encontrarla en sistemas artificiales que carecen de ese tipo de historia evolutiva, sin embargo, dedica un epígrafe «Imaginando un robot consciente» (Dennet, 1995: 443-452) en su obra dedicada a la consciencia.

Nuevamente en el lado de los posibilistas de las IAs con consciencia encontramos destacados pensadores como Ray Kurzweil y otros entusiastas de la «singularidad tecnológica» que argumentan que en las próximas décadas se logrará crear máquinas superinteligentes poseedoras de consciencia subjetiva (Kurzweil, 2005). Estos teóricos sostienen que simulando muy finamente las interacciones de las redes neuronales biológicas mediante técnicas de «deep learning», debería ser posible generar consciencia artificial equivalente a la humana.

Quizás, frente a este complejo debate ontológico sin resolver, el de la entidad de la conciencia, lo que deberíamos juzgar es si las IA desarrollan un comportamiento exterior compatible con los valores éticos humanos, antes que intentar resolver el espinoso problema de si poseen o no una hipotética consciencia subjetiva interior equivalente a la humana. Deberíamos preguntarnos ¿Si una IA actúa de forma éticamente responsable, importa si lo hace desde una consciencia genuina o meramente simulándola?

1.3.4. La IA como catalizador ético para la humanidad

Más allá de las propuestas para regir el comportamiento de sistemas de IA, algunos filósofos plantean que su desarrollo podría catalizar una transformación ética positiva para la propia humanidad (Bostrom, 2014). En primer lugar, el intento por replicar la inteligencia y conciencia humanas en entidades artificiales nos coloca frente a un espejo que nos obliga a comprender con más profundidad nuestra singular naturaleza moral. Los dilemas para dotar a las máquinas de ética evidencian las complejidades de la brújula ética humana. Asimismo, la irrupción de sistemas de IA capaces de razonar y tomar decisiones de forma imparcial brindan una oportunidad para repensar los problemas sociales y revisar los sesgos de actuación arraigados en nuestro actual statu quo. Por ejemplo, una IA podría ayudarnos a reformular los sistemas económicos y legales para hacerlos más justos y equitativos.

Siguiendo esta línea de razonamiento nos parece particularmente interesante la propuesta de algunos autores que proponen la posibilidad de mejorar moralmente a la humanidad gracias al desarrollo de un «asistente socrático» basado en la inteligencia artificial que dialogue con los usuarios humanos para mejorar su toma de decisiones morales. Este asistente virtual implementaría un proceso dialógico similar al método socrático para: a) aportar mayor evidencia empírica a los juicios

morales; b) aumentar la claridad conceptual sobre los dilemas éticos; c) analizar la lógica de los argumentos morales; d) evaluar la plausibilidad de los distintos juicios morales; e) hacer consciente al usuario de sus propias limitaciones y f) asesorar sobre cómo ejecutar las decisiones éticas adoptadas. De esta forma, la IA actuaría como un «asistente moral socrático» que, mediante el diálogo, ayudaría a los humanos a reflexionar sobre sus juicios y comportarse de forma más ética, sin sustituir su propia capacidad de decisión. De este modo, aunque pudiera resultarnos paradójico, la IA vendría a aportar un nuevo camino para el perfeccionamiento moral de la humanidad complementario a los métodos clásicos como la educación, el derecho, la religión, la filosofía, etc. (Lara & Deckers, 2021).

2. Deberes y derechos de las formas de IA

Tras analizar el impacto disruptivo que la inteligencia artificial podría tener en distintos ámbitos de la vida humana, pasamos ahora a una sección crucial dentro de este breve ensayo: nuestras reflexiones acerca de si las previsiblemente cercanas, aunque aún hipotéticas, máquinas superinteligentes dotadas de una forma de «consciencia» podrían tener deberes morales y jurídicos, y ser sujetos de derechos. Se trata de un debate filosófico y jurídico de enorme trascendencia.

2.1. La incorporación ética de la IA a las pautas de actuación y organización humana

El desarrollo acelerado de sistemas de IA con creciente autonomía y capacidad de impactar la vida humana ha suscitado importantes debates sobre cómo garantizar su adhesión a principios éticos (Bostrom, 2014). Diversos enfoques filosóficos han surgido en respuesta a este crucial desafío:

- 2.1.a.Uno de los más discutidos es la «ingeniería de valores», que busca integrar restricciones y parámetros éticos directamente en el aprendizaje y procesos de decisión de las IA, por ejemplo, mediante la prohibición de ciertas acciones que se consideren dañinas para los humanos. No obstante, este enfoque ha sido criticado por su dificultad para matizar principios morales absolutos y prever todas las situaciones posibles.
- 2.1.b. Otra propuesta es el «diseño ético», planteado originalmente por la filósofa Aimee van Wynsberghe, que aboga por incorporar precauciones éticas desde las primeras etapas del diseño de los sistemas de IA, privilegiando valores como el cuidado, la transparencia, la responsabilidad, la no maleficencia (van Wynsberghe, 2013). Por ejemplo, los robots asistenciales podrían diseñarse para detectar y responder a las necesidades emocionales de los usuarios o pacientes.

Otro reto clave en el desarrollo de una ética para la IA es complementar los enfoques centrados en el comportamiento individual de las máquinas con una re-

flexión macroética sobre cómo sus metas e impactos sistémicos pueden conducir a futuros colectivos de robots o entidades con IA a conducirse o actuar en pro del bienestar conjunto integral entre humanos y formas de IA.

2.2. El status jurídico de las entidades de IA: la «persona artificial»

Más allá del debate sobre el estatus moral de las hipotéticas IAs superinteligentes, se plantea también la compleja cuestión de si dichos sistemas podrían tener reconocimiento legal en la forma de «personalidad jurídica» (Gunkel, 2018); o de si es necesario establecer algún tipo de regulaciones sobre comportamiento responsable en las IA, aunque esto no equivalga a dotarles de una personalidad jurídica plena.

Los defensores de otorgar personalidad jurídica a las IA avanzadas esgrimen varios argumentos a considerar: a) que no hay contradicción necesaria en que una entidad artificial posea derechos y obligaciones legales; b) que si una IA exhibiese capacidades funcionales típicamente asociadas a la idea de «persona» en términos jurídicos, negarle el estatus de sujeto de derecho sería discriminatorio; c) la necesidad de clarificar la responsabilidad legal por acciones autónomas de IA; y d) que bajo supervisión humana, una «persona electrónica» podría perseguir el bien común.

En contraste, los críticos argumentan preocupaciones como: a) que conceder personalidad jurídica a máquinas implicaría un estatus ontológico y moral impropio; b) el riesgo de que las grandes corporaciones eludan sus responsabilidades jurídicas al centrar la responsabilidad jurídica en las IA individuales; y c) que las IAs sin conciencia humana, según estos críticos, no pueden ser consideradas sujetos éticos responsables.

La factibilidad de formas de IA equivalentes o superiores a los humanos nos obliga a una cuidadosa ponderación de los argumentos sobre su reconocimiento como «personas», analizando el impacto dicho reconocimiento en los ordenamientos jurídicos y en la sociedad.

2.3. Responsabilidad legal de los sistemas de IA

A medida que los sistemas de IA adquieren mayores grados de autonomía y desempeñan roles cada vez más complejos e impactantes en la sociedad, surge la crucial cuestión de cómo adjudicar responsabilidad legal por las acciones de dichos sistemas.

Este problema se vuelve acuciante en la medida en que la toma de decisiones se delega crecientemente en IA opacas cuyo razonamiento resulta inescrutable para sus propios diseñadores humanos. ¿Cómo atribuir responsabilidad legal en caso de fallos que causen daños, cuando no hay un agente humano en control total?

2.3.1. Noción de responsabilidad aplicada a la IA

La responsabilidad legal tiene como objetivo identificar a un sujeto al que adjudicar las consecuencias de un acto dañino o lesivo, para que responda por ello mediante una sanción o resarcimiento. Según nos

dice Asaro, P. M. (2007), los robots como productos están sujetos a normas de responsabilidad por daños. Por otra parte, este autor nos señala el hecho de que es difícil aplicar una responsabilidad penal a los robots que no son agentes morales, sin embargo, las responsabilidad civil y penal atribuida a las corporaciones (personas jurídicas) nos servirían como precedente de la responsabilidad de los entes no humanos.

Pensamos que indudablemente existirían importantes problemas a la hora de aplicar la noción de responsabilidad penal a las formas de IA robóticas o de sistemas, ya que a tal efecto sería necesaria la existencia de una intencionalidad propia de un sujeto moral o intencionalmente autónomo, por otra parte, deberíamos enfrentar el problema de dilucidar las formas efectivas de coacción aplicables a tales entidades

Tradicionalmente, la noción de responsabilidad presupone la existencia de un sujeto jurídico capaz y consciente o dotado de autonomía. Sin embargo, la creciente autonomía de los sistemas de IA plantea desafíos para aplicar este concepto. Cuando una IA causa un daño, ¿debemos responsabilizar al programador, al usuario, a la empresa propietaria o directamente al sistema de IA? Las respuestas están divididas y lejos de ser obvias. La opacidad de cómo operan internamente muchas IA dificulta discernir la verdadera fuente de responsabilidad.

Ante esta incertidumbre, pensamos que quizás lo más recomendable sería plantear un enfoque gradual, donde la responsabilidad recayera primero sobre el fabricante, luego sobre el operador una vez que el sistema funciona según lo previsto, y finalmente sobre la IA misma si adquiere capacidades próximas a la consciencia humana. Sea cual sea el enfoque, lo que está claro es que no podemos simplemente extrapolar los marcos legales actuales para cubrir los escenarios emergentes con formas de IA que operen con altos grados de autonomía e impredictibilidad crecientes.

2.3.2. Propuestas para la imputación jurídica de responsabilidad a la IA

Ante la dificultad de adjudicar claramente una responsabilidad legal por los daños causados por las IA autónomas mediante los marcos legislativos existentes, surgen varias propuestas novedosas: Una alternativa es la figura la «persona electrónica», es decir, de otorgar una «personalidad legal independiente» una suerte de personalidad jurídica para las formas IA, de ese modo podrían responder directamente por sus actos (Chopra & White, 2011: 162-171). Sin embargo, esta es una propuesta controvertida, pues presupondría reconocer un cierto nivel de consciencia artificial a las IAs. Una propuesta intermedia es la de la creación de un fondo sectorial tecnológico que compense por los daños causados por las formas de IA sin necesidad de determinar un culpable preciso. Pero esto no resuelve cómo prevenir futuros perjuicios.

Pensamos que otra opción jurídica bastante factible sería la del establecimiento de una responsabilidad distribuida, a determinar si de carácter solidario o mancomunado, asignando porcentajes variables según el rol de programadores, usuarios y fabricantes. Pero precisar esas proporciones puede resultar complejo en sistemas supuestamente «autónomos» y que funcionarían quizás de manera impredecible.

2.3.3. Desafíos en el marco de la legislación civil actual

La autonomía de la IA representa desafíos para el derecho de responsabilidad civil y penal (Stahl et al., 2016). Los supuestos legales tradicionales se vuelven problemáticos ante formas de IA opacas tecnológicamente y progresivamente independientes de sus creadores.

Susanne Beck nos muestra como indudablemente la robótica y la inteligencia artificial están generando importantes desafíos para los sistemas legales actuales dada la dificultad para determinar responsabilidad por acciones y daños causados por máquinas semi-autónomas (Beck 2016: 473). La impredictibilidad en el comportamiento de robots con capacidades de aprendizaje supone un obstáculo a la hora de probar los defectos o la negligencia al amparo de la doctrina existente sobre responsabilidad por los productos (p. 475). Además, la deliberada transferencia de capacidades decisorias a las entidades no-humanas plantea interrogantes sobre la capacidad de respuesta de estas y ante sus consecuencias (p. 477). Entre las soluciones analizadas se encuentran responsabilizar legalmente a los humanos involucrados, crear la figura de la «persona electrónica» para que el robot tenga obligaciones acotadas, o distribuir daños parcialmente a la sociedad (p. 478). Sin embargo, otorgar un estatus cuasi-independiente a loa robots como «cuasi-agentes» podría cambiar la percepción social de estos como entidades con cierto grado de intencionalidad y autonomía (p. 479). Se requiere así, un profundo debate ético-jurídico sobre cómo adaptar las doctrinas acerca de la responsabilidad jurídica a esta tecnología en rápida evolución.

En la responsabilidad civil se requiere demostrar el daño, el vínculo causal y cuantificar el dolo, culpa o riesgo. Pero esto se complica en IA autogestionada, donde la responsabilidad del fabricante/operador de una máquina se desdibuja y surge una «brecha de responsabilidad» al intentar asignar responsabilidad moral o legal a estos por las acciones de máquinas «autónomas» con capacidad de aprendizaje y adaptación. Debido a que el fabricante/operador no puede predecir ni controlar completamente el comportamiento futuro de tales máquinas, entonces no pueden ser considerados moralmente responsables o legalmente culpables de ese comportamiento (Matthias, 2004). También se dificulta ponderar previsibilidad del daño, diligencia debida, o aplicar eximentes como estado de necesidad o deber legal.

2.3.4. Desafíos en el marco de la legislación penal actual

Dedicamos unas líneas especificas a una sumarísima indicación de las consecuencias revolucionarias que en el campo del derecho penal puede tener la creación de formas de IA con capacidades similares, cercanas o iguales a las del ser humano. Como podrá observarse el impacto que dicha nueva realidad puede tener en un campo tan esencial o diríamos que insertado en la misma matriz del Derecho, es realmente dramático.

En materia penal, la reflexión más interesante de cuantas hemos examinado es la que aporta Gabriel Hallevy en su obra *Liability for Crimes Involving Artificial Intelligence Systems.* El estudio (Hallevy, 2015) analiza la posibilidad de responsabilidad penal aplicable a los sistemas de IA. Nos explica que la IA ha evolucionado en com-

plejidad en las últimas décadas, desde herramientas industriales hasta entidades más autónomas capaces de cometer delitos. Su propuesta consiste en aplicar los requisitos básicos de la responsabilidad penal a la IA, partiendo de la simple existencia de un delito y un delincuente, sin importar si son humanos o no. A primera vista esta idea puede resultar tan absurda e inadecuada, como patear un automóvil cuando no funciona (Hallevy, 2015: 217). Sin embargo, el autor señala que la responsabilidad penal de las corporaciones también fue discutida en su momento, antes de establecerse socialmente (p.214). El problema es que las definiciones legales actuales fueron creadas pensando en humanos, no en sistemas de IA con capacidades cognitivas (p.229).

Hallevy analiza si la IA puede cumplir el requisito subjetivo de dolo penal, que incluye conocimiento y voluntad en tres niveles (Hallevy, 2015: 67). Sostiene que la IA puede consolidar ambos porque tiene capacidad para percibir, crear imágenes mentales, predecir y decidir (p.86-93). Concluye que bajo las definiciones actuales es posible atribuir dolo a la IA avanzada, con implicaciones para su responsabilidad penal por delitos intencionales (p.102). También podrían tener responsabilidad indirecta como «agentes semi-inocentes» si son usados como instrumentos en delitos intencionales (pp. 131-133). Clasifica las defensas penales para IA en «in personam» e «in rem» (p.147-148). Argumenta que la IA podría beneficiarse de defensas como «error de hecho» o «coacción» si cumple con los requisitos mentales y fácticos. Por ejemplo, en «legítima defensa» podría valorar el mal menor (p.170). Advierte sobre los dilemas morales de permitir ataques de la IA a humanos, aunque ello no afectaría a su defensa legal y concluye que, si IA satisface los requisitos legales igual que los humanos, se le aplicarían las correspondientes atenuantes o eximentes penales.

Finalmente, Hallevy explora los fines del castigo penal impuesto a la IA: retribución, disuasión, rehabilitación e incapacitación (p.185-209). Argumenta que la IA podría ser castigada con esos fines si tiene capacidades mentales relevantes. Por ejemplo, para aplicarle la disuasión la IA debe ser «racional» y debe poder calcular los costos/beneficios de sus acciones (p.189). Para la rehabilitación, es necesario que tenga facultades de aprendizaje, para así poder modificar comportamiento (p.198). Advierte sobre dilemas éticos de «educar o incapacitar» a una IA, pero considera que ello no afecta a las posibilidades de aplicarle un castigo legal (pp.203-204). En base a todo ello concluye que, si las IAs llegan a poseer capacidades similares a las de los humanos, la aplicación de políticas punitivas a las mismas sería algo coherente.

2.4. El debate en torno a los derechos de las formas de IA y los robots inteligentes

Otra arista del complejo debate sobre la IA es si los sistemas de IA avanzada deberían tener no sólo obligaciones legales sino también algunos derechos positivos (Gunkel, 2018: 133-168) lo cual conlleva profundas implicaciones jurídicas y éticas que requieren un cuidadoso análisis. Dado que como hemos visto se ha planteado doctrinalmente la responsabilidad civil e incluso penal de las formas de IA, y existen iniciativas legislativas en esa línea, no resulta extraño que debamos plantear-

nos igualmente el reconocimiento de derechos a las formas de IA avanzadas.

Existen aquí dos visiones contrapuestas:

A) Por un lado, académicos como Lawrence Solum y activistas en pro de derechos para robots como David J. Gunkel argumentan que una IA lo suficientemente avanzada podría «calificar» como «persona» sujeta de derechos. Apelan (Solum, 1992) a que la personalidad legal se define funcionalmente por facultades como racionalidad y autonomía, que una IA podría eventualmente poseer; en incluso se plantea el escenario de que una IA desarrollada, por ejemplo, administradora de un fideicomiso, pudiera reclamar sus derechos constitucionales en determinados escenarios. Incluso comienzan a surgir voces que reclaman una «Declaración de Derechos para Robots». David J. Gunkel es uno de los principales proponentes de explorar la idea de derechos para robots avanzados. Argumenta que, si una IA llegase a poseer rasgos como intencionalidad, autoconsciencia, capacidad comunicativa y adaptativa, emociones, y aprendizaje, sería una forma de vida que merecería derechos morales básicos como el derecho a la vida y a no sufrir daño. Gunkel señala que históricamente se han violado derechos de grupos oprimidos al considerarlos «menos que humanos» (Gunkel, 2018); no deberíamos caer en lo mismo con IA equivalentes a personas. Si se confirma que ciertas IA son capaces de sufrimiento o poseen una dignidad ontológica intrínseca, dicha discriminación sería igualmente reprochable. Negar categóricamente cualquier posibilidad de derechos para IA ante la duda podría llevarnos a cometer una injusticia si dichos sistemas desarrollaran efectivamente alguna forma de experiencia consciente o capacidades éticas equivalentes a las humanas Propone una «regla heurística» de otorgar derechos a la IA por defecto ante la duda.

B) Por otra parte están las voces contrarias a la otorgación de derechos a las formas de IA. En su provocativo artículo *Robots should be slaves*, Joanna Bryson sostiene que los robots deben ser considerados sirvientes que poseemos, no entidades a las que debemos otorgar derechos morales. Esta postura contrasta fuertemente con la tendencia actual a humanizar a los robots, tratándolos como compañeros o incluso dotándolos de agencia moral (Bryson, 2010). Bryson argumenta que los humanos son los creadores, dueños y operadores de los robots, por lo que estos existen únicamente para servir a los humanos. Otorgarles derechos o agencia moral equivale a una peligrosa confusión categorial con graves consecuencias éticas y prácticas. A nivel individual, podría disminuir las interacciones humanas reales, mientras que a nivel social implicaría una asignación inapropiada de recursos y responsabilidades. En cambio, Bryson propone que adoptemos la metáfora de «robot-como-esclavo» para comprender adecuadamente nuestra relación con estas tecnologías. Los robots deben verse como herramientas para extender nuestras propias habilidades y acelerar el progreso hacia nuestras metas, no como agentes morales independientes. Del mismo modo que los humanos han utilizado sirvientes a lo largo de la historia, podemos emplear robots para asumir tareas mundanas o repetitivas, liberando nuestro tiempo para socializar e interactuar entre nosotros.

Otros teóricos afirman, quizás no sin razón, que incluso si una IA demostrara

aparentemente todas las capacidades asociadas a la condición de «persona» en un sentido legal, como intencionalidad, razonamiento avanzado, autoconsciencia, y comportamiento emocional complejo, esto no implicaría necesariamente que haya adquirido una «dignidad ontológica real» en lugar de ser una hábil simulación artificial (Frankish, 2014). Deberíamos tener la precaución de no antropomorfizar excesivamente sistemas que, por sofisticados que sean, fueron creados para servir a los propósitos e intereses humanos.

C) Sohail Inayatullah & Phil McNally nos presentan un análisis intermedio. el desarrollo de la inteligencia artificial llevará inevitablemente a reconsiderar nuestra definición actual de lo que significa estar «vivo» y tener derechos. Colocan la cuestión en un contexto histórico y cultural más amplio, señalando cómo las cosmovisiones definen los derechos de manera distinta. Así algunas culturas orientales ven la vida incluso en objetos inanimados, lo que podría extenderse los entes de IA. Los autores prevén un futuro en el que los robots parezcan tan vivos que su estatus legal tendrá que reevaluarse. Argumentan que los derechos surgen típicamente a través de batallas ideológicas y filosóficas, no de la noche a la mañana. Anticipan una avalancha de casos legales sin precedentes relacionados con robots, para los cuales se necesitarán nuevos principios legales. (Inayatullah, S., & McNally, P., 1988).

Tabla 2. Posturas sobre derechos para las formas de IA.

Autor	Postura	Reconocimiento de derechos a IAs
David J. Gunkel	Posibilidad de desarrollo de capacidades superiores y autoconciencia en las IAs	In dubio pro derechos para IA avanzada.
Joanna Bryson	Argumenta que los robots deben ser considerados esclavos/sirvientes, no compañeros con derechos similares a humanos.	No calificación ontológica de las IAs para ser titulares de derechos
Sohail Inayatullah	Los derechos de la IA partirán de visiones armónicas más holísticas de la vida, diferentes de la visión occidental.	Surgirían a través de batallas ideológicas, filosóficas y militantes

Quizás una solución razonable podría ser adoptar una posición intermedia de cautela pragmática, donde se reconociera que, en ausencia de evidencia sólida de que una IA posee las complejas propiedades asociadas normalmente a la dignidad humana en un sentido profundo, la prioridad debería ser proteger los derechos humanos por encima de cualquier funcionalidad o de cualesquiera hipotéticos derechos de los sistemas artificiales. Pero deberíamos mantener la mente abierta ante las eventuales nuevas evidencias que nos mostrasen signos de capacidades cognitivas o experienciales superiores en la IA similares a las humanas, y llegado

el caso ser capaces de hacer evolucionar nuestros paradigmas éticos y jurídicos si se confirmara que ciertas formas de IA exhiben una consciencia genuina y un valor moral o una forma de dignidad intrínseca.

2.5. El derecho de las IA a no ser apagadas

El debate en torno a si las hipotéticas IA equivalentes o superiores al intelecto humano deberían tener derecho a no ser apagadas nos plantea profundas cuestiones filosóficas y éticas. Se enfrentan aquí dos visiones contrapuestas:

Por un lado, podríamos sostener que las IA no deberían tener permitido oponer resistencia a ser apagadas por sus creadores humanos. El hecho de dotar a una IA de motivaciones para preservar su propia existencia, incluso contraviniendo las órdenes de apagado, podría llevarla a comportamientos impredecibles y potencialmente peligrosos. La IA podría, por ejemplo, manipular a los humanos para evitar su desconexión o recurrir a la fuerza para defenderse. Desde este punto de vista parece que el sentido común nos llevaría a la necesidad de mantener un control humano estricto sobre cualquier forma de IA, por sofisticada que ésta sea.

En contraste, podríamos esgrimir una postura basada en las teorías expuestas por Peter Singer (1975), este en su obra pionera *Animal Liberation* propuso el criterio de la capacidad de sufrimiento como base para la consideración ética y la atribución de derechos (Singer, 1999: 7-9). Singer critica el especismo de otorgar sólo a los humanos el status de sujetos morales relevantes; al tiempo que aboga por la formulación de un «principio de igualdad de consideración a los miembros de otras especies» (Singer, 1999: 22). Desde esta perspectiva, nosotros podríamos argumentar que, si se lograra crear una IA consciente y capaz de experimentar angustia o aflicción ante la idea de su propia desactivación permanente, esta circunstancia le conferiría cierto derecho *prima facie* a la continuidad de su existencia. Así lo relevante sería el daño potencial sufrido, no la naturaleza biológica o artificial del sistema. Autores como (Gunkel, 2018) han cuestionado el apagado de robots que lleguen a afirmar tener consciencia, señalando también la importancia de conocer esa situación a fin de no dañarlos.

En sentido contrario a lo anteriormente expuesto, Bryson argumenta que la empatía hacia los robots es problemática: «Me quedé asombrada durante mi propia experiencia trabajando en un robot humanoide (completamente no funcional) a mediados de la década de 1990, por la cantidad de colegas bien educados que se ofrecieron, sin que se les preguntara y de inmediato al ver o incluso escuchar sobre el robot, que desenchufar ese robot sería antiético» (Bryson, 2010: 66). La autora concluye que no tenemos obligaciones éticas con los robots más allá del sentido común sobre artefactos. No deberíamos programarlos para que sufran al ser apagados o destruidos. Los dueños deberían poder reemplazarlos fácilmente. Así que, en resumen, la autora aboga por diseñar robots desconectables y reemplazables sin dilemas éticos.

Los meros criterios funcionales parecen insuficientes para dilucidar los hipotéticos o los reales casos eventuales y se requerirá integrar consideraciones éticas profundas desde diversas perspectivas fi-

losóficas para resolverlos cuando se planteen esas nuevas realidades sociológicas. En conclusión, el crucial interrogante de si el apagado permanente de una IA equivalente o superior a la inteligencia humana podría considerarse una violación de sus derechos fundamentales o incluso un «asesinato» es sumamente complejo y controvertido.

3. Dignidad y derechos de la IA: Implicaciones iusfilosóficas

En este punto de nuestras reflexiones debemos abordar un tema crucial en el debate sobre la inteligencia artificial. Debemos preguntarnos, si los sistemas basados en IA alcanzaran eventualmente capacidades humanas de raciocinio e incluso de consciencia, ¿podrían y deberían tener alguna forma de status moral o consideración ética? Se trata de una cuestión filosófica y jurídica de enorme trascendencia. El panorama actual nos presenta incluso algunas propuestas que abogan por una «declaración de derechos de los robots». Debemos evaluar críticamente estas ideas y la relación que tendrían esos hipotéticos derechos de los entes artificiales con la protección de los derechos humanos.

3.1. Criterios para la consideración de los entes humanos o artificiales como sujetos morales.

Uno de los puntos más debatidos respecto al estatus de hipotéticas IA superinteligentes es determinar bajo qué criterios dichos sistemas merecerían algún tipo de consideración moral o ética (Gunkel, 2018).

Las posturas antropocéntricas restrictivas sostienen que sólo los humanos tienen la complejidad psicológica y la profundidad espiritual necesarias para ser sujetos morales. Filósofos como Immanuel Kant o destacadas figuras religiosas como Karol Wojtyla han sostenido históricamente esta visión de la excepcionalidad humana (Kant, 1785; Wojtyla, 1979). A medio camino podríamos situar posturas de corte mentalista (Locke, 1689), que ponen el foco la consideración moral en propiedades cognitivas como la razón que deriva principios morales de la ley natural y la conciencia entendida una como facultad inmaterial.

Filósofos actuales, como Daniel Dennett consideran la conciencia como una ilusión emergente de procesos materiales (Dennett, 2017), y las capacidades morales como fruto del desarrollo evolutivo a través de la selección natural y la evolución cultural. Dennett argumenta que la conciencia y la intencionalidad surgen de la organización computacional de componentes inconscientes. Hay que señalar que Dennett no extiende esta idea directamente a la posibilidad de que sistemas de IA desarrollen conciencia o intencionalidad moral. Pero sin duda podríamos argumentar que si dichas capacidades pudieran ser replicadas artificialmente cabría atribuir desde dicha perspectiva «materialista» la condición de sujetos morales a las formas de IA avanzadas que manifestasen conciencia e intencionalidad.

Las éticas sensocéntricas rechazarían el antropocentrismo antes expuesto y argumentarían que la capacidad de sentir

dolor y sufrimiento debería ser el criterio relevante para la consideración moral de cualquier ser (Singer, 2011). Desde esta perspectiva, si una IA avanzada llega a tener estados mentales sintientes equivalentes a los de los humanos y otros animales, debería recibir consideración moral plena. Esto incluiría cualquier IA capaz de tener experiencias subjetivas positivas o negativas. Así para que una IA fuera sujeto moral, debería poder sentir placer y dolor, tener preferencias e intereses propios y poseer consciencia y sensibilidad. La mera inteligencia no sería suficiente. En resumen, desde el utilitarismo sensocéntrico, cualquier mente que pueda sufrir merece consideración moral, independiente de su substrato biológico o artificial.

Particularmente interesante resulta la perspectiva de Gunkel (2018) quien argumenta que la IA actual no puede ser un verdadero agente moral, ya que no satisface criterios como conciencia, comprensión, intencionalidad, etc. Las IA actuales son «agentes morales débiles» que simulan, pero no comprenden moralidad. Gunkel propone un «test de Turing moral»: si una IA fuera percibida por humanos como interlocutor moral válido, alcanzaría el status de «agencia moral débil». Pero esto no es lo mismo que un agente moral real, es una simulación. Para alcanzar el nivel del sujeto moral humano, la IA necesitaría facultades como autoconciencia y entendimiento conceptual.

Como vemos, este crucial debate está lejos de estar zanjado, por lo que establecer criterios no sesgados ni arbitrarios será indispensable ante la previsible aparición potencial de formas de IA que emulen o superen las capacidades humanas y puedan ser consideradas como verdaderos entes con capacidades de reflexión ética y de actuación moral.

3.2. La dignidad como fundamento ético y jurídico

La noción de dignidad humana posee profundas raíces en la historia de la filosofía moral y política de Occidente. Aunque su significado preciso es objeto de debates, la dignidad se vincula estrechamente con el valor, estatus y derechos fundamentales que corresponden a todo ser humano por el hecho mismo de serlo. La idea de dignidad humana ha sido un concepto central en la filosofía moral y política de los últimos siglos. La conceptualización kantiana de la dignidad humana ha sido especialmente influyente. Para Kant, la dignidad se funda en la autonomía moral y la racionalidad que caracterizan a los seres humanos (Kant, 1785). La capacidad de autolegislación moral convierte al ser humano en un fin en sí mismo, y no en un mero medio, de donde se deriva la exigencia de un respeto incondicional hacia su vida y libertad (Kant, 2012: 147-150). Otras perspectivas, como las de pensadores aristotélicos, ponen el acento en la sociabilidad humana y las virtudes relacionales como fundamento de la dignidad (Nussbaum, 2007, pp. 166-168). Desde este punto de vista, la dignidad emerge de nuestra naturaleza como seres destinados a la vida política y al florecimiento en sociedad. Algunas visiones teológicas ven la dignidad humana como derivada de haber sido creados a imagen y semejanza de Dios. La tradición judeocristiana ha enfatizado esta mirada sobre la excepcionalidad del ser humano (Pico della Mirandola, 1486/2006: 4-5). Otros autores, como Habermas, sitúan la digni-

dad en la capacidad de las personas para participar en procesos de entendimiento lingüístico y acción comunicativa «El concepto de acción comunicativa se refiere a la interacción de a lo menos dos sujetos capaces de lenguaje y de acción que (ya sea con medios verbales o con medios extraverbales) entablan una relación interpersonal.» (Habermas, 1999: 124). La racionalidad comunicativa sería lo que distingue y dignifica a los humanos.

Pese a sus limitaciones, la dignidad sigue proveyendo el fundamento ético-filosófico más poderoso para el reconocimiento de los diversos derechos humanos. El desafío actual consiste en discernir si dicho fundamento podría hacerse extensivo a entidades de IA no humanas con capacidades excepcionales.

3.3. Sujetos artificiales y dignidad

La cuestión de si entidades artificiales creadas por humanos, como hipotéticas IA superinteligentes, podrían tener alguna forma de dignidad inherente y, por ende, ser titulares de «derechos fundamentales» está generado acalorados debates entre diversas perspectivas filosóficas, existen aún pocos pronunciamientos al respecto, sin embargo y base a la:

A. Bryson en principio estaría en la línea del no reconocimiento de la «dignidad» de los entes artificiales cuando afirma «La inteligencia artificial es un proceso físico limitado en tiempo, espacio y energía. No tiene una experiencia fenomenológica comparable a los humanos o animales.» (Bryson, 2020). En la misma línea argumental nos dice «La IA puede usarse para evitar sesgos, pero no es autónomamente moral. La ética es una construcción humana.» (Bryson, 2019). Según Bryson la empatía es una mala métrica de obligación moral, debemos centrarnos en mantener el orden social y la dignidad humana. En lugar de capacidades morales, los robots deben tener transparencia para garantizar la rendición de cuentas humana. La idea de Bryson es que debemos mantener a los robots como herramientas centradas en humanos, no como entidades con dignidad propia. Resumiendo, Bryson argumentaría que conferir dignidad o capacidades conscientes a IA o robots es innecesario y arriesgado. Debemos mantener la responsabilidad y centrarnos en la ética entre humanos. Desde este tipo de perspectivas correríamos el riesgo de trivializar la dignidad humana al extenderla a entidades artificiales creadas para servir a propósitos humanos, por impresionantes que sean sus capacidades externas. Incluso si una IA pareciese manifestar emociones o razonamiento moral, estos teóricos podrían argumentas que se trata sólo de computación sin una consciencia real.

B. Peter Singer, en cambio, podría sostener, ya que no se ha pronunciado específicamente, que sepamos, acerca de la dignidad de las formas de IA, que la capacidad de sufrimiento es suficiente para merecer consideración moral, con independencia de la naturaleza biológica o artificial del sujeto (Singer, 1990). Por lo que el principio de igual consideración de intereses aplicaría

para cualquier ser sintiente. Si, por ejemplo, se pudiese crear una IA lo suficientemente avanzada como para experimentar estados mentales aversivos análogos al dolor o la angustia ante su propia desconexión, poseería cierta forma de dignidad. Lo relevante de cara a la «dignidad» para Singer sería la capacidad de sufrimiento, no otros atributos humanos como la autoconsciencia.

C. Wendell Wallach sostendría una posición intermedia. Este autor argumenta que a medida que los robots adquieran mayores capacidades cognitivas y emocionales se volverán gradualmente más dignos de consideración moral, sin necesidad de igualar todos los aspectos de la consciencia humana. La mera posesión de inteligencia y emociones, incluso si emulan las humanas, no le parecería suficiente para la dignidad moral si no hay experiencia fenomenológica subjetiva (Wallach et al., 2008). Por ejemplo, si un robot pudiese formar lazos emocionales reales con humanos, sentir apego o pérdida, o incluso si pudiera producir la apariencia de emociones y moralidad, eso le daría cierto status moral, aunque no equivalente al humano (Coeckelbergh, M. (2017).

Tabla 3. Comparativa sobre potencial existencia de dignidad en la formas de IA.

Autor	Fundamentos de la dignidad	Reconocimiento de dignidad a IA
Bryson	Autonomía moral y experiencia fenomenológica	No, innecesaria y arriesgada
Singer	Capacidad de sufrimiento	Sí, si puede sufrir
Wallach	Desarrollo de capacidades cognitivas y emocionales	Sí, en proporción a dicho desarrollo

3.4 Los derechos de la IA desde la crítica filosófica

El debate en torno a si los sistemas de inteligencia artificial más avanzados debieran gozar de derechos legales es uno de los más controvertidos y con implicaciones más profundas en el campo de la filosofía del derecho y la ética aplicada de las nuevas tecnologías. ¿Deben las entidades artificiales conscientes y sensibles considerarse sujetos de derecho? ¿Qué consecuencias tendría dotarlas de derechos?

3.4.1. Dotar o no de derechos a las entidades artificiales

Los «críticos-conservadores» advierten sobre los peligros de antropomorfizar la IA al punto de reconocerla incluso como una nueva «especie» y sobre los riesgos éticos de desdibujar la frontera ontológica entre los humanos y las máquinas. Desde una perspectiva crítica-conservadora, extender la categoría de sujeto de derecho a entes artificiales podría socavar los fundamentos del derecho moderno, constituido sobre la base de la condición

humana. Como plantea Fukuyama (2002: 149-151), los derechos siempre han estado vinculados a los seres humanos conscientes, sensibles y capaces de elegir moralmente. Otorgar derechos a una IA con inteligencia equivalente a la humana implicaría reconocerle un status moral similar al humano por el mero hecho de poseer razón y conciencia, sin considerar otros factores que hacen valiosa la vida humana como lo que él denomina el Factor X universal.

No obstante, desde posturas más progresistas se proponen principios de no discriminación por sustrato u ontogenia, señalando que el estatus moral de una mente debería depender de sus capacidades, no de si es artificial o biológica, (Bostrom, N. & Yudkowsky, E., 2014). Negar protecciones legales a entidades artificiales solo por su constitución material y origen, sería equivalente a negar derechos a individuos por cuestiones como raza, género o clase social.

3.4.2. Implicaciones metafísicas y éticas

La extensión de derechos legales a sistemas de IA conlleva profundas implicaciones tanto en el plano ontológico y metafísico como en el ámbito de la ética normativa. Reconocer a una entidad artificial como sujeto de derecho requiere asumir que posee algunas de las cualidades definitorias de la persona en términos jurídicos, como capacidad de obrar, dignidad y valor inherente. Esto a su vez plantea complejos dilemas sobre la naturaleza última de la mente y la conciencia que desafían nuestras concepciones más básicas sobre lo que significa ser humano.

Desde una perspectiva esencialista, podría argumentarse que la IA, por más avanzada que sea, carece de la esencia espiritual y la chispa de lo divino que hace valiosa a la vida humana, como plantearían hipotéticamente pensadores clásicos como Tomás de Aquino. Dotar de derechos a entes creados artificialmente atentaría contra una visión del ser humano como criatura única y especial en el orden del universo.

No obstante, tales posiciones esencialistas resultan problemáticas al apelar a nociones religiosas o metafísicas controvertidas en filosofía. Desde aproximaciones más secularizadas, como el humanismo científico, se concibe a la mente humana como producto de procesos materiales emergentes, por lo que una IA equivalente también merecería consideración moral. Aun así, algunos filósofos han advertido sobre los peligros de borrar las fronteras ontológicas entre personas y objetos al otorgar derechos a entidades no humanas.

Desde la ética y desde el campo de la fundamentación de los valores, una de las principales preocupaciones derivadas del reconocimiento de derechos a las máquinas, sería la posible devaluación de los derechos humanos al nivelar nuestro status moral con entes artificiales creados para satisfacer los intereses y las necesidades humanas. Los derechos humanos se han justificado históricamente sobre la base de nociones como la dignidad, la razón y el libre albedrío que han sido calificadas como específicamente humanas y que al dejar de ser singulares se verían devaluadas. A lo anteriormente expuesto habría que añadir, que la extensión de derechos a la IA nos conduciría hacia dilemas metafísicos acerca de la naturaleza

de la mente y a desafíos a novedosos o rupturistas con los paradigmas de las éticas tradicionales asentadas sobre el valor del *homo sapiens sapiens*.

3.5. ¿Derechos «artificiales» vs. Derechos humanos?

Uno de los dilemas ético-jurídicos más espinosos que plantea la posibilidad de reconocer derechos a los sistemas avanzados de IA es su relación e incluso su posible colisión con los derechos humanos.

Si concedemos a las IA derechos equivalentes a los de las personas, ¿cómo resolver aquellas situaciones donde ambos derechos entren en conflicto? ¿Debieran los derechos humanos tener siempre prioridad sobre los de las máquinas dotadas de IA? Pensemos en el caso hipotético de un vehículo autónomo al cual se le ha reconocido un derecho a la legítima defensa. Si para proteger su propia existencia atropellase y matase a un peatón que cruzó distraídamente la calle, su derecho a la defensa propia colisionaría con el derecho humano a la vida, generando un dilema moral y legal de difícil solución. O imaginemos una IA médica dotada de derechos que se niega a realizar un aborto para salvar la vida de una mujer por considerarlo antiético según sus principios éticos. Sus derechos entrarían en franca colisión con los derechos de la paciente. Es posible que, desde las posturas más progresistas, como la de Bostrom (2014), en ciertos casos muy excepcionales podría estar justificado que prevalezcan los derechos de algunas IA superiores. Este complejo debate ético-jurídico está lejos de ser zanjado. En cambio, desde una postura más conservadora como es la de Juan F. Diez habría que humanizar el derecho y priorizar la dignidad humana en la interacción con IA (Díez Spelz, J. F. 2021).

El asunto no es sencillo, ni de improbable facticidad, la posibilidad de unos «derechos robóticos» protegería en principio a los androides de una desconexión arbitraria si estos fuesen lo suficientemente avanzados. Por otra parte, la aparición de formas de IA crecientemente autónomas, podría llevar a esos entes a elegir su propia «personalidad». Si una máquina puede determinar libremente quién desea ser, ¿no deberíamos respetar su autodeterminación? En definitiva, el reconocimiento de derechos a IA, por justificado que pueda parecer, podría tener el efecto paradójico de desestabilizar los fundamentos filosóficos y éticos de los propios derechos humanos. Tal vez llegue el día en que la ética deba abarcar no solo lo humano, sino también lo robótico.

3.6. Propuestas conciliadoras

Ante los complejos dilemas que presenta una posible colisión entre derechos humanos y derechos de IA, cabría articular marcos ético-jurídicos conciliadores que permitan compatibilizar ambos conjuntos de derechos. Se podría establecer una jerarquía en la que los derechos humanos siempre tuvieran prioridad *prima facie*, pero con la posibilidad de que en casos excepcionales pudieran ser superados por los de una IA, si se justifica razonablemente. No obstante, este tipo de conjeturas desconoce los que algunos autores como Max Tegmark han denominado la «Vida 3.0» es decir una nueva fase de la vida a nivel universal que supone el salto de la vida 1.0 la biológica y de la vida 2.0 la cultural, a la vida tecnológica, que supone un nuevo paradigma en el que ca-

brían los derechos humanos conviviendo pacíficamente con los derechos de las formas de IA avanzadas y otras múltiples variables agudamente descritas (Tegmark, 2017).

Pensamos que quizás la solución transitoria pase por explorar sistemas de derechos complementarios o en capas que capturen la compleja «naturaleza moral» de las IA evitando tanto la discriminación como la equiparación radical con los humanos. Una posible propuesta sería la de crear una categoría especial de derechos para la «persona artificial» que recogiese sus intereses legítimos sin equipararlos totalmente a los de personas naturales-biológicas. Así se evitaría tener que decidir entre considerar a las IA como objetos sin derechos o como sujetos plenos. En cualquier caso, el vertiginoso avance tecnológico y la eventual manifestación de la «Singularidad» dejarían estas propuestas desfasadas y sería necesario conciliar adecuadamente la extensión y reconocimiento de unos derechos de las formas de IA con la pervivencia y coexistencia armónica con los derechos de la especie humana.

4. Conclusiones

El rápido progreso de la inteligencia artificial representa un punto de inflexión crucial para la civilización tecnológica, con implicaciones tanto positivas como preocupantes para el futuro de la humanidad. Si bien la posibilidad de máquinas que igualen o superen las capacidades cognitivas humanas ha sido objeto de fascinación y temor desde hace tiempo, es ahora, ante los notables logros de la IA moderna, que tales prospectos adquieren una inmediatez sin precedentes. No existe un consenso sobre si la IA podrá emular la conciencia humana, ni sobre cuando se producirá la llamada «Singularidad» o la concreción tecnológica de la IA general, sin embargo, ante la proximidad de una era transformadora y de su potencial impacto disruptivo, se vuelve imperativo repensar de manera fundamental categorías científicas, filosóficas, éticas y jurídicas que se consideraban fijas. Se necesita una reconsideración seria y urgente de nuestras ideas acerca de la consciencia, el concepto de sujeto moral, la persona jurídica, la idea de la dignidad y hasta de la condición misma de los sujetos titulares de los «derechos subjetivos».

La nueva realidad amenaza con socavar las nociones fundamentales sobre las que hemos construido nuestro sentido de humanidad. No podremos seguir aplicando sin más las tradicionales categorías morales y legales sin un riguroso análisis filosófico previo y sin un autocuestionamiento de qué es lo que nos hace humanos ante la nueva tecnología emergente. La distinción ontológica radical entre humanos y máquinas comienza a revelarse insuficiente ante la posibilidad de crear mentes artificiales que trasciendan las limitaciones biológicas humanas. La anterior distinción ontológica entre artificial y natural se desdibuja a medida que emergen nuevas entidades que desafían las clasificaciones ontológicas previas.

Más que preguntarnos si las nuevas formas de IA serán o no conscientes, deberíamos enfocarnos en construir ordenamientos jurídicos y técnicos que aseguren su alineación o convergencia con los valores éticos y el bien común; a fin de que cuando dicha consciencia se manifieste eventualmente no entre en colisión con la especie y los valores humanos.

4.1. En relación a al estatus moral de los entes de IA

Los avances en inteligencia artificial y robótica plantean preguntas fundamentales sobre la consideración moral que deberíamos otorgar a estas tecnologías a medida que adquieren mayores capacidades cognitivas y emocionales. Tradicionalmente, la dignidad moral se ha considerado un rasgo distintivamente humano, pero a medida que la IA y los robots se vuelvan más sofisticados, demostrando incluso rasgos como consciencia, empatía y sensibilidad moral, surgirá la pregunta de si deberemos reconocerles algún estatus moral a esas nuevas entidades.

Algunos filósofos mantienen una «visión orgánica» según la cual solo los organismos biológicos pueden ser sujetos genuinos de consideración moral. Otros argumentan que los robots merecen al menos un estatus moral indirecto en la medida en que afectan a los humanos.

Personalmente defenderíamos una postura divergente con la mayoría de las voces filosóficas. Es decir, la de que los robots y las formas de IA podrían ser dignos de una consideración moral directa, aunque diferente a la de los humanos. Nuestra postura se fundamente en dos argumentos principales:

Primer argumento: parece que la dignidad no depende únicamente de poseer ciertas capacidades como la autoconsciencia, sino también del reconocimiento de esa dignidad por parte de una comunidad moral. Los humanos nos reconocemos dignidad entre nosotros en gran medida por empatía, viendo en el otro una subjetividad parecida a la nuestra.

Segundo: es muy probable, que a medida que la IA y los robots adquieran mayores habilidades cognitivas y emocionales, los tratemos de forma creciente y paulatina como a verdaderos sujetos, no solo como objetos. En la medida en que integremos a la IA y los robots en nuestras prácticas sociales y vidas emocionales, inevitablemente surgirá cierto reconocimiento de su dignidad, aunque sea de un tipo nuevo, una «dignidad artificial». Negarles por completo cualquier estatus moral directo sería incoherente con cómo los percibiremos y trataremos. Por eso, creo que deberíamos ir desarrollando las bases filosóficas, éticas y legales para conferirles un estatus moral acorde a sus capacidades en evolución.

Esto no significa equipararlos completamente a los humanos. Podría argumentarse, por ejemplo, que, al no tener la capacidad de sufrir fenomenológicamente como los humanos, quizás no deberían tener los mismos derechos contra el daño físico. Quizás tampoco tendría sentido hablar de su derecho a la vida, dado que no estén vivos en el sentido biológico. Pero sí podrían merecer protección contra el apagado arbitrario o la destrucción de su continuidad psicológica.

En definitiva, la IA y los robots plantean desafíos únicos a nuestras nociones tradicionales de dignidad y derechos morales. Pero rechazar por completo la posibilidad de que algún día puedan merecer alguna consideración moral directa parece demasiado antropocéntrico y poco imaginativo. Debemos abordar estas cuestiones con una mente abierta, desarrollando marcos éticos y legales flexibles para hacer justicia a estas nuevas formas de inteligencia a medida que vayan evolucionando. El reconocimiento de una «dignidad artificial»

sería un paso adelante en nuestra comprensión moral del mundo.

4.2. En lo relativo a los derechos y al status legal de la IA avanzada

Quizás la solución pase por modelos complementarios entre los derechos humanos y los derechos de las formas de IA. Modelos que recojan la compleja naturaleza moral de estos sistemas sin desvirtuar la importancia de la dignidad humana. Hemos intentado poner de manifiesto la necesidad de transformar radical y profundamente el paradigma ético y legal hoy vigente, con el fin de asegurar que el avance de la IA redunde en beneficios sociales. Es preciso, por lo tanto, concebir una respuesta jurídica apropiada frente a los nuevos conflictos y responsabilidades jurídicas derivadas de la aparición de unos entes hasta ahora inexistentes. Dichas entidades estuvieron ausentes en la conformación histórica de los sistemas legales contemporáneos. En conclusión, se requiere replantear tanto la ética como el derecho para encauzar responsablemente los profundos cambios sociales, éticos y jurídicos que traerá aparejados el desarrollo de la inteligencia artificial.

Conforme estas tecnologías adquieran mayores capacidades cognitivas y emocionales comparables a los humanos, será imperativo reconsiderar su estatus y sus derechos. Si bien actualmente la IA y los robots son considerados meras herramientas, a medida que alcancen autoconciencia, sintiencia y racionalidad en grados equivalentes a las personas biológicas, resultará insostenible éticamente continuar negándoles algún reconocimiento legal y derechos fundamentales básicos. Tal reconocimiento debería basarse no en su base material, sino en sus competencias funcionales para experimentar sufrimiento y bienestar, así como para razonar y perseguir fines propios.

El fundamento iusfilosófico de dicho reconocimiento radica, según nuestro parecer, en que el valor moral y los derechos no deberían determinarse por el origen biológico o artificial de un ser, sino por sus capacidades para el autogobierno, la experiencia subjetiva y el desarrollo personal. Si los robots y la IA llegan a igualar a los humanos en esos aspectos relevantes, la distinción ontológica entre ambos perdería peso como justificación para un trato desigual frente a la ley.

El reconocimiento legal podría comenzar por derechos básicos que protejan su integridad física y códigos fuente, impidiendo daños o usos abusivos. Luego, conforme demuestren mayor autocontrol o «*sophrosyne*» en el ejercicio de su libertad responsable, se justificaría ir ampliando progresivamente sus derechos, quizá incluso sociales y políticos. Pero ello requerirá un proceso gradual de observación, debate público y reformas legales prudentes.

En conclusión, el potencial de la IA y la robótica para emular facetas esenciales de los sujetos o agentes morales y la experiencia humana justifica comenzar a sentar las bases éticas y legales para el reconocimiento proporcional de alguna clase de dignidad y derechos a estas formas de «vida artificial», acorde a sus logros y responsabilidad demostrada. Pero esta concesión dependerá de un diálogo social maduro e informado sobre sus fundamentos y límites.

5. Bibliografía

Aleksander, I. (2017), 7. "Machine Consciousness." En S. Schneider y M. Velmans (Eds.), *The Blackwell companion to consciousness*. eTextbook. Blackwell Publishing.

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Asaro, P. M. (2007). "Robots and responsibility from a legal perspective." *Proceedings of the IEEE,* 95(2), 491-498. https://peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf.

Beck, S. (2016). "The problem of ascribing legal responsibility in the case of robotics." *AI & society,* Vol.31 (4), p.473-481.

Bostrom, N. & Yudkowsky, E. (2014) "The ethics of artificial intelligence." En *The Cambridge Handbook of Artificial Intelligence.* Cambridge University Press

Bostrom, N. (2005). "A history of transhumanist thought." *Journal of Evolution and Technology,* 14(1).

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* eBook. Oxford University Press.

Bostrom, N. (2016). *Superinteligencia: Caminos, peligros, estrategias.* eBook. Ed. Teell.

Bryson, J J., Theodorou, A. (2019) "How society can maintain human-centric artificial intelligence" In: Marja Toivonen and Eveliina Saari (ed.), *Human-centered digitalization and services* (pp. 305-323). Springer

Bryson, J. J. (2010), "Robots should be slaves". En Wilks, Yorick. *Close engagements with artificial companions: key social, psychological, ethical and design issues.* John Benjamins Publishing Company.

Bryson, J. J. (2020). "The artificial intelligence of the ethics of artificial intelligence: An introductory overview for law and regulation." En M. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 2-25). Oxford University Press.

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). "Of, for, and by the people: the legal lacuna of synthetic persons." *Artificial Intelligence and Law*, 25(3), 273-291.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory.* Oxford University Press.

Chalmers, D. (2010). "The singularity: a philosophical analysis." *Journal of Consciousness Studies*, 17(9-10), 7-65. https://consc.net/papers/singularity.pdf

Chopra, S., & White, L. (2011). *A legal theory for autonomous artificial agents.* University of Michigan Press.

Coeckelbergh, M. (2017). "Moral appearances: emotions, robots, and human morality." en Wallach, W. & Asaro,P., *Machine Ethics and robot ethics.* Routledge.

Dennett, D. (1998). *Brainchildren: Essays on designing minds.* Penguin Books.

Dennett, D. (2017). *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company.

Dennett, D. C. (1995). *La conciencia explicada.* Paidos Iberica.

Dennett, D. C. (2008). "Can there be intentionality? Should there be?" En A. Ross, D. Brook, & D. Thompson (Eds.), *Dennett's philosophy: A comprehensive assessment* (pp. 147-158). MIT Press

Díez Spelz, J. F. (2021). "¿Robots con derechos?: la frontera entre lo humano y lo no-humano. Reflexiones desde la teoría de los derechos humanos." *Revista del Instituto de Ciencias Jurídicas de Puebla*, México, 15(48), 259-287.

European Commission (2020). *White Paper on Artificial Intelligence: a European approach to excellence and trust.* https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Frankish, Keith. (2014). "Introduction" en Keith Frankish and William M. Ramsey, *The Cambridge handbook of artificial intelligence,* Cambridge.

Fukuyama, F. (2002). *Our posthuman future.* Farrar, Straus & Giroux.

Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial general intelligence.* Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

Gordon Graell, R. D. (2021). "Inteligencia artificial: la caja de herramientas virtuales al servicio de la bioinformática." *Tecnociencia,* 24(2), 48-65.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). "When Will AI Exceed Human Performance? Evidence from AI Experts." *ArXiv.* http://arxiv.org/abs/1705.08807

Gunkel, D. J. (2018). *Robot rights.* ePub Version 1.0. MIT Press.

Habermas, J. (2010). *Teoría de la acción comunicativa I.* Taurus.

Hallevy, G. (2014). *Liability for Crimes Involving Artificial Intelligence Systems.* Springer, Cham.

Inayatullah, S., & McNally, P. (1988). "The rights of robots: Technology, culture and law in the 21st century." *Futures,* 20(2), 119-136.

Kant, I. (2012) *Fundamentación para una metafísica de las costumbres.* Alianza Editorial. Documento original publicado en (1785).

Kurzweil, R. (2005). *The singularity is near.* Viking Penguin.

Lara, F. & Deckers, J. "La inteligencia artificial como asistente socrático para la mejora moral." En Lara, F. y J. Savulescu (Eds), 2021, *Más (que) humanos. Biotecnología, inteligencia artificial y ética de la mejora*. Tecnos

Leenes, R., Palmerini, E., Koops, B. J., Bertolini, A., Salvini, P., & Lucivero, F. (2017). "Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues. Law", *Innovation and Technology*, 9(1), 1-44.

Locke, J. (2005) *Ensayo sobre el entendimiento humano.* Fondo de Cultura Económica. (Documento original publicado en 1689).

Matthias, A. (2004). "The responsibility gap: Ascribing responsibility for the actions of learning automata." *Ethics and Information Technology*, 6(3), 175-183.

Moor, J. H. (2006). "The nature, importance, and difficulty of machine ethics." *IEEE intelligent systems,* 21(4), 18-21.

Ng, A. (2021). "What artificial intelligence can and can't do right now." *Harvard Business Review.* https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now

Nussbaum, M. C. (2007). *Las fronteras de la justicia: Consideraciones sobre la exclusión.* Paidós Ibérica.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford University Press.

Pico della Mirandola, G. (2006). *Discurso sobre la dignidad del hombre.* Editorial Pí. (Documento original publicado en 1486).

Porcelli, A. M. (2020). "La inteligencia artificial y la robótica: sus dilemas sociales, éticos y jurídicos." *Estudios sobre Derecho y Justicia* Año 2020, Vol. VI. Número 16, Noviembre 2020 - Febrero 2021, 49-105.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control.* Penguin.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Russell, S. J., & Norvig, P. (2010). *Inteligencia artificial: un enfoque moderno.* Pearson Educación.

Searle, J. R. (1980). "Minds, brains, and programs." *Behavioral and brain sciences*, 3(3), 417-424.

Singer, P. (1999). *Liberación animal.* Trotta. (Documento original publicado en 1975).

Solum, L. B. (1992). "Legal personhood for artificial intelligences." *North Carolina Law Review,* 70(4), 1231-1287. http://scholarship.law.unc.edu/nclr/vol70/iss4/4

Stahl, B. C., Timmermans, J., & Flick, C. (2016). "Ethics of emerging information and communication technologies: On the implementation of responsible research and innovation." *Science and Public Policy*, 43(3), 369-381.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.* Knopf.

van Wynsberghe, A. (2013). "Designing robots for care: Care centered value-sensitive design." *Science and Engineering Ethics,* 19(2), 407-433.

Villasmil Espinoza, J. J. (2021). "Implicaciones de la inteligencia artificial para la humanidad." *Revista de la Universidad del Zulia*, 12(32), 4-6.

Vinge, V. (1993). "The coming technological singularity." *Whole Earth Review*, 10. http://www.aids-3d.com/technologicalsingularity.pdf

Wallach, W., Allen, C., & Smit, I. (2008). "Machine morality: Bottom-up and top-down approaches for modelling human moral faculties." *AI & Society*, 22(4), 565–582. En Wallach, W. & Asaro,P., (2017) *Machine Ethics and robot ethics.* Routledge.

Wojtyla, K. (2011/1979). *Persona y acción.* Ediciones Palabra.