Ed UVa EDICIONES Universidad Valladolid

¿Por qué ética para la Inteligencia Artificial? Lo viejo, lo nuevo y lo espurio *

Why Ethics for Artificial Intelligence? The old, the new and the spurious

TXETXU AUSÍN
Instituto de Filosofía
CCHS-CSIC
Albasanz 26-28, 28037 Madrid, España
txetxu.ausin@cchs.csic.es
ORCID: https://orcid.org/0000-0003-2098-5540

DOI 10.24197/st.Extra_2.2021.1-16

RECIBIDO: 14/06/2021
ACEPTADO: 20/08/2021

Resumen: Nuestra realidad y nuestra vida se configuran ya como un sistema sociotécnico en el que interactuamos personas, dispositivos, datos, algoritmos, robots. La inteligencia artificial, la ciencia de datos y la robótica constituyen tecnologías disruptivas que están reconfigurando la vida. Y ya sabemos que los artefactos tecnológicos incorporan valores y tienen política. Por ello se hace indispensable un análisis ético de dichas tecnologías, identificando los peligros que queremos evitar e introduciendo desde el diseño mismo los valores que queremos promover. No como un fin en sí mismo, sino para generar confianza ciudadana en las tecnologías y en las instituciones que las impulsan, para favorecer su aceptación y apropiación social. La ética constituye una poderosa herramienta para el empoderamiento tecnológico de la ciudadanía, previniendo las injusticias algorítmicas (discriminación, perfilamiento, sesgos) e impulsando un necesario nuevo contrato "tecno-social".

Palabras clave: ética; inteligencia artificial; cuidado; confianza; bien común

Abstract: Our reality, our life, is already configured as a socio-technical system in which we interact with people, devices, data, algorithms, robots. Artificial intelligence, data science and robotics are disruptive technologies that are reconfiguring life. And we already know that technological artifacts incorporate values and have politics. That is why an ethical analysis of these technologies is essential, identifying the hazards we want to avoid and introducing from the very design the values we wish to promote. Not as an end in itself, but to generate citizen trust in the technologies and in the institutions that promote them, to foster their social acceptance and appropriation. Ethics constitutes a powerful tool for the technological empowerment of citizens, preventing algorithmic injustices (discrimination, profiling, biases) and promoting a necessary new "techno-social" contract.

Keywords: ethics; artificial intelligence; care; trust; common good

* Este trabajo se ha realizado en el marco de los proyectos INBOTS CSA network: *Inclusive Robotics for a better Society* (EU H2020 G.A. 780073), EXTEND: Bidirectional Hyper-Connected Neural System (EU H2020 G.A. 779982) y BIODAT: *Datos de salud: Claves ético-jurídicas para la transformación digital en el ámbito sanitario* (Fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia – Ref. 20939/PI/18).

1. CONTEXTO: EL INTERNET DEL TODO

La llamada Inteligencia Artificial (IA) forma parte de ese conjunto de tecnologías "convergentes" que están transformando radicalmente nuestro mundo y nuestra vida en lo que se conoce como Cuarta Revolución industrial o Revolución 4.0. Supone la convergencia de tecnologías digitales, físicas y biológicas que evolucionan a gran velocidad. Son las llamadas NBIC (nano-bio-info-cogno): nanotecnologías, biotecnologías, tecnologías de la información y ciencias cognitivas (IA, ciencia de datos, robótica, interfaces cerebro-máquina, biología sintética, nanotecnología). Se trata de una enorme revolución económica y social que ayuda a una toma de decisiones más atinada e informada y que permite identificar y predecir tendencias y correlaciones. Esto es especialmente relevante, por ejemplo, en ámbitos de interés social como la salud, para predecir la expansión de epidemias, descubrir efectos secundarios en los medicamentos, establecer medidas contra la contaminación ambiental, etc.

La automatización podría liberar al ser humano de tareas peligrosas o alienantes. La economía de la transformación digital puede cambiar nuestros modos de ver y hacer las cosas y generar nuevos modelos de emprendimiento, como la innovación a través de la cooperación (inteligencia colectiva y experimentación abierta) y nuevas oportunidades de activismo social.

Asimismo, las políticas públicas basadas en datos contribuirán a modelos decisionales menos especulativos, a reducir riesgos e incertidumbres. Las políticas de datos abiertos contribuyen a la transparencia y la rendición de cuentas de las administraciones públicas, favoreciendo la participación y el compromiso ciudadano con las políticas públicas, y la IA puede mejorar la calidad de la administración pública (Ramió 2019). La pandemia actual ha acelerado la relevancia de la digitalización y la importancia de la IA para afrontar desafíos y problemáticas de todo tipo, no solo con respecto a la salud pública sino también con relación al trabajo, la gestión administrativa, las políticas sociales o la participación ciudadana.

La IA se basa en una explosión de artefactos e instrumentos, altamente interconectados que recogen enormes cantidades de información, de todos los objetos (Internet de las Cosas) y también de nuestros cuerpos (Internet de los Cuerpos), que incorporan infinidad de sensores que registran *todos* nuestros datos (Internet del Todo): dispositivos en todos los objetos cotidianos (domótica), en los vehículos, en los teléfonos y ordenadores, pero también dispositivos médicos, píldoras y medicinas digitales, una variedad de dispositivos de seguimiento del estilo de vida y la forma física, y una gama cada vez más amplia de dispositivos adheridos o incorporados al cuerpo que se despliegan en escenarios empresariales, educativos y recreativos. La mayor parte de la información es generada por los consumidores a través de la interacción con los servicios basados en Internet y telefonía móvil. 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: Se han desarrollado tecnologías biométricas que se instalan en espacios públicos y privados, como los sistemas de reconocimiento facial, los sensores de huellas dactilares y los escáneres de retina, que se centran en la recogida y el procesamiento de los datos de una gran población o grupo, en lugar de individuos particulares.

La IA procesa, cruza y reutiliza esta ingente cantidad de datos mediante algoritmos. Un algoritmo es una lista más o menos larga de instrucciones, un conjunto ordenado y finito de pasos que puede emplearse para hacer cálculos, resolver problemas y alcanzar decisiones. Dados un estado inicial y una entrada, siguiendo los pasos sucesivos se llega a un estado final y se obtiene una solución.[1]

La interacción de los seres humanos con la IA y, en general, con las tecnologías convergentes, 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 los artefactos en una suerte de simbiosis entre la inteligencia humana y la artificial (*Human-Machine-Interaction*). En este sentido, la IA es inteligencia colectiva y social, constituyendo sistemas inteligentes multi-agentes, por lo que sería más apropiado hablar de Inteligencias Artificiales, en plural.

Este ímpetu tecnológico-industrial de carácter acelerado y rápido[2] ha provocado impactos ecológicos masivos, hasta el punto de que se habla ya de una

[1] Los algoritmos son inherentes a la vida humana porque operan, en sentido amplio, como los procedimientos o pasos que se establecen para conseguir un propósito cualquiera, es decir, un método de resolución de problemas y/o de toma de decisiones. Desde un punto de vista matemático, recordemos que el término deriva del erudito árabe Al-Juarismi (s.VIII-IX), quien compiló los saberes algebraicos previos y dio lugar a esta forma poderosa de cálculo. Pero, en términos generales, donde haya una codificación de medios y fines que permita procesar informaciones con fines pragmáticos (es decir, computar), se trabaja con algoritmos. Son muy útiles, aunque a la vez deben analizarse en sus diversos registros y aplicaciones para evitar que se extralimiten, pues podrían colonizar de hecho y de derecho buena parte de la conciencia y de las actividades humanas. (Espinosa 2020).

[2] "La Gran Aceleración es como se conoce al fenómeno de rápidas transformaciones socioeconómicas y biofísicas que se inició a partir de mediados del siglo XX como consecuencia del enorme desarrollo tecnológico y económico acontecido tras el final de la Segunda Guerra Mundial. (…) este fenómeno habría sumido al planeta Tierra en un nuevo estado de cambios drásticos inequívocamente atribuible a las actividades humanas. Así, el enorme crecimiento del sistema económico-financiero mundial, junto al desarrollo tecnológico y al proceso de globalización, habrían posibilitado un acoplamiento a escala planetaria entre el sistema socioeconómico y el sistema biofísico de la Tierra que representaría el comienzo de la era de los

nueva época geológica bautizada como "Antropoceno", caracterizada por el hecho de que la humanidad se ha convertido en uno de los más influyentes factores geológicos. El Antropoceno obliga a redefinir los horizontes del desarrollo tecno-industrial uno de los cuales y más relevantes es la revolución del *big data* y la IA.

2. ÉTICA PARA LA IA

Hay que recordar una vez más que los artefactos tecnológicos tienen moralidad, no son neutrales (Winner 1980). No se trata de un "mal uso" de la tecnología sino de su mismo diseño y desarrollo. Una tecnología no solo debe ser evaluada por la forma en que contribuye a la eficiencia y la productividad, sino también por la forma en que puede crear ciertas formas de poder y autoridad. Las tecnologías transforman los objetos pero igualmente los hábitos, costumbres o relaciones.

Precisamente, la IA es una tecnología disruptiva porque transforma profundamente los sistemas, ya sean sociales, económicos o naturales. Y esta vocación transformadora crea conflictos éticos en múltiples fases del desarrollo tecnológico. Su auto-organización compleja crea propiedades emergentes que tienen efectos incontrolados y un fuerte impacto en la sociedad, en los individuos y en el medio ambiente.[3]

Asimismo, la revolución digital, el *big data* y el desarrollo de sistemas y máquinas cada vez más inteligentes o, al menos, capaces de realizar tareas que hasta ahora solo los seres humanos ejecutaban, van a reconfigurar la democracia y la organización social humana como ahora las conocemos. Este nuevo ecosistema tiene aplicaciones extensivas y el potencial de cambiar nuestra vida, derechos y libertades fundamentales. Se trata de un claro contexto de ciencia "post-normal" (Funtowicz & Ravetz 2000) donde los desafíos que introduce la moderna tecnociencia (IA, biología sintética, interfaces cerebro-máquina, robótica, *big data*) producen importantes desacuerdos entre los expertos e implican decisiones cruciales a nivel individual y colectivo y la asunción de riesgos en contextos de incertidumbre. Es un contexto caracterizado por la incertidumbre sobre los hechos, los valores en disputa, los enormes desafíos (de carácter sistémico, como las pandemias o la emergencia climática) y la necesidad de tomar decisiones urgentes (debemos gestionar el

humanos." [Antropoceno]. Mateo Aguado, *Vivir bien en un planeta finito* (tesis doctoral). Tomado de Riechmann (2016): 73.

[3] The Ethics of Socially Disruptive Technologies Research Programme: https://www.esdt.nl/ [consultado el 26/03/2021].

desconocimiento). Por todo ello, se requiere debate ético, deliberación pública, transparencia y políticas; esto es, buena gobernanza.

Precisamente, la UE ha formulado una estrategia general de investigación e innovación responsables (RRI: *responsible research and innovation*). La RRI es una retórica radical sobre la apertura y socialización de los procesos tecnocientíficos que se concreta en cuatro principios de gobernanza: anticipación, reflexividad, deliberación y responsabilidad. Esta apertura y socialización se concretan en la idea de participación pública, que es esencial para los investigadores en términos de mejores soluciones éticas sobre cuestiones difíciles que facilitan la aceptabilidad y la confianza en la investigación.

La RRI pretende un nuevo modelo de gobernanza de la investigación que reduzca la brecha entre la comunidad científica y la sociedad, incentivando que los distintos grupos de interés (comunidad educativa, científica, empresa, industria, entidades de la sociedad civil, política) trabajen juntos en todo el proceso de investigación e innovación. La idea es que la cooperación entre distintos actores (relativamente autónomos pero altamente interdependientes) permita alinear el proceso de investigación y sus resultados con los valores, necesidades y expectativas de la sociedad.

Aunque el término RRI se acuñó hace una década, recientemente, ha cobrado protagonismo debido a su inclusión en el Programa SWAFS *Science with and for Society* (ciencia con y para la sociedad) impulsado por la Comisión Europea en el marco de la estrategia de investigación Horizonte 2020.

La RRI comprende 6 agendas políticas a tener en cuenta durante todo el proceso de investigación e innovación:

- a) Participación ciudadana, para fomentar que múltiples actores se involucren en el proceso de investigación desde su concepción hasta su desarrollo y obtención de resultados.
- b) Igualdad de género, para promover el equilibro entre hombres y mujeres en los equipos de trabajo.
- c) Educación científica para mejorar los procesos educativos y promover vocaciones científicas entre los y las más jóvenes.
- d) Ética para fomentar la integridad científica, con el fin de prevenir y evitar prácticas de investigación inaceptables.
- e) Acceso abierto a la información científica, para mejorar la colaboración entre grupos de interés y el diálogo abierto con la sociedad.
- f) Acuerdos de gobernanza, para proporcionar herramientas que fomenten la responsabilidad compartida entre grupos de interés e instituciones.

3. LO VIEJO

Existe ya una larga reflexión sobre aspectos éticos de la ciencia y la tecnología en autores como Evandro Agazzi, Jürgen Habermas, Gilbert Hottois o Carl Mitcham entre otros (Echeverría 2007; Linares 2008).[4] En este sentido, la IA hereda una serie de deberes éticos que comparte con otras tecnociencias. Si cabe, estos deberes son más acuciantes en el caso de la IA en tanto en cuanto, como hemos visto, el potencial transformador de esta actividad, especialmente en convergencia con otras tecnologías físicas, digitales y biológicas, es muy grande (disruptivo).

Estos deberes éticos se plantean como requisitos mínimos y esenciales que protegen al individuo, a la sociedad y al medio ambiente de los daños e impactos negativos que puede ocasionar la IA. Se entienden como un amparo y escudo frente a la fragilidad y vulnerabilidad que se puede experimentar con relación a la IA —desde una perspectiva ética del cuidado que conlleva obligaciones o deberes para impedir, minimizar o mitigar el daño y las áreas o espacios de vulnerabilidad.[5]

3.1. Minimizar los daños (seguridad y protección)

Es decir, el deber de no hacer daño, que abarca no sólo las acciones intencionadas sino también la imposición de riesgos de daño (por lo que se aplican la precaución y la prevención). Así pues, la seguridad es un valor social que debe destacarse cuando los seres humanos interactúan con cualquier tecnología y, claro, también cuando lo hacen con sistemas autónomos. Una norma sobre seguridad debe establecer un procedimiento claro para medir, probar y certificar la funcionalidad de un sistema de IA basado en datos. Daño no solo relativo a los individuos sino también a los colectivos, con especial atención a los individuos y grupos vulnerables y a los riegos existenciales o sistémicos.

3.2. Maximizar los beneficios

Los beneficios potenciales deben superar el posible riesgo para los sujetos, las comunidades y el medio ambiente. Esto se relaciona con la promoción del bienestar, el bien común y los objetivos de desarrollo sostenible (Vinuesa et al. 2020).

[4] https://es.unesco.org/themes/etica-ciencia-y-tecnologia/ [consultado el 26/03/2021].

[5] Establezco estos deberes mínimos desde una perspectiva ética del cuidado (Tronto 1993) si bien podrían justificarse también desde otros enfoques como el consecuencialismo, la ética de las virtudes o el deontologismo (Boddington 2017).

Los tratamientos de *big data* a través de la IA han de acreditar su necesidad, efectividad y orientación al bien común, acotando un propósito bien definido y determinado (proporcionalidad).

Se han de compartir los beneficios con las poblaciones desfavorecidas, especialmente si la investigación se realiza en países en desarrollo.

3.3. Respeto y autonomía

Los procesos de IA han de estar centrados en las personas y respetar el derecho de las mismas a conservar la facultad de decidir qué decisiones tomar y ejercer la libertad de elegir cuando sea necesario.

El principio de respeto conlleva escuchar y prestar atención a los individuos no solo como sujetos pasivos en relación con el uso y reutilización de sus datos sino como colaboradores y participantes activos (control, docilidad) en una estrategia para innovar y generar mayor valor público.

Asimismo, se protegerá a los trabajadores que analizan y entrenan los algoritmos, reconociendo sus derechos laborales y vigilando las cadenas de subcontratación. Es un hecho que la IA utiliza "trabajadores fantasma" (*ghost workers*), que realizan micro-trabajos en pésimas condiciones laborales para la gestión de ingentes cantidades de datos, imágenes o textos y para "entrenar" a los algoritmos.[6]

3.4. Garantizar la privacidad y la identidad personal

Sofisticados algoritmos y técnicas de Inteligencia Artificial se aplican a grandes cantidades de datos para encontrar patrones estadísticos recurrentes que pueden ser usados para predecir y entender el comportamiento de las personas (perfilado) y manipularlas neuro-emocionalmente ("troquelado de mentes"). Estos sistemas basados en IA pueden vigilar y modificar adaptativamente el cerebro, transformando la experiencia fenomenológica del usuario, lo que afecta a su propio sentido de autonomía e identidad y, en última instancia, a la forma en que se ven a sí mismos y a sus relaciones con los demás.

En consecuencia, se ha de proteger a los individuos contra el uso coercitivo de estas tecnologías y la posibilidad de que la tecnología pueda ser utilizada sin su consentimiento. Por un lado, se protegerán los datos de carácter personal (privacidad) evitándose la re-identificación (los individuos pueden volverse identificables a partir de datos que, en primera instancia, son anónimos), la fuga de datos y la falta de transparencia en la recogida de datos.

---

[6] Trabajadores fantasma / *click workers*: https://www.rtve.es/rtve/20200910/red-controla-estreno-nueva-temporada-noche-tematica/2041828.shtml [consultado el 26/03/2021].

Por otro lado, se preservará la identidad personal y la continuidad del comportamiento personal frente a modificaciones no consensuadas por terceros (privacidad mental) —en el marco de lo que se han llamado nuevos "neuroderechos" protectores de la libertad cognitiva o autodeterminación mental: el derecho a la privacidad mental, el derecho a la integridad mental y el derecho a la continuidad psicológica (Ausín, Morte, Monasterio 2020).

3.5. Proteger del medio ambiente y las generaciones futuras

El desarrollo de la tecnología y de los entornos digitales, entre ellos las infraestructuras vinculadas a la IA (data centers, nubes, centros de cálculo), deberá perseguir la sostenibilidad medioambiental y el compromiso con las generaciones futuras. Por ello, se promoverá la eficiencia energética en el entorno digital, favoreciendo la minimización del consumo de energía y la utilización de energía renovable y limpia.

La IA y la economía digital en general se presentan como un desarrollo respetuoso con la protección de medio ambiente, que no abusa de las materias primas escasas y que genera poca contaminación. Sin embargo, no hay "desmaterialización" sino que más bien la fabricación y mantenimiento de redes y productos electrónicos supera con creces a otros bienes de consumo. Por ejemplo, el gasto en combustibles fósiles utilizados en la fabricación de un ordenador de sobremesa supera las 100 veces su propio peso mientras que para un coche o una nevera la relación entre ambos pesos (de los combustibles fósiles usados en su fabricación y del producto en sí) es prácticamente de uno a uno.

Pero además, los grandes centros de computación y de almacenamiento de datos en la nube requieren enormes cantidades de energía y tienen una alta huella por emisiones de CO2, con un impacto medioambiental muy elevado. El consumo eléctrico es tan grande que las emisiones de carbono asociadas son ingentes, colosales, hasta el punto de poner en cuestión todo el desarrollo de la IA (Winfield 2019).

3.6. Promover la inclusión y la justicia

La justicia en las políticas de ciencia y tecnología se refiere de manera central a dos cuestiones básicas: el reconocimiento y la forma de asignar los escasos recursos a las personas.

El reconocimiento (inclusividad) supone tener en cuenta a todos los interesados (la diversidad de agentes y valores) en el proceso. Como señala la RRI, en todas las etapas de los procesos de investigación e innovación deben utilizarse metodologías inclusivas y participativas (Monasterio Astobiza et al. 2019)

La asignación requiere la búsqueda de la igualdad (Sen 1992) y la accesibilidad (equidad). Se ha de evitar una nueva brecha de desigualdad a causa de

la IA, sobre la ya existente brecha digital, y por ello se impulsarán los algoritmos abiertos y el conocimiento compartido para favorecer la distribución de los beneficios de la IA (Luengo-Oroz 2019):

> Solidarity as an AI principle should imply the following: (1) sharing the prosperity created by AI, implementing mechanisms to redistribute the augmentation of productivity for all, and sharing the burdens, making sure that AI does not increase inequality and nobody is left behind; and (2) assessing the long-term implications before developing and deploying AI systems. Solidarity should be a core ethical principle in the development of AI.

4. LO NUEVO

Junto con los anteriores deberes éticos, podemos añadir tres elementos novedosos que, de algún modo, singularizan más la reflexión ética sobre la IA.

4.1. Explicar y auditar

Uno de los principales requisitos éticos de la IA y la ciencia de datos es la explicabilidad, ya que la toma de decisiones basadas en IA puede afectar a ámbitos sensibles e importantes de la vida (como la atención sanitaria, los derechos civiles y sociales, el derecho penal, el crédito...). Por tanto, los algoritmos que manejan los datos deben ser auditables.

Para conciliar este principio con la protección de la innovación científica e industrial y la propiedad intelectual, debiera considerarse la atribución de competencias de fiscalización a autoridades independientes. Más aún, cuando se pongan en juego valores fundamentales como los mencionados (salud, vigilancia policial, sistema penal, ayuda social…), los sistemas deberían ser trazables, esto es, se podrá identificar el proceso de decisión, las personas implicadas y las consecuencias que se deriven.

Sin embargo, las aplicaciones de aprendizaje profundo tienen el problema de que no pueden explicar fácilmente el proceso.

En todo caso, más allá de la mera transparencia, se facilitará la comprensión e interpretabilidad de los sistemas de IA, evitando en lo posible las “cajas negras” y actuando proactivamente para ofrecer información a la ciudadanía.

4.2. Evitar sesgos

Unido a lo anterior, se han de evitar los sesgos tanto en la recopilación de los datos, como en su manejo y en el entrenamiento con datos de algoritmos e instrumentos de inteligencia artificial.

Se entiende por sesgo a un prejuicio en nuestro modo de procesar la información, una tendencia a percibir de modo distorsionado la realidad, una predisposición.

La IA se ha planteado como la panacea para la toma de decisiones más acertada, imparcial y eficiente, que evitaría los errores humanos, como un nuevo paradigma en la obtención de conocimiento. Sin embargo, ha obviado algo básico: que los sesgos no desaparecen nunca aumentando el tamaño de la muestra; por ello son sesgos y no confusores.

Pues bien, los algoritmos que manejan enormes cantidades de datos nos condenan a repetir o incluso empeorar los errores que queríamos evitar, ya que replican y hasta multiplican los prejuicios (los sesgos). Y ello debido principalmente a dos motivos:

- Los sesgos implícitos en los datos.

Las muestras que se recogen para la IA incorporan en los datos seleccionados prejuicios que no hacen sino amplificar el sesgo. Así sucede, por ejemplo, con los bancos de imágenes que se utilizan para entrenar a las máquinas de inteligencia artificial: Si se parte de un corpus sesgado de género, los modelos predictivos aumentan este sesgo. En una reciente investigación sobre el aprendizaje de las máquinas de reconocimiento visual,[7] los varones protagonizaban un 33% de las fotos que contenían personas cocinando. Tras entrenar a la máquina con estos datos, el modelo dedujo que el 84% de la muestra eran mujeres. Conclusión: "Si está en la cocina, es una mujer".[8] Y siguiendo con el sesgo de género, otro estudio de la Universidad Carnegie Mellon descubrió que las mujeres tienen menos posibilidades de recibir anuncios de trabajos en Google.[9]

- El software (el algoritmo) hace suyo el prejuicio o la tendencia subyacente en la sociedad para poder acertar.

Por ejemplo, la policía de varias ciudades norteamericanas usa algoritmos y análisis de *big data* para pronosticar los lugares en los que es más probable que haya delincuencia. De este modo, acuden más a estas zonas, detienen a más gente cometiendo delitos allí y refuerzan el ciclo negativo de las mismas.

---

[7] Zhao et al. (2017): https://scirate.com/arxiv/1707.09457 [consultado el 26/03/2021].

[8] Javier Salas, EL PAÍS 22/09/2017: https://elpais.com/elpais/2017/09/19/ciencia/1505818015_847097.html [consultado el 26/03/2021].

[9] Véase: https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html [consultado el 26/03/2021].

Aunque los algoritmos se presentan bajo una apariencia de neutralidad, el caso es que no dejan de ser opiniones "encapsuladas" (O'Neil 2016). Y los ejemplos de discriminaciones debidas a los algoritmos de IA son cada vez más numerosos y conocidos: El primer certamen de belleza juzgado por un ordenador colocó a una única persona de piel oscura entre los 44 vencedores. Facebook permite a los anunciantes que excluyan de su target comercial a minorías étnicas y que incluyan, en cambio, a jóvenes identificados como vulnerables y depresivos o a personas antisemitas. Y, en un plano todavía más grave, un programa que usa el Departamento de Justicia de USA para pronosticar la reincidencia de los presos, etiqueta doblemente peor a los acusados negros que a los blancos, de modo que aquellos eran tratados más duramente por el sistema penal.[10]

La introducción de sesgos en las inteligencias artificiales da lugar a discriminación (por razón de género, raza, estatus económico…) y puede consolidar prejuicios y estereotipos existentes, reforzando la exclusión social y la estratificación.

Asimismo, en la medida en que cada vez más decisiones en la sociedad se basan en el uso de algoritmos, se puede caer en una "dictadura de datos", donde ya no somos juzgados sobre la base de nuestras acciones reales, sino sobre la base de lo que los datos y la IA indiquen que serán nuestras acciones probables (enfermedades, conductas, accidentes de tráfico, ayudas sociales…).

A todo ello tenemos que añadir que la interacción entre los humanos y las máquinas puede producir además nuevos sesgos, como el llamado sesgo de automatización; esto es, la tendencia de operadores humanos para restarle importancia a sus propios juicios a favor de la información o recomendación ofrecida por la máquina. Por ejemplo, se han detectado sesgos de automatización en los pilotos de avión, así como en las clínicas donde los médicos confían en las herramientas automáticas de diagnóstico más que en otros juicios o percepciones propias.

Finalmente, el miedo a dejar "huellas digitales" —que nos "perfilen"— y que luego puedan perjudicarnos en un futuro (por ejemplo, a la hora de encontrar un trabajo, obtener un préstamo o un seguro, etc.) hace que restrinjamos la búsqueda de puntos de vista alternativos y que limitemos el intercambio libre de opiniones, debilitando la democracia y la libertad de expresión. Esto hace que la población sólo se exponga a contenidos que confirman sus propias actitudes y valores, con los riesgos de polarización que ello comporta. Los motores de búsqueda, por ejemplo, han sido criticados por crear burbujas de filtro alrededor

[10] ProPublica, "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks": https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [consultado el 26/03/2021].

de los usuarios. Este tipo de fenómeno, también conocido como cámaras de eco, puede ser explotado como parte de campañas de desinformación.

4.3. Atribuir responsabilidad

A medida que la IA y la robótica adquieren mayor protagonismo surge la cuestión de la atribución de responsabilidad (que no es sólo una cuestión ética sino también práctica y jurídica). La robótica y la IA no se limitan a un solo artefacto. Es el problema de "muchas manos y muchas cosas conectadas". Entre ellas figuran la ética de la ingeniería de los diseñadores, los fabricantes y los sistemas de mantenimiento, los aspectos éticos de los propios artefactos y las actitudes éticas de los usuarios (ya hemos dicho que constituimos sistemas socio-técnicos híbridos de seres humanos y artefactos autónomos inteligentes). Todos los tipos de interesados que participan en la IA pueden considerarse redes de responsabilidad moral distribuida (responsabilidades compartidas en un proceso evolutivo); la responsabilidad de los artefactos es paralela a la inteligencia de los artefactos. En este sentido, podría considerarse que existen diferentes grados de responsabilidad artefactual, así como también distinguimos grados de competencia humana.

Aquí surgen tres cuestiones entrelazadas:

- La definición de la autonomía de la máquina y la consiguiente determinación de las responsabilidades y obligaciones asociadas a esas entidades autónomas, sus diseñadores y sus gestores.

- El análisis de la eficacia del marco jurídico actual para abordar cuestiones de gobernanza de la IA.

- La necesidad de prever nuevos instrumentos jurídicos para abordar los problemas relacionados con tecnologías emergentes.

5. LO ESPURIO

Una enorme variedad de organizaciones, entidades e instituciones han desarrollado toda una pléyade de códigos éticos, recomendaciones, documentos y guías de buenas prácticas éticas para la IA (Jobin, Ienca, Vayena 2019). De acuerdo con Floridi, existen más de 70 recomendaciones sobre ética e IA formuladas en los últimos años (Floridi 2019) y sobre las más variadas cuestiones en el campo de la IA: agencia, autonomía, moralidad, derechos, valores, virtudes, confianza, transparencia, rendición de cuentas…

Como dice acertadamente Margarita Robles (2020), esta proliferación y trabajo intenso indica el apreciable interés que se otorga a la dimensión ética de la IA pero, desgraciadamente, genera confusión, duplicación del trabajo y ruido, reduciendo las posibilidades de transparencia del mismo debate. Más aún, el

diálogo ético internacional sobre la IA está dominado por una minoría de países con un alto nivel de desarrollo tecnocientífico, lo que excluye de esta deliberación a otros muchos países y agentes que luego se verán sin duda afectados.

También sucede que se interpreta la ética de la IA como un recetario o conjunto de principios y buenas intenciones con respecto a su uso, renunciando al sentido primigenio y profundo de la reflexión ética que no es otro sino interrogar, analizar y cuestionar el sentido mismo de los desarrollos tecnocientíficos (no tanto su uso), por lo que la ética debe de estar presente desde el mismo diseño de la tecnología y en constante diálogo con su desarrollo (*ethics by design*). La ética es básicamente pregunta.

Pero lo peor es la utilización por parte de las grandes empresas tecnológicas de la ética como coartada para evitar las regulaciones (Ochigame 2019). Existe un discurso casi ubicuo de la "ética para las tecnologías disruptivas" que, por ejemplo, no estaba en los orígenes de la IA y que de repente se ha instalado con fuerza. Un discurso que muchas veces se alinea con los esfuerzos de Sillicon Valley por evitar restricciones legales a tecnologías controvertidas (*whitewashing*). Millones de dólares de Facebook, Amazon, Microsoft, IBM… nutren centros e investigaciones en "IA ética", entendida como prácticas responsables y voluntarias (frente a la regulación legal). La mayoría de los documentos relativos a ética e IA descansan en la auto-regulación y los compromisos individuales en vez de promover una mejora de las regulaciones. Se obvian de este modo cuestiones muy relevantes con relación a los riesgos y oportunidades de la IA: cuestiones de poder, de pluralismo razonable, de legitimidad, de propiedad o de bien común.

En este sentido, la ética (y la subsiguiente regulación legal) no deberían contemplarse como frenos a la capacidad de innovación sino más bien como promotoras de un uso socialmente responsable de la IA que contribuya a la paz, la prosperidad, el bienestar y la justicia, en consonancia con los Objetivos de Desarrollo Sostenible y la Agenda 2030.

6. CONCLUSIONES: PENSAR ÉTICAMENTE LA IA

Existe una enorme ambivalencia con respecto a la IA. Por un lado, se plantean grandes esperanzas y expectativas, próximas a un "solucionismo" tecnológico según el cual todos nuestros problemas (económicos, sociales, ambientales) se arreglarán con el desarrollo y uso de la IA y las demás tecnologías convergentes. Por otro lado, existe un enorme desconocimiento social (en la ciudadanía, en las empresas, en las instituciones y organizaciones) sobre la IA, sobre sus posibilidades, sus potenciales beneficios y sus riesgos. No se percibe socialmente la importante cadena de valor que proporciona usar, compartir y reutilizar los datos

mediante algoritmos de IA, para generar políticas públicas anticipatorias, servicios, actividad económica y acciones de denuncia social, tanto a nivel macro (instituciones y administraciones), como meso (organizaciones y empresas) y micro (asociaciones y personas). Y el desconocimiento se traduce en una percepción distorsionada de esta tecnología, produciendo miedo, desconfianza y rechazo.

Para lograr el objetivo de una IA justa y socialmente responsable es preciso, en primera instancia, superar las barreras o los condicionantes denominados "no tecnológicos" que pueden intervenir en la aplicación de las tecnologías. Es decir, es necesario generar confianza entre los usuarios y las empresas e instituciones que desarrollan las tecnologías. Es necesario que los usuarios acepten y se apropien socialmente de estas tecnologías. La persistencia de la desconfianza terminará convirtiéndose en incertidumbres y riesgos que pueden afectar de manera decisiva al desarrollo y la utilización de la tecnología. Inicialmente, la introducción de una tecnología en un determinado entorno es sólo una "novedad"; esta novedad se convierte en "innovación" cuando es adoptada por la comunidad o el grupo social al que va dirigida. Las características de este proceso de adopción son muy variadas, ya que el acercamiento que una comunidad o grupo hace a la innovación propuesta se produce a través de conjuntos de prácticas, representaciones sociales y valores éticos sobre dicha innovación.

En este sentido, un elemento fundamental para la aceptación social y apropiación de la IA es la ética. La práctica del uso de una tecnología está estrechamente relacionada con los principios éticos que incorpora y es la condición básica para la apropiación y aceptación de la tecnología por parte de la comunidad. Así, la reflexión ética sobre la IA constituye un instrumento privilegiado para el imprescindible empoderamiento tecnológico de la ciudadanía. Asimismo, y como necesario complemento de la reflexión ética, los poderes públicos y, en general, las instituciones y otras organizaciones, han de promover la formación y sensibilización de la sociedad sobre el potencial de la IA para generar valor económico y social, haciendo hincapié en los deberes de cuidado y protección mencionados anteriormente y, a la vez, en las oportunidades que un tratamiento correcto y ético de la IA ofrece para la investigación biomédica, la salud, la gestión administrativa, los servicios sociales, la atención a colectivos desfavorecidos, el desarrollo económico, la innovación…; esto es, para el bien común —que legitima precisamente este uso socialmente responsable de la IA y la ciencia de datos masivos.

Sólo así será posible un nuevo pacto "tecno-social" (entre ciudadanos, organizaciones y estados) que, basándose en una reflexión ética, evite las injusticias algorítmicas (por falta de inclusividad, por desigualdad, por discriminación) y promueva en bien común. El desafío es desarrollar la IA (y demás tecnologías disruptivas) con equidad y participación, alineada con el modelo europeo de investigación e innovación responsables (RRI) y que favorezca su apropiación social en términos de lo que se ha denominado *engaging technologies* o "tecnologías

entrañables" (Quintanilla et al. 2017): abiertas, versátiles (interoperables), controlables, comprensibles, sostenibles, respetuosas con la privacidad, centradas en las personas y socialmente responsables (con especial cuidado de los individuos y colectivos más desfavorecidos). De ese modo, se generarán entornos tecnológicos "amigables", que no aíslen del mundo fuera de línea ni de nuestro cuerpo y que incrementen la conexión social que contribuye efectivamente a la vida comunal.

REFERENCIAS BIBLIOGRÁFICAS.

Ausín, T.; Morte, R.; Monasterio Astobiza, A. (2020). Neuroderechos: Derechos Humanos para las neurotecnologías. *Diario La Ley*, Nº 43, Sección Ciberderecho, 8 de Octubre de 2020, Wolters Kluwer.

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Oxford: Springer.

Echeverría, J. (2007). *Ciencia del bien y del mal*. Barcelona: Herder.

Espinosa, L. (2020). El ser humano y los algoritmos. *Dilemata. Revista Internacional de Éticas Aplidadas*, 33, 235-250.

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, 32.2, 185-193.

Funtowicz, S.O. & Ravetz, J.R. (2000). *La ciencia posnormal: ciencia con la gente*. Barcelona: Icaria.

Jobin, A.; Ienca, M.; Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence,* 1, 389-399.

Linares, J.E. (2008). *Ética y mundo tecnológico*. México: FCE.

Luengo-Oroz, M. (2019). Solidarity should be a core ethical principle of AI. *Nature Machine Intelligence*, 1, 494.

Monasterio Astobiza, A. et al. (2019). Bringing inclusivity to robotics with INBOTS. *Nature Machine Intelligence* 1**,** 164.

Ochigame, R. (2019). The invention of 'Ethical AI'. How Big Tech Manipulates Academia to Avoid Regulation, *The Intercept*: https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/ [último acceso: 26 marzo 2021].

O'Neil, C. (2016). *Armas de destrucción matemática*. Madrid: Capitán Swing.

Quintanilla, M.A. et al. (2017). *Tecnologías entrañables*. Madrid: Catarata.

Ramió, C. (2019). *Inteligencia artificial y administración pública*. Madrid: Catarata.

Riechmann, J. (2016). *¿Derrotó el smartphone al movimiento ecologista?* Madrid: Catarata.

Robles, M. (2020). Artificial Intelligence: From ethics to law. *Telecommunications Policy*, 44.6, DOI: https://doi.org/10.1016/j.telpol.2020.101937

Sen, A. (1992). *Inequality Reexamined*. Harvard University Press.

Tronto, J. (1993). *Caring Democracy: Markets, Equality and Justice*. New York University Press.

Vinuesa, R. et al. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11.1, 1-10.

Winfield, A. (2019). On the simulation (and energy costs) of human intelligence, the singularity and simulationism. En A. Adamatzky y V. Kendon (eds), *From Astrophysics to Unconventional Computation* (pp. 397-407). Cham: Springer

Winner, L. (1980). Do artifacts have politics? *Daedalus,* 109.1, 121–136.