L'IRCOCERVO
PRIMA RIVISTA ELETTRONICA ITALIANA DI METODOLOGIA GIURIDICA - TEORIA GENERALE DEL DIRITTO E DOTTRINA DELLO STATO
FIRST ITALIAN DIGITAL JOURNAL OF LEGAL METHODOLOGY - GENERAL THEORY OF LAW AND DOCTRINE OF THE STATE

Ética de la virtud para una IA compasiva

María Teresa García-Berrio

Universidad Complutense de Madrid

Abstract: **Virtue Ethics for a Compassionate AI**
As Artificial intelligence systems permeate the social fabric of contemporary communities, ethical considerations surrounding their impact on human well-being come to the fore. Academia is calling for the integration of virtue ethics into the design of AI systems, which plays a critical role in shaping a landscape of compassionate and human-friendly AI. The aim of study is to embark on a philosophical journey on Virtue ethics – more precisely, kindness, empathy and compassion – as applied to the field of Artificial intelligence (AI).

Keywords: AI, Virtue, Ethics, Emotions, Nussbaum, Algorithmic Biases.

Resumen: 1. Inteligencia Artificial y Condición humana – 2. La ruptura del Antropocentrismo ante la irrupción de la IA – 3. Emocionalidad, Vulnerabilidad y Virtud frente a la Tecnología – 4. Hacia una IA Compasiva – 5. Principio de Beneficencia ante el riesgo de discriminación por sesgos algorítmicos – 6. Principio de No-maledicencia: La prevención del daño y la preservación de la dignidad humana ante el riesgo de alexitimia virtual.

1. Inteligencia Artificial y Condición humana

La llamada Inteligencia Artificial (en lo sucesivo, IA) forma parte de ese conjunto de tecnologías "disruptivas" que están transformando radicalmente nuestro mundo a través de la convergencia de las llamadas tecnologías NBIC (NANO/BIO/INFO/COGNO) – es decir, nanotecnologías, biotecnologías, tecnologías de la información y ciencias cognitivas – las cuales están evolucionando a un ritmo imparable. La IA procesa, cruza y reutiliza ingentes cantidades de datos mediante algoritmos. Un algoritmo es una lista más o menos larga de instrucciones; dicho en otros términos, un conjunto ordenado y finito de pasos que puede emplearse para hacer cálculos, resolver problemas y alcanzar decisiones.

La interacción de los seres humanos con las tecnologías disruptivas está acelerando nuestra configuración como entornos socio-técnicos, donde se difuminan las fronteras entre los sujetos humanos y la tecnología y donde los seres humanos trabajamos con artefactos virtuales en una suerte de simbiosis entre la inteligencia humana y la artificial (*Human-Machine-Interaction*). En este

sentido, los retos a los que nos enfrenta la aplicación de sistemas de IA, nos sitúan forzosamente ante una nueva andadura de colaboración transversal entre especialistas en ética, tecnólogos y responsables políticos con el objeto de que, desde un prisma interdisciplinar, seamos capaces de generar los cauces adecuados para el progreso ético de la IA y, en un segundo estadio, de construir puentes para la promoción de los valores humanos aplicados a estas *tecnologías disruptivas*.

La IA es una disrupción antropológica sin precedentes, con un impacto directo en todas las estructuras naturales, económicas y sociales de las comunidades humanas. Los últimos avances tecnológicos han supuesto un revulsivo para nuestra consciencia ética frente a la ilusión de la autonomía individual postmoderna, fuertemente marcada por el elevado precio pagado en las sociedades industrializadas por el proceso conocido como *individualización* como marca sustantiva de la "modernidad reflexiva" que propugnan algunos de los sociólogos y filósofos contemporáneos más destacados como Gilles Lipovetsky, Elizabeth Beck o Zygmunt Bauman, quienes a su vez se han nutrido de la tradición sociológica de Norbert Elias y Émile Durkheim.

En la obra *Modernidad líquida*, Zygmunt Bauman conceptualiza la llamada "individualización" como aquel proceso que consiste "en hacer que la identidad humana deje de ser un 'dato' para convertirse en una tarea, y en cargar sobre los actores la responsabilidad de dicha tarea y de las consecuencias (y efectos secundarios) de su actuación"[1].

El término "individualización" se refiere pues a aquel proceso social que es consecuencia, por un lado, del desarrollo extremo del individualismo que ha caracterizado a la modernidad tardía de la segunda mitad del siglo XX – calificada en términos sociológicos como "modernidad reflexiva" – y, por otro, del triunfo de la lógica libertaria en las primeras décadas del siglo XXI, en base a la cual los individuos son considerados como dueños y responsables absolutos de su vida a través de un proceso de reflexión que premia ante todo el desarrollo de las capacidades de autodeterminación.

Aplicada a los procesos tecnológicos, la "individualización" supone para algunos sociólogos, como Ulrich Beck, una transformación radical de la estructura de la personalidad propia de las sociedades, ya que se hace creer al individuo aislado que es factible su liberación del encorsetamiento de las estructuras sociales tradicionales y que, por ende, puede disponer de un control completo sobre el desarrollo de su vida a través de las decisiones que toma en los procesos de racionalización tecnológica[2]. Sin embargo, el efecto de esta "individualización" es devastador, ya que arranca al individuo de la comunidad, desgastando sus vínculos de confianza hasta dejarle sin protección en un mundo virtual en el que cada vez es más difícil el pleno desarrollo autónomo de las

[1] Z. Bauman, *Modernidad Líquid*a, Fondo de Cultura Económica de Argentina, Buenos Aires, 2003, p. 20.
[2] U. Beck, *La sociedad del riesgo mundial: En busca de la seguridad perdida*, Paidós, Barcelona, 2008.

personas y cuyo reverso es el riesgo asociado al uso masivo de medios digitales para la vigilancia y el control social[3].

La IA se sustancia en una sofisticada combinatoria de instrumentos y dispositivos altamente interconectados que recogen enormes cantidades de información de todos los objetos digitales que nos rodean en el día a día – el llamado "Internet de las Cosas" – y que registran todos nuestros datos a través de dispositivos adheridos o incorporados al cuerpo que se despliegan en escenarios empresariales, educativos y recreativos. Lo paradójico es pues que la mayor parte de la información de la que se nutre la IA es generada por los propios usuarios – por nosotros mismos – a través de la interacción con los servicios basados en Internet y telefonía móvil. Como señala el tecno-filósofo Txetxu Ausín

> […] la información no solo se obtiene de registros, públicos o privados, sino que muchas veces se consigue de fuentes abiertas (redes sociales) y de transacciones electrónicas que no relacionamos en términos de investigación: actualizar información de una app, usar o simplemente llevar encima un teléfono móvil (geolocalización), participar en medios sociales como Facebook o Twitter, el registro de viajeros o simplemente moverse por el espacio público[4].

En este mismo sentido se pronuncia la famosa socióloga norteamericana Shoshana Zuboff cuando nos advierte del advenimiento de un modelo de capitalismo – al que califica como *Capitalismo de la vigilancia* – en el que los procesos automatizados realizados por medios digitales y tecnológicos han ido reemplazado progresivamente en el aprendizaje social de las personas los vínculos interpersonales propios de las relaciones humanas por algoritmos que buscan erradicar el sentimiento de "interdependencia humana". Y es ahí precisamente donde reside, a juicio de esta autora, *el oscuro corazón del capitalismo de la vigilancia* que puede llegar a tener efectos desgarradores tanto en la biografía personal de las personas, como en su comprensión de la realidad social y de las normas cognitivas. Por todo ello, admite Zuboff, la manera de generar y recibir información desde medios digitales – tan inmediata en tiempo, tan difusa en fuentes y tan automática en su origen – está logrando alterar la propia capacidad crítica y cognitiva del receptor de la información. A este tipo de poder lo califica esta autora como "poder instrumentario para una tercera modernidad" o *instrumentarismo*, construido a partir de la combinación de los términos predicción, monetarización y control.

> En esa definición – apunta esta autora – la instrumentación hace referencia al títere: la arquitectura material ubicuamente conectada con la computación

[3] U. Beck, E. Beck-Gernsheim, *La individualización: El individualismo institucionalizado y sus consecuencias sociales y políticas*, Paidós, Barcelona, 2003.
[4] T. Ausín, "¿Por qué ética para la Inteligencia Artificial? Lo viejo, lo nuevo y lo espurio", in *Sociología y Tecnociencia*, 11 (2021), n. 2, pp. 2-3.

> sensible que transfiere, convierte, interpreta y acciona la experiencia humana. La instrumentalización, por su parte, denota las orientaciones sociales que orientan a los titiriteros hacia la experiencia humana cuando el capital de la vigilancia se vale de las máquinas para transformarnos en medios de los fines mercantiles de otros[5].

Se hace pues imperativo el abordaje ético de las tecnologías asociadas a la IA, identificando para ello, por un lado, los daños y peligros a evitar y promocionando, por otro, aquellos valores de *interdependencia humana* que sean capaces de establecer un ecosistema de confianza para los ciudadanos y usuarios frente al empleo "potencialmente nocivo" de la IA. En esta línea de actuación, la Unión Europea ha apostado desde hace años por una estrategia general de investigación e innovación responsable en materia de tecnociencias – la cual responde a las siglas RRI (*Responsible, Research & Innovation*) – *para describir los procesos de investigación científica y desarrollo tecnológico que tienen en cuenta los efectos y posibles impactos sobre el medio ambiente y la sociedad de las tecnologías*. Se trata de una iniciativa sin precedentes que busca instaurar un nuevo modelo de gobernanza de la investigación capaz de reducir la brecha entre la comunidad científica y la sociedad, incentivando para ello procesos de socialización de los entornos tecno-científicos en los que la cooperación entre la sociedad civil y los tecnólogos permita alinear el proceso de investigación científica con los valores, necesidades y expectativas de la sociedad. En concreto, la llamada *RRI* comprende seis líneas de actuación: (i) participación ciudadana a lo largo del proceso de investigación; (ii) igualdad de género en los equipos de trabajo; (iii) educación científica para mejorar los procesos educativos y promover vocaciones científicas entre los más jóvenes; (iv) concienciación ética para fomentar la integridad científica, con el fin de prevenir y evitar prácticas de investigación inaceptables; (v) acceso abierto a la información científica para mejorar el diálogo abierto con la sociedad y (vi) acuerdos de gobernanza, con el objeto de proporcionar herramientas que fomenten la responsabilidad compartida entre grupos de interés e instituciones.

Esta estrategia de investigación e innovación *éticamente responsable* en materias de tecno-ciencias ha cobrado especial protagonismo en los últimos años debido a la presentación, con fecha de 21 de abril de 2021, de la *Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificia*l (lo que se ha venido a conocer en nomenclatura anglosajona como *Artificial Intelligence Act*). Esta propuesta legislativa supone un hito normativo a nivel comunitario, que tiene por objeto estructurar un marco regulatorio de la IA (i) *transnacional*, ya que supone la primera regulación jurídica transversal que es directamente aplicable en todos los Estados miembros de la Unión Europea sin necesidad de elaborar para ello las

[5] S. Zuboff, *La era del capitalismo de la vigilancia: La lucha de un futuro humano frente a las nuevas fronteras del poder*, Paidós, Barcelona, 2020, p. 439.

subsiguientes normas de transposición nacional; (ii) *con eficacia universal*, ya que se trata de un sistema regulatorio universal que extiende su ámbito de aplicación a todos los sistemas de IA que funcionan como componentes de productos o, que siendo productos en sí mismos, se pretenden comercializar en el mercado de la Unión European *y* (iii) orientado al control y la gestión del riesgo.

En efecto, la Propuesta de Reglamento europeo en IA incorpora un sistema basado en la gestión del riesgo que establece diferentes obligaciones de información para los proveedores de sistemas en función del nivel de riesgo asociado al sistema de IA respecto a las garantías de los derechos fundamentales de los usuarios; distinguiéndose así entre: (iii.1) riesgo limitado, (iii.2) alto riesgo y (iii.3) riesgo inaceptable. (iii.1) El Reglamento establece una primera obligación de *información mínima* para aquellos sistemas de IA considerados de "riesgo bajo" o "riesgo limitado". En concreto, estos sistemas de IA deben cumplir unos requisitos mínimos de transparencia que permitan a los usuarios, tras interactuar con las aplicaciones, tomar decisiones con conocimiento de causa y bajo su consentimiento. Por tanto, estaríamos ante un sistema de IA de "riesgo limitado" cuando los usuarios son conscientes de que los contenidos de imagen, audio o vídeo que se les ofrecen han sido generados por una aplicación de IA. A este respecto, en la propuesta de Reglamento se incluyó una mención especial para las aplicaciones de IA generativa, como ChatGPT, para las que se establece que deberán cumplir con requisitos adicionales de transparencia de cara a poder ser catalogadas dentro de las aplicaciones de "riesgo limitado". En concreto, el Reglamento impone como requisitos específicos a los sistemas de IA generativa: (a) que revelen siempre al usuario que el contenido ha sido generado por una IA, (b) que se obliguen a publicar resúmenes periódicos de los datos protegidos por derechos de autor empleados para el entrenamiento y (c) que prohíba la difusión de contenidos ilegales. (iii.2) Aquellos sistemas de IA que pudieran afectar negativamente a la seguridad o a la garantía y salvaguarda debidas de los derechos fundamentales son considerados en el Reglamento como sistemas de "alto riesgo". La Propuesta de Reglamento europeo en IA distingue a ese propósito entre dos categorías de sistema de "alto riesgo". (iii.2.1) En primer lugar, los sistemas de IA que se utilicen en productos sujetos a la legislación de la UE sobre seguridad de los productos de consumo (por ejemplo juguetes, aviación, automóviles, dispositivos médicos o ascensores). (iii.2.2) En segundo lugar, los sistemas de IA que permitan (a) la identificación biométrica y categorización de personas físicas, (b) la gestión y explotación de infraestructuras críticas, (c) la educación y formación profesional, (d) el empleo, gestión de trabajadores y acceso al autoempleo, (e) el acceso y disfrute de servicios privados esenciales y servicios y prestaciones públicas, (f) la gestión de la migración, el asilo y el control de fronteras y (h) la asistencia en la interpretación jurídica y la aplicación de la ley. Todos estos sistemas deberán ser evaluados antes de su comercialización y a lo largo de su ciclo de vida. (ii.3) Por último, aquellos sistemas de IA que supongan una amenaza directa para las personas y para la garantía de sus derechos fundamentales son considerados en el Reglamento como

sistemas de "riesgo inaceptable" y serán prohibidos. Dicha prohibición se extiende a tres modalidades esenciales de sistemas de IA: (iii.3.1) Sistemas de manipulación cognitiva del comportamiento de personas o grupos vulnerables, como la infancia y adolescencia [por ejemplo, es el caso de posibles juguetes animados que pudieran fomentar comportamientos peligrosos en los niños o sistemas de IA que pudieran inducir a comportamientos suicidas en los adolescentes]. (ii.3.2) Sistemas que emplean algoritmos para la generación de sesgos identitarios que permitan clasificar a las personas en función de su estatus socioeconómico o características personales – tales como la raza, el género, la nacionalidad, la orientación sexual, la religión, etc. (ii.3.3) Por último estarían aquellos sistemas de identificación biométrica, tanto en tiempo real como a distancia, que emplean el reconocimiento facial.

A pesar del refrendo sustancial que la Propuesta de Reglamento ha obtenido en la Eurocámara, una de las cuestiones que mayor controversia que ha suscitado durante la fase de su tramitación legislativa ha sido precisamente la de establecer un rango de prohibición adecuado para aquellos sistemas de IA que constituyen un "riesgo inaceptable" para la garantía de los derechos fundamentales de las personas y para la salvaguarda de principios éticos. En este sentido, las arduas discusiones parlamentarias acontecidas entre los Grupos de expertos durante el proceso de tramitación legislativa han dado como resultado la ampliación del listado de sistemas de IA que deben ser considerados prohibidos a cinco nuevas modalidades: (a) los sistemas de identificación biométrica remota "en tiempo real", cuando se realiza en espacios de acceso público que permitirían una vigilancia masiva; (b) los sistemas de identificación biométrica remota "en diferido", con la única excepción de que el empleo de dichos sistemas fueran realizados por las fuerzas y cuerpos de seguridad del Estado para la persecución de delitos graves y mediante autorización judicial previa; (c) los sistemas de previsión de riesgo de comisión de ilícitos penales o administrativos; (d) los sistemas de IA predictivos que permiten inferir las emociones de una persona física en los ámbitos de la aplicación del Derecho y la gestión de fronteras, en lugares de trabajo y en centros educativos y (e) los sistemas que emplee técnicas subliminales más allá de la conciencia de una persona con el fin de distorsionar materialmente el comportamiento de la misma.

Asimismo, especial atención requiere el tratamiento que ofrece el artículo 5.1.a) de la Propuesta de Reglamento europeo en materia de IA en su última redacción cuando, entre los sistemas de IA prohibidos, recoge expresamente: "[…] la comercialización, puesta en servicio o uso de un sistema de IA que emplee técnicas subliminales más allá de la conciencia de una persona o técnicas deliberadamente manipuladoras o engañosas, con el objetivo o el efecto de distorsionar materialmente el comportamiento de una persona o de un grupo de personas, al afectar considerablemente la capacidad de la persona para tomar una decisión informada, causando así que la persona tome una decisión que de otro modo no habría tomado de una manera que cause o sea probable que cause daño significativo a esa persona, a otra persona o a un grupo de personas. La

prohibición de un sistema de IA que emplea técnicas subliminales mencionada en el primer párrafo no se aplicará a los sistemas de IA destinados a ser utilizados con fines terapéuticos aprobados, siempre y cuando se obtenga un consentimiento informado específico de las personas expuestas a ellos o, en su caso, de su tutor legal"; la trascendencia ética de esta cláusula normativa es sustancial ya que cualquier tentativa de intervención en nuestros procesos mentales profundos o inconscientes mediante técnicas subliminales que van más allá de la conciencia de una persona, o cualesquiera técnicas deliberadamente manipuladoras o engañosas que se empleen en dispositivos de IA con el propósito de influenciar en nuestras decisiones como usuarios o consumidores sobre qué comprar, qué consumir, qué despreciar o qué apreciar, debe ser prohibida y declarada nula de pleno derecho.

Pese a las numerosas deficiencias en que incurre la *Propuesta de Reglamento del Parlamento Europeo y del Consejo en materia de Inteligencia artificial*, se vislumbra un hilo de esperanza al haber introducido en su clausulado una nueva garantía de *supervisión humana* que emplaza a aquellas personas físicas que ejerzan la gestión de los sistemas de IA a que sean conscientes del riesgo de sesgo, de automatización o confirmación que este tipo de aplicaciones supone. En este sentido, la Eurocámara exhorta a los gestores de los sistemas de IA a que cumplan con la obligación legal "de proporcionar especificaciones para los datos de entrada o cualquier otra información pertinente en cuanto a los conjuntos de datos utilizados en los sistemas de IA, incluida su limitación y supuestos, teniendo en cuenta la finalidad prevista y el mal uso previsible y razonablemente previsible del sistema".

Si aceptamos pues que el inconsciente humano merece ser protegido como bien jurídico, no basta con prohibir sólo aquellas técnicas subliminales deliberadamente engañosas o manipuladoras empleadas por los sistemas de IA con fines lucrativos, las cuales causan una afectación considerable sobre la capacidad para tomar una decisión informada en los sujetos. Como nos revela Ignasi Beltrán de Heredia en su obra *Inteligencia artificial y neuroderechos: La protección del yo inconsciente de la persona*[6], el escenario descrito en el artículo 5.1.a) del la propuesta de Reglamento europeo nos obliga a plantearnos si el concepto de "acto propio y voluntario" queda en entredicho y, de forma derivada, si también impacta en el de responsabilidad social. Por ello, si estamos hablando de herramientas efectivas que trascienden la conciencia humana, es posible que necesitemos no sólo un marco jurídico que nos dé amparo frente a quienes quieran aprovecharse de nuestros actos por debajo de dicho umbral de la consciencia, sino, lo que es más importante, precisamos de un marco de protección favorable a la condición humana en el seno de los sistemas de IA. A este respecto, el reconocimiento del papel clave que juega la *intersubjetividad humana* ante las aplicaciones de IA – sobre todo, tratándose de IA generativa – requiere cuatro

[6] I. Beltrán de Heredia, *Inteligencia artificial y neuroderechos: La protección del yo inconsciente de la persona*, Aranzadi, Madrid, 2023.

líneas de acción prioritarias. (i) En primer lugar, se debe apostar por el desarrollo de un marco legislativo internacional consensuado que subordine el diseño, producción y desarrollo de la IA a la dignificación de la condición humana. (ii) En segundo lugar, una vía prioritaria de protección de la condición humana mediante la penalización de cualquier tentativa de *instrumentalización* a través de la irrupción en los procesos mentales profundos de técnicas subliminales que van más allá de la conciencia de una persona, o cualesquiera técnicas deliberadamente manipuladoras o engañosas que se empleen por la implementación de aplicaciones de IA, con el objeto de transformar la conciencia ética y crítica de las personas para ser y hacer libremente. (iii) En tercer lugar, se debe apostar por el reconocimiento con carácter universal de un *derecho humano a la preservación de la mente inconsciente* como base irrenunciable de la individualidad humana y fuente de validez del consentimiento libre exigible en cualquier acto jurídico. (iv) Por último, dado que es altamente probable que la gestión de los sistemas de IA sean monopolizados en los próximos años por elites políticas y económicas y que, en consecuencia, el acceso a estas herramientas tecnológicas acrecienten las desigualdades y los mecanismos de exclusión social, se deben adoptar todas la medidas normativas necesarias para perseguir y castigar el uso de la IA como dispositivo de control y manipulación social.

La implantación masiva de sistemas de IA en los próximos años nos situará inevitablemente ante la compleja encrucijada del transhumanismo y ante el riesgo exponencial de cuestionarnos – una vez más en la historia de la filosofía moral – sobre la veracidad del paradigma filosófico y ontológico antropocéntrico, según el cual, el ser humano es el único ser en el mundo dotado de conciencia y voluntad para actuar con libertad y que, en consecuencia, el ser humano es el único que merece un tratamiento ético. Cabe pues preguntarse, a continuación, *¿si la IA expone la consciencia colectiva de la responsabilidad social el elevado precio de la individualización (Zuboff)?¿Si estamos ante la ruptura del Antropocentrismo?* Y lo que es más relevante, *¿si es basta con apelar al principio de responsabilidad colectiva (Hans Jonas) para favorecer un marco ético sustentado en la interdependencia humana frente a la ilusión de la autodeterminación y autosuficiencia que fomenta la IA?*

2. La ruptura del Antropocentrismo ante la irrupción de la IA

Las éticas tradicionales son netamente antropocéntricas – también calificadas como *éticas de la proximidad* – cuyas normas están circunscritas más al ámbito de lo personal. Se trata pues de éticas en la que la realidad de la persona y su condición son consideradas como constantes en su esencia y para las que el alcance de las prescripciones éticas se sustancia tradicionalmente en relaciones de intersubjetividad humana y de conexión con el prójimo. Como afirmara Martin Luther King en la carta que escribiera 16 de abril de 1963 durante su estancia en la cárcel de Birmingham: “Estamos atrapados en una *red ineludible de*

reciprocidad, ligados en el tejido único del destino. Cuando algo afecta a una persona de forma directa, afecta indirectamente a todas"[7].

Todas las éticas tradicionales dan cuenta de tres premisas básicas: (i) la condición humana permanece inmutable para siempre; con base en el presupuesto anterior, (ii) se puede determinar con certeza el bien humano y, por último, (iii) el alcance de la acción humana y de su consecuente responsabilidad es susceptible de ser delimitado.

Sin embargo, con la irrupción de la tecnología el esquema clásico de las "éticas de proximidad" se transforma completamente, ya que las consecuencias que se derivan del empleo de técnicas disruptivas superan con creces a las tres premisas básicas de las éticas tradicionales, obligándonos a cuestionarnos de nuevo acerca del "principio de responsabilidad" que tratara Hans Jonas en su *Ensayo de una Ética para la civilización tecnológica*[8]. Como apunta este autor, la intervención tecnológica cambió drásticamente el componente de ineludible reciprocidad de la "ética de la proximidad" al poner la naturaleza al servicio del hombre, quien la alterará radicalmente en tan sólo cuatro décadas con efectos nocivos y de degradación inimaginables. Así pues, el nuevo poder de la acción humana sobre el mundo natural supone, a juicio de Jonas, la transformación de la propia naturaleza de la ética. A partir del momento en que el ser humano tiene el poder material de destruir la humanidad adquiere una nueva responsabilidad de cara a las generaciones futuras*: la responsabilidad por lo venidero*. Siendo que, en el ámbito de la ciencia y tecnología, la *responsabilidad hacia lo venidero* es siempre *intergeneracional* y *colectiva*.

Son múltiples los sentidos tradicionales asociados al término de "responsabilidad": (i) sea responsabilidad como imputación causal de los actos cometidos – éste es el sentido normalmente asumible en teoría penal al poder coercitivo de la sanción, a la responsabilidad legal y, en el plano de la axiología, a la responsabilidad moral; (ii) sea responsabilidad *de lo debido* orientada hacia el futuro – este segundo sentido es el normalmente asociado a la ética; (iii) sea responsabilidad por los actos cometidos – en el sentido de lo que se hace (en tiempo presente) y de lo que se ha hecho (en tiempo pasado); (iv) sea responsabilidad de padre-hijo – en el sentido de obligación por la responsabilidad que se atribuye a la potestad. Todos estos sentidos tradicionales de "responsabilidad" se encuentran asociados a éticas deontológicas, es decir, *éticas del deber*.

Por su parte, la constatación creciente en la postmodernidad de la vulnerabilidad de la naturaleza ante el yugo de la intervención tecnológica del hombre revela uno de los más importantes desafíos a los que se enfrenta el pensamiento ético con respecto a la condición humana: el ser humano pasa del

[7] Aa.Vv., *Informe 2013: El estado de los derechos humanos en el mundo*, Editorial Amnistía, Madrid, 2013, p. 11.
[8] H. Jonas, *El principio de responsabilidad: Ensayo de una Ética para la civilización tecnológica*, Herder Editorial, Barcelona, 1995.

plano de la "ineludible reciprocidad" entre prójimos – *ama a tu prójimo como a ti mismo* – a una *responsabilidad para con la naturaleza* de marcado carácter teleológico-aristotélico y opuesta, por tanto, a las éticas fundadas en la convicción.

El imperativo ejemplar kantiano – *Actúa de tal modo que el principio de tu acción se transforme en una ley universal* – será adaptado por Jonas a través de nuevas formulaciones de un imperativo colectivo, intergeneracional y de acción en la esfera pública: "Actúa de tal modo que los efectos de tu acción sean compatibles con la permanencia de una vida humana auténtica"; o en su formulación negativa, "No pongas en peligro la continuidad indefinida de la humanidad en la Tierra"; o, lo que es su formulación más programática, "Incluye en tu elección presente, como objeto también de tu querer, la futura integridad del hombre"[9].

Por todo ello, ante el poder extraordinario de transformación de los paradigmas antropológicos que en las últimas décadas nos ha aportado la biotecnología, la neurociencia y las tecnologías disruptivas como la IA, es imperativo formular los presupuestos de una *responsabilidad por lo venidero* que repose – como preconizara el filósofo español Fernando Savater en su ensayo *Ética para Amador*[10] – no sólo en los imperativos categóricos de la biología humana aislada, sino en atributos diferentes a la inteligencia, tales como la voluntad, la solidaridad o capacidad de cooperar, la capacidad de sacrificio y, lo que es más importante, las emociones.

Si emulamos un recorrido por la historia de la ética como el que magistralmente elabora Jonas advertiremos, junto al filósofo alemán, que son muchos los testimonios en filosofía moral que evidencian la necesidad de coaligar la razón con la emoción para que el bien objetivo adquiriera poder sobre la voluntad. En este sentido, existe una consolidada tradición de investigación en antropología moderna y filosofía moral que apuesta por la determinación del elemento emocional de la ética aplicada a las tecnociencias; destacan en este ámbito las aportaciones de Evandro Agazzi, Jürgen Habermas, Paul Ricoeur, Gilbert Hottois o Carl Mitcham, entre otros. En este sentido, la mayoría de las corrientes filosóficas surgidas desde mediados del siglo XIX favorables a la emoción, cuestionan el binomio positivista *razón vs. emoción*, según el cual las emociones deben estar fuera del campo de la razón. Nietzsche es seguramente el exponente más claro del abandono de este modelo racionalista y de la disolución de la dicotomía *razón vs. emoción* al considerar que la emoción anticipa a la razón y que las personas deciden primero por emociones. En este mismo sentido se pronuncia Antonio Damasio, cuando en su obra *El error de Descartes*[11] considera que "la cognición y las emociones no solo están estrechamente entrelazadas, sino que además, la emoción es el primer mecanismo para la racionalidad". A juicio de

[9] *Ivi*, p. 188.
[10] F. Savater, *Ética a Amador: Una invitación a vivir sin odio ni miedo*, Ariel, Barcelona, 2011.
[11] A. Damasio, *El error de Descartes*, Booket, Barcelona, 2001, p. 57.

este autor, si se integran adecuadamente cognición – en el sentido de racionalidad – y emoción, los seres humanos estarían adecuadamente dirigidos a la toma de decisiones racionales.

Asimismo, cada día son más numerosas las voces que muestran la necesidad de abordar la ética aplicada a las tecnociencias y la IA desde la hermenéutica – crítica que promociona la *interdependencia* a través la transversalidad del enfoque emocional en el conocimiento humano. En este sentido, compartimos la propuesta del filósofo valenciano Jesús Conill Sancho de abandonar la reducción cientificista del positivismo naturalista y el factualismo fenomenológico heideggeriano, para abrazar la vía de la hermenéutica-critica – bien conocida en autores como Habermas, Apel o la propia filósofa española Adela Cortina – en un intento por *re-interpetar la realidad de la persona humana desde su fragilidad y vulnerabilidad con relación a la tecnología*[12].

3. Emocionalidad, Vulnerabilidad y Virtud frente a la Tecnología

La llamada *emocionalidad*, mundo emocional o dimensión emocional de la persona es aquella dimensión en la que tiene cabida nuestro mundo interno de sentimientos, lazos afectivos y otras pulsiones emotivas que facilitan nuestra interacción social en los colectivos humanos. A través de las emociones nos vinculamos con el mundo y por tanto, las emociones resultan indispensables no sólo para nuestro bienestar como individuos, sino para poder garantizar una interactuación social adecuada en los colectivos sociales.

En este sentido, la rama conocida como *Sociología de la emoción* distingue entre dos tipos posibles de emociones. (i) En primer lugar, las llamadas *emociones primarias*, que se consideran respuestas fisiológicas, biológicas y neurológicas innatas a los seres humanos e inmediatas ante determinados estímulos. Suelen tratarse de impulsos o reacciones de corta duración a nuestro entorno más inmediato – tales como la ira, el miedo o la alegría. Por consiguiente, el miedo ante un peligro a nuestra integridad física o la ira inmediata ante la visión de un acto de injusticia que nos impulsa a gritar e insultar a alguien, serían ejemplos de esta primer tipología de emociones primarias. (ii) En segundo lugar, las llamadas *emociones secundarias* adquieren una naturaleza siempre *relacional.* Se trataría pues de aquellas emociones construidas desde el colectivo social y que, por tanto, son siempre portadoras de significados dependientes de consideraciones sociales y culturales que definen los momentos y situaciones sociales que los seres humanos vivimos. Por ejemplo, las situaciones sociales en las que el sujeto se siente con adecuado nivel de poder y/o estatus dan lugar siempre a emociones positivas sobre el propio sujeto, emociones que le proporcionan seguridad, satisfacción y respeto. Por el contrario, las situaciones sociales en las que el sujeto se siente con

[12] J. Conill, *Intimidad personal y persona humana. De Nietzsche a Ortega y Zubiri*, Taurus, Madrid, p. 216.

escaso o nulo poder y/o estatus dan lugar siempre a emociones negativas sobre el propio sujeto, emociones que le proporcionarán inseguridad, frustración, miedo e ira. La emoción, por lo tanto, puede llegar a convertirse en un dispositivo político de poder y en un efectivo mecanismo de control social. Quien mayor poder tiene y quien menos poder goza, viven en mundos diferentes y antagónicos, no solo físicos sino sociales y emocionales.

En su libro *Emociones Políticas ¿Por qué el amor es importante para la justicia?*[13], la filósofa norteamericana Martha Nussbaum despliega una detallada argumentación sobre el rol y la repercusión de este segundo tipo de emociones para la vida pública de las democracias actuales. En primer lugar, Nussbaum trata de identificar tanto aquellas *emociones secundarias positivas* que ayudan a consolidar la estabilidad de las democracias contemporáneas (por ejemplo, la compasión) para, a continuación, contrastarlas con aquellas *emociones secundarias negativas* que conspiran contra la estabilidad de la vida pública (por ejemplo, el asco, el miedo y la envidia). A juicio de esta autora, las democracias actuales deben favorecer y promover el cultivo de aquellas emociones secundarias que conducen a la benevolencia y el florecimiento virtuoso – *eudaimonía* – de todos los seres humanos en la esfera pública. En segundo lugar, Nussbaum aspira a proponer estrategias de participación ciudadana en el cultivo de la virtud que permitan fomentar las emociones secundarias positivas en detrimento de las segundas (las llamadas negativas).

Para Martha Nussbaum el *cultivo de las emociones* es condición necesaria para alcanzar la calidad democrática que se le predispone a las sociedades contemporáneas occidentales, cuya estabilidad no sólo se hace depender de su adecuación a unos principios de justicia, sino del desarrollo de "virtudes cívicas" que reposan en emociones y sentimientos. Tanto es así que la estabilidad de las democracias actuales dependería, a juicio de Nussbaum, no sólo mecanismos procedimentales como los que aportan los sistemas normativos – equilibrando los intereses de cada uno y reaccionando en caso de conflicto entre los mismos –, sino a través también de la promoción de lo que Will Kymlicka[14] define como un cierto grado de "emocionalidad" en la esfera pública correspondiente a sentimientos ciudadanos y al espíritu público.

Por consiguiente, la llamada *ética de la virtud* conforma un contexto válido respecto a otras tradiciones filosóficas – centradas en el consecuencialismo de las acciones – para reflexionar sobre el significado ético y universal de aquellas virtudes cívicas – tales como la bondad, la empatía y la compasión – que contribuyen significativamente al bienestar de las comunidades. La *ética de la virtud* trasciende pues las muestras esporádicas de benevolencia en la comunidad, supera las afiliaciones personales y los gestos momentáneos para construir así un

[13] M.C. Nussbaum, *Emociones políticas. ¿Por qué el amor es importante para la justicia?*, Paidós, Barcelona, 2014.
[14] W. Kymlicka, *Ciudadanía multicultural. Una teoría liberal de los derechos de las minorías*, Paidós, Barcelona, 1996.

compromiso cívico más profundo con la empatía, la comprensión y la buena voluntad activa hacia los demás. Al adoptar esta perspectiva, las personas activan la *interdependencia humana* y fomentan la empatía, la compasión y la comprensión hacia "el otro".

A este respecto, Martha Nussbaum nos ofrece un estudio transversal de lo que califica como "filosofía de la bondad" en el que cobra un papel esencial la *compasión como virtud en la promoción del florecimiento humano y la vida ética*. La exploración de la bondad por parte de Nussbaum subraya su valor inherente en la toma de decisiones éticas y en el florecimiento moral, ya que contribuye al desarrollo de un *yo virtuoso y compasivo*.

En su obra monumental *Paisajes del pensamiento: La inteligencia de las emociones*[15], Nussbaum analiza la compasión como virtud como primera etapa fundacional hacia la bondad. La compasión representa, a su juicio, una virtud esencial que contribuye al bienestar general de las personas y las comunidades, que trasciende los meros actos de benevolencia y que abarca una preocupación genuina por el bienestar y la dignidad de los demás. Es más, la compasión desempeña para esta autora un rol esencial en el fomento de la cohesión social y en el cultivo de relaciones interpersonales. Nussbaum se adhiere a la tradición aristotélica de la compasión como aquella "pasión penosa suscitada por el dolor o el sufrimiento de otro"[16] o por "el pesar o dolor físico y los grandes males que provoca la fortuna"[17]. Asimismo, en la formulación de su teoría cognitiva de las emociones, Martha Nussbaum formula un concepto crucial, el concepto de la llamada *compasión política universal*, para referirse a aquellos *impulsos individuales de compasión*, los cuales nos llevan a atender las necesidades materiales de los demás y que sirven de base para la generación de acuerdos institucionales, objetivos políticos y políticas públicas. De este modo, argumenta Nussbaum que la compasión, convenientemente guiada por la racionalidad, puede contribuir a la realización del bienestar colectivo.

Por su parte, Alasdair McIntyre, en su reconocida obra *Tras la virtud*[18], coincidiría con la perspectiva de Nussbaum al entender que la bondad es una disposición de carácter que emana de una preocupación genuina por el bienestar de "los otros", reconociéndoles su dignidad y que, en consecuencia, moldea nuestra perspectiva ética y nuestros compromisos morales. Para MacIntyre, como también lo es para Nussbaum, la bondad trasciende las acciones aisladas, los intereses personales y los gestos benévolos. Sin embargo, McIntyre discrepa con Nussbaum cuando afirma que la filosofía de la bondad no debe separarse del discernimiento moral y la razón, ya que la bondad requiere una reflexión meditada

[15] M.C. Nussbaum, *Paisajes del pensamiento: La inteligencia de las emociones*, Paidós, Barcelona, 2008.
[16] Aristóteles, *Retórica*, 1385b13 y ss.
[17] M.C. Nussbaum, *La fragilidad del bien: Fortuna y ética en la tragedia y la filosofía griega*, Visor, Madrid, 1995, p. 476.
[18] A. MacIntyre, *Tras la virtud*, Ed. Crítica Ediciones Grijalbo, Barcelona, 1984.

y una acción deliberada cuando nos esforzamos por adoptar comportamientos que promuevan el florecimiento virtuoso de los demás.

4. Hacia una IA Compasiva

La Inteligencia artificial puede cambiar radicalmente el mundo en el que vivimos, pero es la ética detrás de ella la que determinará cómo será ese cambio. En este sentido, la transversalidad del enfoque de la *ética de la virtud* nos obliga a entablar un debate en torno a cuáles son las virtudes éticas fundamentales en las sociedades digitales y si la *ética de la virtud* debería vincularse a la *responsabilidad de lo venidero frente al desarrollo tecnológico* que apuntara el filósofo alemán Hans Jonas.

Al embarcarnos en el viaje trascendente que nos aporta el enfoque de la *ética de la virtud*, comprobamos que no se trata simplemente de crear máquinas inteligentes, capaces de sustituir e incluso superar al ser humano en su raciocinio y capacidades cognoscitivas, sino de fomentar una *IA virtuosa* que refleje lo mejor de las aspiraciones morales del ser humano.

Debemos empezar pues reconociendo la responsabilidad moral que tenemos como arquitectos de un marco ético de la IA. Integrando bondad, empatía y compasión en el diseño de los sistemas de IA ayudaremos decisivamente a dar prioridad al bienestar de los usuarios, promover la equidad, la transparencia y la responsabilidad, y garantizar que las tecnologías de IA sirvan a los intereses de los ciudadanos y no así de corporaciones tecnológicas o poderes fácticos. Esto último se debe traducir en diseños de algoritmos de IA que den prioridad a la empatía, al respeto y a la preservación de la dignidad humana frente a la construcción de sesgos discriminatorios.

En efecto, cuando imbuimos a las máquinas de inteligencia y capacidad de decisión, las virtudes que podemos inculcar en ellas se convierten en la verdadera piedra angular del desarrollo ético de la tecnología, sobre todo a la hora de hacer frente a las posibles injusticias sistemáticas que se pudieran derivar del uso malicioso de sistemas de IA y de las injusticias sistémicas que implican los sesgos en los datos y en los algoritmos discriminatorios.

Los seres humanos somos particularmente vulnerables ante la IA por el potencial peligro que entraña para las personas el empleo de dispositivos técnicos que inducen a una violencia psicológica encaminada al control de la voluntad humana – la cual ha sido calificada por los especialistas en neurociencia como "violencia neuronal".

La Propuesta de *Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de Inteligencia artificial* es consciente de que la vulnerabilidad humana es central en cualquier contexto de disrupción como el que aporta la IA. Por este motivo, prohíbe (i) en primer lugar, aquellos sistema que desplieguen técnicas subliminales con el objetivo de distorsionar el comportamiento de una persona de manera que pueda causarle

daños físicos o psicológicos a ella o a otros y (ii) en segundo lugar, aquellos sistemas de IA que exploten vulnerabilidades de un grupo específico de personas por su edad, discapacidad o situación social o económica, de tal forma que se distorsionen el comportamiento de estas personas y probablemente les causen daños a ellas o a terceros.

Por tanto, para hacer realidad una *IA confiable* no solo es suficiente con cumplir con la ley, si no que es necesario adoptar una serie de estándares éticos en su uso que generen en la sociedad civil la confianza debida de que los sistemas de IA no provocarán un daño, no solo doloso, si no involuntario a sus usuarios.

Sin embargo, construir un marco ético universal para los sistemas de IA no es tarea fácil y nos enfrenta al manido embate postmoderno del relativismo ético de que cada comunidad y cada persona pueden tener concepciones distintas de lo que constituye "lo moralmente aceptable". El enfoque colectivo que hemos explorado antes a través de la *responsabilidad por lo venidero* de Jason, nos proporciona una justificación más explícita de la asunción de un paradigma moral universal, al basarlo en experiencias humanas compartidas y en los principios duraderos de las tradiciones filosóficas que han guiado el fundamento ético a lo largo de la historia. Por este motivo, la construcción de un marco ético compartido para la IA, lejos de parecer una utopía, se vislumbra como una necesidad cada vez más acuciante.

La adopción de un paradigma moral y de un marco ético universal para el desarrollo de la IA obedece a dos presupuestos. (i) En primer lugar, un marco ético universal nos sirve de parapeto frente los inconvenientes de la reducción cientificista del positivismo naturalista y el factualismo fenomenológico heideggeriano, y nos impulsa en la dirección de una hermenéutica-crítica que promociona la *interdependencia* a través la transversalidad del enfoque emocional en el conocimiento humano. La construcción de un marco ético compartido fomenta pues una comprensión colectiva de *aquellos límites que la IA no debe sobrepasar*. (ii) En segundo lugar, la búsqueda de un marco de referencia ético universal en la IA se alinea con el propio *telos* de la tecnología. La palabra "tecnología" procede etimológicamente del griego τέχνη ("téchnē") en el sentido de *arte*, *oficio* o *destreza.* Desde una perspectiva más amplia, la tecnología se identifica con *un proceso o una capacidad de transformar o combinar algo ya existente para construir algo nuevo y permitir, de este modo, mejorar y profundizar la existencia humana*. Esta segunda manera de abordar el sentido de la tecnología tiene su punto de partida en el presupuesto aristotélico de que la manera debida de valorar los bienes, instituciones y prácticas sociales depende de los propósitos o fines de dichas practicas o instituciones. Y, precisamente, la idea de que podemos descubrir las virtudes – tanto éticas como dianoéticas – apropiadas a un proceso, oficio o destreza como la tecnología, cuando tratamos de comprender el fin, propósito o *telos* de dicho proceso, constituye el núcleo de la teoría aristotélica de la Justicia y el fundamento de la *ética de la virtud*.

Si hacemos depender el desarrollo de una IA *con factor humano* – respetuosa pues con la dignidad humana – al reconocimiento de marcos éticos

virtuosos, no solo mitigamos los riesgos de la divergencia anti-cognitivista, sino que garantizamos que las creaciones tecnológicas contribuyan positivamente al bienestar de las personas y de la sociedad en general. Asimismo, para poder encontrar un equilibrio adecuado entre los intereses individuales y el bienestar colectivo, es necesario explorar a fondo las consideraciones que en Bioética se han llevado a cabo entorno a la naturaleza de la virtud y los límites éticos[19].

5. Principio de Beneficencia ante el riesgo de discriminación por sesgos algorítmicos

El llamado *principio de beneficencia* se refiere a la obligación moral de prevenir o aliviar el daño – por tanto, *de hacer el bien* – y el deber de ayudar al prójimo por encima de los intereses particulares; en otras palabras, *obrar en función del mayor beneficio posible, procurando el mayor bienestar general*.

Como enuncia Kottow en su *Introducción a la bioética*[20], los elementos que se incluyen en este principio son todos los que implican *una acción de beneficio que haga o fomente el bien, prevenga o contrarreste el mal o daño y, adicionalmente, todos los que implican la omisión o la ausencia de actos que pudiesen ocasionar un daño o perjuicio.*

El llamado "principio de beneficencia", a juicio de Beauchamp y Childress, favorece que los individuos y las instituciones sientan la obligación ética de contribuir activamente al bienestar de la comunidad[21]. Destacan en este contexto, en particular, la promoción de virtudes cívicas como el altruismo, la solidaridad, la compasión y la responsabilidad social en las acciones humanas.

Por su parte, el filósofo Peter Singer, en su obra *Vivir éticamente: Cómo el altruismo eficaz nos hace mejores personas*, revitaliza el razonamiento moral utilitario, según el cual los individuos tienen la obligación moral de evitar el sufrimiento y promover el bienestar en la medida de sus posibilidades. Según la propuesta de Singer, disponemos de *libre albedrío* para elegir el curso de nuestras

[19] H. Engelhardt, *Fundamentos de la bioétic*a, ed. 2, Paidós, Barcelona, 1995. Engelhardt formula en su tratado introductorio cuáles son los cuatro principios básicos de la Bioética: *autonomía, beneficiencia, no-maleficiencia y justicia*. (i) La autonomía es entendida como la capacidad de las personas de deliberar sobre sus finalidades personales y de actuar bajo la dirección de las decisiones que pueda tomar. (ii) La beneficencia significa "hacer el bien", entendida como la obligación moral de actuar en beneficio de los demás. Curar el daño y promover el bien o el bienestar. (iii) La no-maleficencia significa no producir daño y prevenirlo. Incluye no matar, no provocar dolor ni sufrimiento, no producir incapacidades. Es un principio de ámbito público y su incumplimiento está penado por la ley. (iv) Por último la justicia en su dimensión aristotélica, entendida como la equidad en la distribución de cargas y beneficios. El criterio para saber si una actuación es o no ética, desde el punto de vista de la justicia, es valorar si la actuación es equitativa, con un rechazo a cualquier forma de discriminación por cualquier motivo.

[20] M. Kottow, *Introducción a la bioética*, Editorial Universitaria, Santiago de Chile, 1995, p. 72.

[21] T. Beauchamp, J. Childress, *Principles of Bioethical Ethics*, ed. 2, Oxford University Press, Oxford, 1994, pp 148-149.

acciones y en esta autonomía de la voluntad debemos sobrepujar aquellas decisiones que maximizan el bienestar general y que, en contrapartida, reducen el sufrimiento innecesario[22].

No obstante, el *principio de beneficencia* nos enfrenta ante el principal escollo de doctrina moral del Utilitarismo bethamiano cuando formula "promover el bienestar y actuar de forma que maximice la felicidad del mayor número posible de personas". En este sentido se pronuncia el filósofo norteamericano Michael J. Sandel, quien considera que el punto débil más clamoroso del razonamiento moral utilitarista es precisamente su falta de respeto a los derechos individuales. "La lógica utilitaria, si se aplica coherentemente, refrenda maneras de tratar a las personas que violan normas de decencia y respeto que creemos fundamentales"[23].

Esta última aproximación sobre las deficiencias de la lógica utilitaria es todavía más evidente en el contexto de la IA, y nos obliga a centrar nuestros esfuerzos en mecanismos de beneficiencia que permitan mitigar daños innecesarios asociados a los sistemas de IA, sobre todo aquellos que pudieran comprometer seriamente el bienestar colectivo. A este respecto, un aspecto crucial con el que deben lidiar los desarrolladores de sistemas de IA para dar cumplimiento a este *principio de beneficiencia* es la cuestión de mitigar el impacto de los sesgos en los algoritmos de IA.

Lamentablemente, los sistemas de IA tampoco son inmunes a los sesgos y prejuicios que existen en la sociedad. Estos prejuicios tienden a impregnar los algoritmos y a propagar la discriminación. En consecuencia, si efectuamos una mala selección de los datos que se emplean para entrenar a un sistema de IA, el resultado puede suponer que la predicción que nos facilita el algoritmo llega a decisiones injustas que inducen a la discriminación o estigmatización de determinados colectivos y/o personas a través de la generación de creencias preconcebidas, predilecciones o prejuicios inconscientes – los llamados "sesgos" – que podemos haber adquirido a lo largo de nuestras vidas fundadas en los estereotipos socioculturales con los que se nos ha educado.

Ejemplos de esto último los encontramos en el recurso a una modalidad de algoritmos que se emplean en aplicaciones de IA para el reconocimiento facial, los cuales muestran un sesgo discriminatorio significativo a la hora de identificar a personas de diferentes orígenes raciales y étnicos. Asimismo, una segunda muestra la encontramos en los llamados algoritmos predictivos de control *ante facto*. (i) En el caso de los primeros, algoritmos para el reconocimiento facial, se ha demostrado que este sesgo obedece a la representación desequilibrada que los desarrolladores de sistemas de IA para el reconocimiento facial emplean en los datos de entrenamiento, en los que predominan los rostros masculinos y de piel clara. En consecuencia, estos sistemas presentan elevadas tasas de error en la

[22] P. Singer, *Vivir éticamente: Cómo el altruismo eficaz nos hace mejores personas*, Paidós, Barcelona, 2017.
[23] M. Sandel, *Justicia ¿Hacemos lo que debemos?*, Ed. Debate, Barcelona, 2011, p. 48.

identificación, al no haber mitigado los sesgos discriminatorios por razones étnicas o raciales. (ii) En el caso de los segundos, la función de los algoritmos predictivos *ante facto* es la de determinar el grado de riesgo, por posible reincidencia, que puede tener una persona a lo largo de las diferentes etapas en que se desarrolla un procedimiento judicial o proceso administrativo. Así pues, desde un punto de vista estrictamente normativo, el empleo de algoritmos predictivos de riesgo implica graves impedimentos para que el legislador, en su tarea de regulación jurídica con ánimo de permanencia futura, pueda reprimir y combatir eficazmente las situaciones vulneradoras de los derechos fundamentales. Asimismo, supone también un problema para los operadores jurídicos, ya que los sistemas de IA predictivos buscan siempre establecer virtualidades de comportamiento humano y eso puede derivar a la postre en una patronización de modelos erróneos que acabe desembocando en sesgos cognitivos y discriminatorios en base al género, la nacionalidad, los orígenes étnicos, la raza o la religión.

Tanto en la administración de justicia como en la actividad policial se han implementado en estos últimos años una serie de *algoritmos predictivos de riesgo* aplicables a aquellas personas que responden a estereotipos delictivos asociados a diferentes grupos raciales u orígenes étnicos que pueden hacer aumentar la idea de culpabilidad o que están vinculados a un mayor nivel de delincuencia, lo que daría lugar a diferencias de trato en investigaciones policiales y sentencias en función del llamado *prejuicio judicial*. Este fue el caso del empleo durante años del llamado modelo COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) que se utilizó en Estados Unidos para evaluar la probabilidad de reincidencia de las personas implicadas en el sistema de justicia penal. El sistema COMPAS reveló una discriminación algorítmica manifiesta hacia varones de origen afroamericano, los cuales representan a la población penitenciaria en Estados Unidos que aglutina penas de reclusión más largas – incluyendo cadena perpetua – respecto a cualquier otra combinación de raza y género. Tras diversas investigaciones se pudo demostrar que el sistema COMPAS se sustanciaba en un marcado sesgo racial, ya que era más probable que se etiquetara erróneamente a varones afroamericanos – dado su perfil en el sistema de un mayor riesgo de reincidencia – en comparación con los varones blancos. El caso COMPAS fue objeto de gran atención por parte de la opinión pública y suscitó preocupación por el posible impacto discriminatorio de los algoritmos de IA cuando se emplean en las decisiones jurisdiccionales en el orden penal.

Un caso que nos es más próximo que el anterior fue el del denominado *Sistema de Indicación de Riesgos* – conocido bajo las siglas SyRI – que el gobierno neerlandés utilizó para prevenir y combatir el fraude de prestaciones de la Seguridad Social. Dicho sistema permitía que las administraciones públicas neerlandesas recurrieran a la elaboración de informes de riesgo de aquellos demandantes de subsidios por hijos menores de edad a su cargo para evitar, de este modo, que se hiciera un uso ilegal de fondos gubernamentales en el campo de la Seguridad Social.

El Sistema SyRi se había configurado a partir del marco normativo que aportaba una ley nacional – la llamada *Ley sobre la estructura de organización de ejecución del trabajo y los ingresos* –, cuyo artículo 65.2 contenía una extensa lista de categorías de información que podían procesarse en el sistema SyRI; en concreto: género, historial de empleo, impuestos, propiedad de bienes, educación, seguro médico, permisos del gobierno, el nivel de endeudamiento, el historial de prestaciones públicas percibidas, las sanciones administrativas, tales como multas de tráfico, entre otros.

Para calcular posibles irregularidades de evasión y fraude, los algoritmos del sistema SyRI enlazaban todos los datos personales de los solicitantes almacenados por instancias gubernamentales para, a continuación, cotejarlos con un "perfil de riesgo" generado a partir de la información de otros ciudadanos con antecedentes penales. Una vez establecidas las posibles similitudes y/o divergencias, el sistema confeccionaba informes de riesgo sobre una lista de nombres como "potenciales defraudadores" que las autoridades podían conservar hasta dos años.

El sistema SyRi se sustanciaba en *proyectos de vecindad* en los que las agencias gubernamentales identificaban aquellos distritos municipales en los que se quería implementar este sistema de evaluación de riesgos. Esto en la práctica se tradujo la implantación del SyRI exclusivamente en aquellos barrios y distritos municipales económicamente más desfavorecidos, caracterizados por sus altas tasas de población inmigrante. Como consecuencia de ello, las autoridades administrativas holandesas acusaron erróneamente de fraude a centenares de familias perceptoras de la prestación por el mero hecho de tener ascendencia marroquí o turca. Resulta pues evidente a través de este ejemplo que el modelo de riesgo "ante facto" elaborado por SyRI alicienta una evidente discriminación de la ciudadanía con ascendencia árabe.

Este caso dio lugar a la primera decisión judicial en Europa en la que se puso a examen un sistema algorítmico de evaluación del riesgo, la conocida como *Netherlands Committee of Jurists for Human Rights vs. State of the Netherlands*. En dicha sentencia de 6 de marzo de 2020, el tribunal concluyó que el sistema SyRI no sólo había producido la afectación del derecho humano a la vida privada, sino que vulneraba adicionalmente la exigencia de transparencia del artículo 8 del Convenio Europeo de Derecho Humanos. Asimismo, el alto tribunal también analiza la legitimidad del recurso gubernamental a configurar informes de riesgo de la ciudadanía para calcular la atribución de prestaciones sociales, concluyendo que el sistema SyRi no era "ni transparente ni verificable"; no sólo porque un sistema así podía ser empleado en la elaboración de perfiles de datos de las personas para otros fines – lo que está prohibido por ley –, si no también porque los modelos de riesgo empleados por el Gobierno neerlandés nunca fueron publicados; asimismo, los interesados no fueron notificados con carácter previo sobre dicho particular cuando sus datos fueron introducidos en el sistema SyRi para la confección de su *perfil de riesgo* ante la Administración pública. Por último, por lo que respecta a la prueba de equilibrio, el tribunal llegó a la conclusión de que un informe de riesgo tiene un efecto jurídico no desdeñable

sobre el derecho a la privacidad de la persona objeto de escrutinio algorítmico y si bien un informe así no puede utilizarse a efectos sancionatorios, esto no excluye para que una información tan sensible pueda ser empleada en ulteriores procedimientos y comunicaciones de la ciudadanía con la Administración pública. En base a este razonamiento, el tribunal desestimó el "interés declarado del Gobierno neerlandés".

Los dos ejemplos que hemos seleccionado nos sitúan ante la principal amenaza que acompaña a la automatización de los algoritmos en los sistemas predictivos de riesgo que emplea la IA. Así pues, el recurso de los Gobiernos a configurar "informes de riesgo" de sus ciudadanos mediante sistemas de IA predictiva lamina uno de los pilares epistemológicos sobre los que se fundamenta la propia definición legal de Estado de Derecho: la noción de autonomía y autodeterminación de las personas.

Los defensores de las concepciones contemporáneas de autodeterminación se sirven del ideal kantiano de autonomía moral para luchar contra la visión estereotipada de aquellos que critican fuertemente la concepción libertaria de autonomía personal, tachándola de individualista o incluso egoísta. Los libertarios, en efecto, conciben la autonomía personal a expensas de deseos o de preferencias subjetivas y en detrimento del bien común. Por este motivo, liberales postmodernos como Roberts o Joseph Raz han apostado por una noción de *autonomía socializada* que permite coaligar de forma efectiva el ideal clásico kantiano de autonomía de la voluntad con los retos que nos ofrecen las nuevas aplicaciones de los algoritmos en los sistemas de IA. Si nuestra autonomía y capacidad de actuar y elegir libremente está comprometida mediante el empleo de algoritmos, como el sistema SyRI, ya no actuamos conforme a una máxima que nos damos a nosotros mismos, si no que lo hacemos en cumplimiento de una máxima que la comunidad debe darse a sí misma, por el bien común.

A pesar de las deficiencias apuntadas, hemos de tener en cuenta que parte del poderoso atractivo que las aplicaciones la IA tienen hoy en día es el de venderse como un mecanismo de superación de la subjetividad humana, o incluso de erradicación de estereotipos y prejuicios sociales, a través del recurso a la certeza y neutralidad que nos ofrecen los algoritmos.

Sin embargo, ejemplos como el llamado sistema SyRI nos hacen considerar que existe la amenaza potencial de que los sistemas algorítmicos de evaluación de riesgos acaben siendo explotados por instancias gubernamentales con vistas a instaurar un sistema represivo que fuera en detrimento de las libertades públicas y garantías constitucionales de los ciudadanos; o incluso, como nos advierten los profesores Lorenzo Cotino Hueso y Jorge Castellanos Claramunt, que se tratase de un sistema vulnerador de la dignidad y los derechos humanos, sobre todo tratándose de minorías étnicas y población inmigrante[24].

[24] L. Cotino Hueso, J. Castellanos Claramunt (coords.), *Algoritmos abiertos y que no discriminen en el sector público*, Tirant lo Blanch, Valencia, 2023.

Cualquier persona susceptible del "escrutinio algorítmico" podría ser potencialmente perseguida en base a eventuales virtualidades de comportamiento susceptibles de riesgo. No seríamos, así pues, perseguidos por los actos que hayamos realizado, sino por determinadas predicciones de comportamientos *pro-futuro* o *ante-facto* en razón de sesgos identitarios asociados a algoritmos que tiene en cuenta desde nuestro nivel de ingresos y de endeudamiento, pasando por nuestra interacción en redes sociales, hasta nuestro lugar de residencia, nuestra religión o nuestro origen étnico.

La problemática que subyace en el "escrutinio algorítmico" que aplican los sistemas de IA de control de riesgo pone de relieve la necesidad del abordaje ético de los sistemas de IA. También nos cuestionamos sobre si la promoción de virtudes como la bondad, la empatía y la compasión en el diseño de los sistemas de IA nos ayuda verdaderamente a generar algoritmos capaces de mitigar los sesgos históricamente asociados a determinadas minorias por razones étnicas, religiosas y raciales. Esta segunda apreciación nos lleva a considerar el segundo de los principios bioéticos anteriormente enunciado: el principio de *no-maledicencia*.

6. Principio de No-maledicencia: La prevención del daño y la preservación de la dignidad humana ante el riesgo de alexitimia virtual

El segundo principio que debemos considerar en nuestra apuesta por una ética de la virtud aplicada a la IA es el *principio de no-maleficencia*. Este principio se inscribe en la tradición de la máxima clásica *primum non nocere – lo primero es no dañar* – y hace referencia a la obligación de no infringir daño intencionadamente. El *telos* de este segundo principio bioético de *no dañar a otros* – por ejemplo, no robar, no lastimar o no matar –, es claramente diferente de la obligación de beneficencia, la cual busca proteger intereses personales o promover el bienestar colectivo.

Aplicado a los sistemas de IA, el *principio de no-maledicencia* garantizaría que dichos sistemas diesen prioridad, en primer lugar, a la la seguridad de las personas o de los potenciales usuarios y, en segundo lugar, a promocionar la dignidad humana, minimizando así el riesgo y maximizando la transparencia y la explicabilidad.

Son numerosos los casos en los que los dispositivos de IA han incorporado el *principio de no-maledicencia* para promover la seguridad de los usuarios. Un ejemplo recurrente lo encontramos en los medios de automoción, en particular en la implementación de sistemas de IA en vehículos autónomos con el objetivo de reducir los accidentes de tráfico y mejorar la seguridad vial. Multinacionales como Tesla o Uber han desarrollado en estos últimos años prototipos de coches automatizados que emplean algoritmos de IA para percibir el entorno y anticipar a continuación decisiones de conducción automatizadas, o incluso controlar el propio vehículo durante desplazamientos por carretera. Lo que se persigue es

precisamente la superación del error humano para mejorar la seguridad vial ya que, de esta manera, los vehículos autónomos tienen el potencial de reducir significativamente los accidentes causados por errores humanos; ya sea intencionados como no intencionados como el déficit de atención, la fatiga visual o la falta de reflejos de los conductores. Se trataría aquí, por tanto, de un ejemplo de implementación del *principio de no-maledicencia*, ya que el objetivo de los dispositivos de IA es el de minimizar los daños potenciales, maximizando la seguridad de las personas en la carretera.

Al contemplar las posibles aplicaciones del *principio de no-maledicencia* a los sistemas de IA debemos cuestionarnos acerca del papel que juegan virtudes como la bondad, la empatía y la compasión en el aprendizaje automátizado de los sistemas de IA, en el procesamiento del lenguaje natural y en la comunicación afectiva para el diseño de las aplicaciones de IA que sean capaces de percibir, comprender y responder a las emociones humanas. En efecto, en el proceso de *Machine Learning*, los sistemas de IA adquieren la capacidad de aprender de ingentes recopilaciones de datos para hacer predicciones de comportamientos humanos *ante facto*. Por tanto, si entrenamos a los sistemas de IA a recopilar con carácter prioritario datos que ejemplifiquen las virtudes esenciales a la condición humana – tales como la bondad, la empatía, la solidaridad, el coraje, la prudencia y la compasión – podemos así imbuirles la capacidad de reconocer y responder a situaciones de tal forma que se promuevan el bien común en las comunidades humanas.

El aprendizaje automatizado nos ofrece posibilidades infinitas para instruir a la IA en virtudes como la bondad, la empatía o la compasión. Pero, *¿cuáles son las limitaciones que debemos superar en esta tentativa por implantar virtudes éticas en la IA?*

El afán de la ética de la virtud de incorporar en los sistemas de IA valores esenciales a la condición humana se enfrenta al riesgo ineludible de la *alexitimia*. Estimular el aprendizaje afectivo en los procesos automatizados de la IA, o impulsar computación afectiva, evidentemente hace aumentar el potencial de los sistemas de IA para percibir, comprender y responder a las emociones humanas. No obstante, debemos aceptar dichos avances con recelo ya que, si bien los sistemas de IA puedan simular empatía, bondad y compasión, carecen en realidad de la conexión emocional que se deriva de la experiencia humana; o dicho en términos kantianos, de la *condición de humanidad*.

Cualquier tentativa de construcción de un marco ético de la IA encierra una promesa inconmensurable de "humanizar la tecnología" a través de la promoción de virtudes esenciales a la condición humana. En esta ingente apuesta por una *IA con factor humano*, las virtudes éticas se nos revelan como mecanismos válidos a la hora de afrontar algunos de los riesgos más acuciantes de la implementación de los sistemas de IA en nuestro quehacer cotidiano. Así es, entreverar las tres virtudes éticas de la bondad, empatía y compasión en el desarrollo de la IA allana el camino para el desarrollo de una IA con un profundo sentido de humanidad.

Como hemos tratado de justificar a lo largo de nuestra investigación, (i) la bondad aplicada al diseño de los sistemas de IA exige equidad, fomentando así el reconocimiento de un principio de *no-maleficiencia* para evitar los daños potenciales que se pueden derivar del empleo de sesgos algorítmicos. (ii) La promoción de la empatía obliga a los desarrolladores de sistemas de IA a dar buen uso del *principio de beneficencia* para, de este modo, compensar los prejuicios ocultos en los algoritmos de prevención de riesgos *ante facto*, así como los efectos discriminatorios en base a identidades étnicas y raciales que derivan inevitablemente en la segregación de las minorías. Y, por último, (iii) la promoción de la compasión en la IA es el motor que impulsa la cada vez más demandada responsabilidad ética de los programadores y desarrolladores de IA ante las secuencias algorítmicas de sesgo y los efectos maliciosos de las últimas utilidades de IA generativa.