EL DISEÑO ÉTICO DE LA INTELIGENCIA ARTIFICIAL PARA NO DISCRIMINAR NI LESIONAR DERECHOS[1]

Idoia Salazar*

1 Estudio realizado en el marco del proyecto MICINN "Claves para una justicia digital y algorítmica con perspectiva de género" PID2021-123170OB-I00, U. Valencia.

* Presidenta de OdiseIA. Profesora de la Universidad CEU San Pablo.

I. INTRODUCCIÓN

El hecho de que ciertos sistemas Inteligencia Artificial (en adelante IA) necesitan incluir, desde su diseño inicial, límites éticos es cada vez menos cuestionado a nivel internacional. Sin embargo, ser drástico al respecto e intentar incluir principios universales, no parece que sea lo más apropiado en este caso. El sistema en cuestión debe personalizarse con unos u otros límites éticos en función de la labor que va a desempeñar (caso de uso), la situación específica o el ámbito geográfico en el que se implementa. Organismos como el Parlamento Europeo han aprobado guías y recomendaciones referentes a la ética en la Inteligencia Artificial. Todos ellos tienen en común la protección de los derechos fundamentales de los seres humanos, especialmente en cuanto a privacidad, libertades en la toma de decisiones o el riesgo de manipulación.

Entre otras iniciativas, a nivel internacional, dedicadas a la formulación de principios éticos para estas peculiares tecnologías autónomas, como el documento *Ethically Aligned Design* (IEEE, 2020) elaborado por el IEEE (*Institute of Electrical and Electronics Engineers*), la cumbre global *AI for Good* organizada por ITU (*International Telecommuication Union*) o las conferencias internacionales AAAI/ACM sobre Inteligencia Artificial, Ética y Sociedad. También empresas privadas como IBM, Microsoft y Google Deep Mind, entre otras muchas, han desarrollado sus propios códigos éticos en IA y, ante los grandes retos que supone esta tecnología, unen sus fuerzas con iniciativas como la llamada *Partnership on AI* u *Open AI*.

Asimismo, los 193 Estados miembros de la Conferencia General de la UNESCO adoptaron en noviembre de 2021 el primer acuerdo mundial sobre la Ética en la IA (UNESCO, 2021). Su objetivo: promover los derechos humanos y la dignidad humana. Además, pretende sentar una base normativa global que permita velar por el Estado de derecho en el mundo digital. Para conseguir estos objetivos la organización recomienda políticas específicas tanto nacionales como internacionales, así como marcos regulatorios para garantizar que estas tecnologías emergentes beneficien a la humanidad en su conjunto.

Todas las iniciativas de ética e IA creadas hasta ahora tienen una base común: la IA debe estar al servicio de los intereses de los ciudadanos, y no al revés. Es decir, el humano siempre como centro.

II. INTELIGENCIA ARTIFICIAL. VENTAJAS Y OPORTUNIDADES

El término sistema de inteligencia artificial no es simple de definir. Según la propuesta de regulación de la Comisión Europea, se define como: un sistema diseñado para funcionar con un cierto nivel de autonomía y que, basándose en datos proporcionados por máquinas y/o personas, infiere cómo lograr un conjunto determinado de objetivos utilizando enfoques basados en el aprendizaje de máquinas y/o en la lógica y el conocimiento. Así, produce resultados generados por el sistema, como contenidos, predicciones, recomendaciones o decisiones, que influyen en los entornos con los que interactúa (UE, 2022).

Otros se refieren al concepto general de Inteligencia Artificial como: el estudio de la informática que se centra en el desarrollo de estrategias y tácticas por el cual sistemas y máquinas manifiestan una inteligencia humana (Briega, 2017).

En definitiva, son muchos los que, a lo largo de los siglos XX y XXI, han intentado abordar el intento de dotar de significado a estas dos palabras: Inteligencia Artificial.

Muchos han sido los que, desde entonces, han intentado que sus programas informáticos superaran el *Test de Turing*. En 1990 se creó el concurso Premio Loebner en el que intentaban superarlo, lo cual no fue posible hasta 2010 cuando el robot Suzette, de Bruce Wilcox, lo consiguió. Suzette tenía 16.000 reglas de conversación y era capaz de mantener 40 horas de conversación ininterrumpida. Tenía una personalidad coherente y respondía emocionalmente. Hoy día las teorías de Turing siguen aplicándose a los fundamentos de la robótica y la inteligencia artificial.

El estadounidense Marvin Minsky también revolucionó este campo durante el siglo XX y sus descubrimientos y pensamientos supusieron un antes y un después. En 1956 acuñó el término "Inteligencia Artificial" –ya ideado por Turing– y estableció sus fundamentos guiado por el sueño visionario de dotar a los ordenadores con la capacidad para razonar. Él la definía de esta manera: "es la ciencia de hacer que las máquinas hagan cosas que requerirían inteligencia si las hubiera hecho un humano". Dos años más

tarde ingresó como profesor en el *Massachusetts Institute of Technology* (MIT), en Boston, donde trabajó más de cuatro décadas y donde creó su Laboratorio de Inteligencia Artificial, incubadora de muchos inventos robóticos actuales. Creó el primer prototipo de una máquina capaz de aprender de manera autónoma *Snarc*, diseñó las primeras manos con sensores táctiles y formó parte del equipo que diseñó ARPAnet, el precursor de lo que hoy conocemos como Internet.

A pesar de estos avances durante el siglo XXI, la verdadera ola actual de progreso y entusiasmo por la IA comenzó alrededor de 2010, impulsada por cuatro principales factores, íntimamente ligados el uno al otro (Salazar, 2019):

— La disponibilidad del *big data*, gratis o a precio reducido, para el comercio electrónico, empresas, redes sociales, medios de comunicación, ciencia y gobierno.

— Esta materia prima (*big data*), de la que se nutren la IA, mejora considerablemente el llamado *machine learning* y los algoritmos.

— El procesamiento de datos mejora exponencialmente gracias a las altas capacidades de los ordenadores.

— La recuperación económica y el aumento significativo, por parte de las empresas tecnológicas, en inversión en IA. La confianza en las oportunidades de esta nueva tecnología es extremadamente alta. Además, tiene aplicaciones transversales en prácticamente cualquier área.

Desde entonces los progresos en IA han sido sorprendentes. Entre los avances más destacados está el reconocimiento de imágenes, con casi un 100% de fiabilidad o las mejoras en el lenguaje natural en el que son capaces de expresarse y comprender los 'robots'. Sin embargo, aún nos encontramos en una fase muy temprana de la IA, la conocida como IA débil, consistente en desarrollos muy específicos como juegos estratégicos, traducción de idiomas, reconocimiento de imágenes, o los primeros pasos en la autonomía de los vehículos. Aun así, en el momento actual, ya ofrece grandes ventajas, por ejemplo, en los diagnósticos médicos, sistemas de recomendación y orientación de anuncios o planificadores de viajes.

La llamada IA general o IA fuerte se refiere a la Inteligencia Artificial del futuro, en el que el comportamiento inteligente de la máquina estará mucho más desarrollado en tantas competencias como podría estarlo en una persona. Este es uno de los grandes miedos que despierta esta tecnología en la actualidad: el hecho de que los robots lleguen a ser algún día más inteligentes que los humanos. Frente esto hay opiniones de expertos que auguran que estos sistemas inteligentes del futuro podrían crear otros algoritmos aún más inteligentes que ellos, hasta llegar a la llamada 'Singularidad Tecnológica', el momento en que los robots dotados de IA superarían en todos los aspectos a la raza humana. Tomarían sus propias decisiones independientemente del criterio humano y, en el caso de descontrolarse, el mundo se sumiría en un caos, con pocas opciones para las personas.

Existe una visión más positiva sostenida por muchos investigadores que ve en el desarrollo de la IA como un compañero de viaje de los humanos. Una ayuda. Un asistente capaz de operar de manera segura y con ética.

En cualquier caso, es muy necesario que los distintos gobiernos y organismos internacionales conozcan muy bien estas tecnologías., sus ventajas y sus posibles riesgos, y ayuden a minimizar el posible impacto negativo con suficiente antelación.

Mientras avanzamos en una dirección u otra, la evolución exponencial de la IA continúa, con retos y riesgos, que serán necesarios controlar si queremos aprovechar también todas las ventajas que trae esta tecnología.

III. RIESGOS Y RETOS REALES DE LA INTELIGENCIA ARTIFICIAL

Existen muchos prejuicios en torno a la Inteligencia Artificial alimentados, en buena parte, por la ciencia ficción. Estas cuestiones pueden suponer, a medio plazo, la ralentización de su implantación a nivel internacional, motivado por la falta de confianza. Algunas carecen de fundamento. Otras no. Los riesgos actuales más relevantes de estas tecnologías de IA empezarían por los 'efectos secundarios' no deseados. Los denominados FATE (*Fair, Accountable, Transparent and Explainable*). En estas cuestiones, la posible discriminación es uno de los puntos más importantes por su impacto social. Aunque el aprendizaje

automático (*machine learning*) es capaz de resolver tareas complejas con un alto rendimiento, puede utilizar información que, desde el punto de vista de la sociedad o de la humanidad, no es deseable. Por ejemplo, conceder un préstamo a personas en función de la raza o la religión. Aunque es posible eliminar esos atributos "no deseados" de los conjuntos de datos, hay otros atributos menos obvios que están altamente correlacionados con esos atributos "no deseados" cuya eliminación es menos sencilla. El aprendizaje automático es objetivo y encuentra cualquier relación que haya en los datos, independientemente de las normas y valores específicos.

Otro de los grandes riesgos es que el conjunto de los datos utilizados para entrenar al sistema de IA esté sesgado de manera consciente o inconsciente. El aprendizaje automático basa sus conclusiones en los datos. Si no se usan muestras lo suficientemente representativas de todos los sectores implicados la conclusión no será válida. Aparte del sesgo en los datos de entrenamiento, el sesgo también puede provenir del algoritmo. Un algoritmo de *Machine Learning* trata de ser lo más preciso posible al ajustar el modelo a los datos de entrenamiento. Si no se elige bien el modelo puede provocar "falsos positivos" y/o "falsos negativos".

La explicabilidad es otro de los grandes problemas a los que se enfrentan los sistemas de IA de *Deep Learning* en la actualidad. Para algunas aplicaciones, la explicación de las decisiones es una parte esencial de la propia decisión, y su falta hace que la decisión sea inaceptable. Por ejemplo, un sistema de IA que ayuda a un juez en su toma de decisiones, extrayendo conclusiones sobre un litigio entre un cliente y una aseguradora de salud es inaceptable sin la explicación de la decisión. Esto se conoce como el problema de la "interpretabilidad" (O´Neil, 2016).

Otras de las cuestiones de mayor relevancia en cuento a los efectos secundarios en el uso de la IA son: privacidad, transparencia y control de los datos. Todos los datos y sistemas de IA explotan datos, y muchas veces estos son datos personales. Utilizarlos tiene como efecto secundario que la privacidad puede verse comprometida, aunque sea de forma involuntaria. El escándalo de *Cambridge Analytica*/Facebook (Wylie, 2020) muestra que este es un problema mayor de lo que podríamos haber pensado. En este caso, ocurrido en 2018, El investigador de la Universidad de Cambridge

Aleksandr Kogan, recopiló datos de decenas de miles de usuarios de Facebook y sus amigos, a través de una aplicación que cumplía con las condiciones de Facebook estipuladas para una investigación concreta. A través de la app consiguió información sobre más de 70 millones de usuarios de esta red social. El problema empezó cuando este investigador transfirió, de manera ilegal (es decir, en contra de la política de datos de Facebook), estos datos a la empresa *Cambridge Analytica*, especializada en el marketing político *online*, sobre todo en campañas electorales, usando tecnologías de *big data*. Esta empresa prestó servicios a la campaña de Trump para influir en las elecciones americanas de 2016 y también a la campaña pro *Brexit* en el referéndum del Reino Unido.

III.1. La explicabilidad

Otro efecto secundario importante, se presenta con la explicabilidad. En la actualidad, los sistemas de IA dependen íntimamente de su programador inicial. Es esta persona la que decide qué sistemas usar y con qué datos entrenar el modelo. Por lo tanto, en el tema de la "explicabilidad" de las conclusiones a que llegan estos sistemas, el primer responsable es el ingeniero.

Los resultados de los algoritmos más potentes de la IA, los de *deep learning* son difíciles de entender, ya que el proceso que desarrolla es opaco. Se les conoce como *Black Box*, caja negra. Sin embargo, hay muchos algoritmos de aprendizaje automático que son perfectamente inteligibles (cajas blancas) para saber cómo ha llegado a una determinada conclusión. Por ejemplo, los llamados "árboles de decisión" son perfectamente comprensibles.

Aunque, hasta la fecha, la mayoría de los ingenieros prefieren usar *deep learning* (una caja negra) por sus prestaciones superiores, cada vez son más las empresas que consideran usar un algoritmo interpretable (caja blanca) si su prestación es suficientemente buena para resolver el problema de negocio, aunque sea su rendimiento un poco peor que un algoritmo de caja negra. La razón es la explicabilidad, que, en ciertas aplicaciones, debe tener un peso importante, como en el sector médico. De todas formas, hoy en día hay mucha investigación trabajando en mejorar la explicabilidad de los sistemas de *deep learning* (Darlington, 2017)

III.2. Responsabilidad

Por otro lado, cuando los sistemas se vuelven autónomos y autodidactas, la responsabilidad del comportamiento y las acciones de esos sistemas se hace menos evidente. En el mundo anterior a la IA, el uso incorrecto de un dispositivo es responsabilidad del usuario, mientras que el fallo del dispositivo es responsabilidad del fabricante. Cuando los sistemas se vuelven autónomos y aprenden con el tiempo, algunos comportamientos podrían no ser previstos por el fabricante. Por lo tanto, no está claro quién sería responsable en caso de que algo vaya mal. Un ejemplo de ello son los coches sin conductor. Según la Comisión Legislativa del Reino Unido la culpa pasaría a ser del desarrollador o fabricante del coche (Raposo, 2020). Este argumento se relaciona con la fabricación del propio producto, el creador hace verificaciones y distintas pruebas con el fin de que cuando sea lanzado al mercado, sea un producto o servicio totalmente seguro. Sin embargo, se tiene que analizar en cada producto o servicio el grado de inteligencia artificial que tiene para poder responder a la pregunta. Por otro lado, presenta una amenaza a la democracia, y crea constantemente las denominadas cámaras de *eco online* (CDE, 2020). Consiste en sistemas que se basan en los comportamientos previos de una persona a nivel digital, y luego archiva esos datos con el fin de mostrarles contenido hiperpersonalizado y un entorno único en el que lo que la persona observa es lo que desea. Por lo cual, el entorno habitual en el que los seres humanos nos movemos va desapareciendo. El debate, el público pluralista, la diversidad de contenidos son cosas que poco a poco se vuelven menos frecuentes. Todo esto conlleva a la separación y polarización a nivel público, puesto que lo único que hace es perfilar a los seres humanos, enseñándoles únicamente contenido de su interés, creencias, gustos, y acostumbrándolos aquello. Se puede evidenciar a su vez puesto que el 62,7% de la encuesta realizada, afirma que le gustaría que le muestren únicamente contenido personalizado en base a sus intereses.

III.3. El futuro del trabajo y de las prestaciones sociales

Otro de los grandes riesgos es el relativo al futuro del trabajo y las prestaciones sociales. La IA puede asumir determinadas tareas repetitivas o peligrosas. Pero si esto ocurre a escala masiva, tal vez desaparezcan muchos puestos de trabajo y el desempleo se dispare.

Si cada vez trabaja menos gente, el gobierno recibirá menos impuestos sobre la renta, mientras que los costes de las prestaciones sociales aumentarán debido al incremento del desempleo. Este tipo de cuestiones ha desembocado en sugerencias referentes a un necesario ingreso básico universal (UBI) para todos.

III.4. Relaciones personas-IA

Por otro lado, empiezan a surgir otra serie de cuestiones relacionadas con la interacción 'hombre-máquina', que deben tratarse para que no suponga un riesgo para la persona. Por ejemplo: ¿cuál debería ser la relación (permitida) entre la IA y las personas? ¿Podría un jefe ser un sistema de IA? En Asia, los robots con IA (y sin ella) ya cuidan de las personas mayores, acompañándolas en su soledad. Y, ¿podrían las personas casarse con una IA, como ya ha ocurrido en algunos lugares de Japón?

Algunas de las primeras interacciones serán las siguientes (Benjamins; Salazar, 2021):

a) Interacciones a largo plazo: en las que los robots cohabitan con los humanos en sus hogares, y lugares de trabajo, por ejemplo.

b) Robots en la educación, la terapia, la rehabilitación y el apoyo a las personas mayores. La robótica asistencial, con y sin inteligencia artificial, es un ámbito de aplicación creciente para los robots de servicio. Implica cuestiones críticas de seguridad y ética, por ejemplo, cuando los robots asumen un papel de asistencia a personas vulnerables o con necesidades especiales. Es un hecho que hoy en día, aproximadamente el 10% de la población mundial tiene más de 60 años; en 2050 esta proporción se habrá duplicado con creces. Necesitamos incorporar técnicas de inteligencia artificial para apoyar a los adultos mayores y ayudarles a afrontar los cambios del envejecimiento, en particular el deterioro cognitivo.

c) Interacciones multimodales, expresividad y habilidades conversacionales en las interacciones: la investigación destinada a dotar a los robots de características y cualidades similares a las de los humanos se está ampliando. Se intenta crear robots con una apariencia similar a la nuestra. Y no sólo

esto. También estamos intentando dotar a "sus caras" de expresiones naturales que normalmente sólo corresponderían a los humanos.

d) Cooperación y colaboración en equipos humano-robot (IA): la IA y los humanos no sólo convivirán, sino que deberán trabajar codo con codo, cada uno desarrollando tareas específicas, pero ayudándose mutuamente. Esta interacción, sobre todo al principio, podría ser bastante difícil teniendo en cuenta el concepto de "máquina", y todas sus implicaciones, que tiene la humanidad hoy en día.

e) Detección y comprensión de la actividad humana: esto también será un problema en el futuro próximo para esta interacción. Muchas veces, los humanos hacen cosas siguiendo diferentes razones que podrían estar fuera de la lógica para una máquina.

III.5. Concentración de poder y riqueza en unas pocas empresas muy grandes

La cuestión de la concentración de poder y riqueza en las mayores empresas que generan datos es un claro riesgo. Entre ellas, las GAFAM (Google, Amazon, Facebook, Microsoft, Apple) y algunas grandes empresas chinas, como Alibaba. Esto podría conducir a un oligopolio. Aparte de la falta de competencia, existe el peligro de que esas empresas mantengan la IA como conocimiento propio, sin compartir nada con la sociedad en general más que por el precio más alto posible. Otra preocupación de esta concentración es que esas empresas pueden ofrecer IA de alta calidad como un servicio, basado en sus datos y algoritmos propios (caja negra).

III.6. El uso malicioso de la inteligencia artificial

Las consecuencias negativas citadas hasta ahora son no intencionadas. Sin embargo, pueden existir también las intencionadas. Por ejemplo: los ciberataques, usando IA. En julio de 2021 más de 200 empresas estadounidenses fueron atacadas con el mismo método. Esta vez el ataque no fue directamente dirigido a empresas que dan servicios finales, sino a un suministrador de *software*. Al ser infectado este *software*, todas sus empresas clientes que usan este *software* fueron afectadas. Otro ejemplo son los ataques físicos.

Por ejemplo, la inteligencia artificial puede dar a los drones la capacidad autónoma para volar solos y actuar de manera autónoma como por ejemplo grabar o incluso disparar. En el ámbito militar, los sistemas de armas autónomas letales son otro ejemplo de crear daños físicos. En el futuro también se pueden esperar ataques que tienen tanto un componente digital como un componente físico, por ejemplo, un ciberataque a vehículos autónomos para hacer daño físico con estos vehículos 'secuestrados'. También es posible usar IA para manipular a la opinión pública, de manera maliciosa.

En cualquier caso, es importante estar preparados para poder asumir todos estos riesgos/retos, porque la IA no es el futuro. Es el presente. Cada vez son más las empresas y organizaciones que se suman al uso de la IA como herramienta para mejorar la eficiencia de sus diferentes procesos de negocio. Es importante actuar rápidamente en la ética y normativa para abordar esta tecnología de manera conveniente y conseguir impacto positivo. Veamos la situación internacional de la IA, en este sentido.

IV. SITUACIÓN INTERNACIONAL DE LA INTELIGENCIA ARTIFICIAL

Ya ha quedado lejos la posición de la IA como tecnología del futuro. La automatización, precisión, y rapidez en el análisis de datos complejos son elementos clave que estos sistemas dominan a la perfección.

Además, esta tecnología fomenta el incremento de los ingresos –y minimización de gastos– gracias a la realización de predicciones de alta precisión, basadas en patrones. De hecho, se espera que en 2025 las inversiones en este sector a nivel mundial sean nueve veces superiores a las actuales, pasando de los 6.000 millones de euros a los 52.000 millones (*elEconomista*, 2019)

En el escenario de superabundancia de productos y servicios actual, unido a la también superabundancia de datos almacenados y aquellos generados en vivo (*big data*), cada vez será más normal el desarrollo de sistemas de IA como herramientas que nos ayuden a lidiar con esta cantidad ingente de datos de manera eficiente.

No es de extrañar, por tanto, que cada vez sean más los países que crean estrategias sobre IA a nivel nacional, con diferencias particulares para cada

uno, pero con unos contenidos comunes como son la utilización de la IA en los servicios públicos y el Gobierno; la educación, así como las competencias ligadas a la misma; I+D+I (Investigación, Desarrollo tecnológico, Innovación); las distintas y variadas infraestructuras; y por último (aunque no menos importante) la utilización de manera ética tanto de los datos como de los sistemas inteligentes de cada país.

En este sentido, la mayor parte de las estrategias en IA tienen especial foco en la ética y la refieren a problemas que pueden surgir de la utilización por el público general de la misma en los aspectos de carácter social, económico, político y legal. En cualquier caso, la ética en el uso de la IA sigue siendo un gran reto, como veremos más adelante.

V. SITUACIÓN DE LA INTELIGENCIA ARTIFICIAL EN ESPAÑA Y LA CARTA DE DERECHOS DIGITALES

Centrándonos en España, existe una clara necesidad de impulso a esta tecnología, manteniendo la responsabilidad y la ética. Así, el Gobierno de España publicó a finales de 2020 la 'Estrategia Nacional en Inteligencia Artificial', la cual marca seis prioridades a poner en práctica a través de la Administración Pública en todos los niveles de la misma y con la colaboración en su desarrollo de los sectores públicos y los sectores privados. De manera explícita indica que las aplicaciones creadas deben de tener 'especial cuidado' con la discriminación de género, raza u otras. Además, contiene siete recomendaciones que requieren la participación de otros ámbitos o sectores y departamentos ministeriales, de manera interdisciplinar. De estas recomendaciones se desprende que las Administraciones que tengan la competencia para ello, han de elaborar una Estrategia Nacional para la IA, que ha de contemplar el mercado laboral, el modelo educativo, la legislación que esté en vigor, así como, las conexiones existentes en la sociedad con las nuevas aplicaciones o productos que se desarrollen más allá de la propia I+D+I.

Las Prioridades que vienen establecidas en la Estrategia (ENIA) son:

I. Conseguir una organización estructural para hacer posible la realización de un conjunto de reglas de I+D+I en Inteligencia Artificial y cuantificar su repercusión.

II. Decidir los campos relevantes en los que se han de centrar los recursos de las tareas de I+D+I.

III. Proporcionar la transmisión de los conocimientos y su devolución a la sociedad.

IV. Hacer un plan con las acciones formativas y de profesionalización en el área de la IA.

V. Llevar a cabo una estructura digital de datos y poner valor a las dotaciones existentes.

VI. Desde el lado de la I+D+I considerar la ética de la IA.

Mientras, las recomendaciones marcadas en la Estrategia son:

I. Promover la difusión de una Estrategia Nacional que lleve a cabo la instauración de medidas concretas en los campos relevantes de la IA y cuya valoración de dichas medidas sea llevada a cabo por un Observatorio Nacional creado para tal fin.

II. Usar la IA para lograr los objetivos de desarrollo sostenible establecidos por la Agenda 2030.

III. Plantear y poner en funcionamiento medidas concretas para la transmisión de los conocimientos a nivel social y económico.

IV. Fomento, recuperación y atracción del talento, así como, usar y añadir el conocimiento de la IA en el mercado laboral.

V. Inteligencia Artificial para usar los datos de las distintas Administraciones Públicas de manera óptima por medio de la creación de un Instituto Nacional de Datos.

VI. Introducir la IA en el sistema educativo español en los distintos niveles como medio de cambio a nivel tecnológico del país adaptando las necesidades y mejorando las competencias necesarias.

VII. Cuidar que todas las medidas y los resultados de las mismas en los diferentes campos de aplicación de la IA en la sociedad, hacen un uso responsable y ético, cumpliendo la legislación nacional y europea.

Además de la Estrategia Nacional en IA, el Gobierno de España, con el asesoramiento de un grupo asesor de expertos, elaboró en 2021 la llamada 'Carta de Derechos Digitales'. En ella se afianzan los derechos fundamentales de todos los ciudadanos en los nuevos escenarios que están dibujando las nuevas tecnologías. En concreto, respecto a la IA, dicha carta expone lo siguiente:

I. La inteligencia artificial deberá asegurar un enfoque centrado en la persona y su inalienable dignidad, perseguirá el bien común y asegurará cumplir con el principio de no maleficencia.

II. En el desarrollo y ciclo de vida de los sistemas de inteligencia artificial:

   a) Se deberá garantizar el derecho a la no discriminación cualquiera que fuera su origen, causa o naturaleza, en relación con las decisiones, uso de datos y procesos basados en inteligencia artificial.

   b) Se establecerán condiciones de transparencia, auditabilidad, explicabilidad, trazabilidad, supervisión humana y gobernanza. En todo caso, la información facilitada deberá ser accesible y comprensible.

   c) Deberán garantizarse la accesibilidad, usabilidad y fiabilidad.

III. Las personas tienen derecho a solicitar una supervisión e intervención humana y a impugnar las decisiones automatizadas tomadas por sistemas de inteligencia artificial que produzcan efectos en su esfera personal y patrimonial.

En definitiva, el impulso a la IA, tanto a nivel nacional como internacional, a través de la inversión masiva y medidas concretas hará que se produzca una implementación masiva de esta tecnología en prácticamente todos los sectores. La cuestión no es que se haga, si no que se haga de manera correcta, y sin repercusiones negativas tanto a nivel profesional, como personal. Siempre

teniendo los derechos fundamentales como guía. En este sentido, pasemos a analizar específicamente la ética en la Inteligencia Artificial y las regulaciones que se están fraguando en este ámbito.

VI. LA ÉTICA EN LA IA Y LA IMPORTANCIA DEL CONTEXTO CULTURAL EN LA TOMA DE DECISIONES

Son muchos los expertos, científicos y autoridades en la materia que abogan por una regulación en la Inteligencia Artificial. Nuevas leyes que "limiten" el uso y aplicaciones de esta revolucionaria tecnología. Pero es difícil hacerlo cuando el cambio es continuo y aún se desconocen muchas de las implicaciones reales. A cambio, se han creado multitud de códigos éticos, muestra de nuestro interés y preocupación. Sin embargo, no son de obligado cumplimiento, muestra de nuestra conciencia y reflexión. Si se incumplen, no ocurre nada. No hay castigo real. Es lo que algunos autores denominan un "lavado (de conciencias) ético" (Sebio, 2020). Y el tremendo potencial económico que promete la Inteligencia Artificial no es una baza a favor de la responsabilidad, el cuidado, la ética y la cautela, cuando tratamos con este tipo de tecnologías. De todas formas, cuando hablamos de ética, muchas veces damos por hecho, de manera incorrecta, que la ética es algo universal, mientras que es algo que puede depender de la cultura.

Es un hecho que no tiene la misma visión del mundo, ni las mismas perspectivas sobre las cosas, un japonés, un alemán o un español (Salazar; Gómez de Agreda, 2019). En este sentido, tener en cuenta la cultura, costumbres históricas, del lugar donde se ejecute el algoritmo, es muy importante si se pretende que esta tecnología gane en aceptación social. Uno de los experimentos más conocidos que demuestran, en sus conclusiones, este punto es el llamado *The Moral Machine*, desarrollado por el *Massachusetts Institute of Technology* (MIT).

En cualquier caso, hoy por hoy no parece plausible enseñar ética a una máquina. Por lo que serán a las empresas y sus desarrolladores a quienes hay que exigir que el desarrollo y el uso de la IA sean realizados de una manera ética, en concordancia con los derechos humanos internacionales. Y esto es el camino dominante que hoy en día están empujando los organismos internacionales. Trabajan en estos códigos éticos en Inteligencia Artificial para

dirigir las actuaciones de las empresas que implementan estas tecnologías en productos y servicios.

En este sentido, en abril de 2019, la Comisión Europea presentó las líneas maestras para el desarrollo de una IA fiable y segura en la Unión Europea, resaltando los aspectos principales a tener en cuenta para evitar posibles errores y consecuencias negativas. Son las siguientes (CE, 2019):

I. Supervisión humana: los sistemas de IA deben permitir sociedades equitativas y proteger los derechos fundamentales de los humanos. No deberán disminuir o limitar la autonomía humana.

II. Robustez y seguridad: la Inteligencia Artificial requiere que los algoritmos sean lo suficientemente seguros, confiables y sólidos como para enfrentar errores o inconsistencias durante sus procesos de toma de decisiones.

III. Privacidad y gobernanza de los datos: los ciudadanos deben de tener un control total sobre sus propios datos. Además, los datos que los conciernen no se utilizarán para perjudicarlos ni discriminarlos.

IV. Transparencia: debe garantizarse la transparencia en el proceso de toma de decisiones del sistema de Inteligencia Artificial (*Explainable AI*).

V. Diversidad, no discriminación y equidad: los sistemas de IA deben considerar toda la gama de habilidades y requisitos humanos, y garantizar la accesibilidad.

VI. Bienestar social y ambiental: los sistemas de Inteligencia Artificial deben utilizarse para mejorar el cambio social positivo y mejorar la sostenibilidad y la responsabilidad ecológica.

VII. Rendición de cuentas: deben establecerse mecanismos para garantizar la responsabilidad y la rendición de cuentas de los sistemas de IA y sus resultados.

Sin embargo, aunque la Comisión Europea es una referencia internacional con sus pautas éticas, muchas otras organizaciones, de distinta índole, han

publicado principios éticos para la Inteligencia Artificial. En marzo de 2019, la Universidad de Harvard estudió los principios de 32 organizaciones en todo el mundo que en ese momento habían publicado principios de IA (Fjeld; Nagy, 2020). Los analizaron en nueve dimensiones e hicieron una puntuación de cada organización para cada dimensión: derechos humanos, valores humanos, responsabilidad profesional, control humano de la tecnología, justicia y no discriminación, transparencia y explicabilidad, seguridad, rendición de cuentas y privacidad. Analizaron organizaciones de distintos tipos como organizaciones civiles, gobiernos, organizaciones intergubernamentales y compañías privadas.

También la ETH Universidad de Suiza examinó los principios de la IA de 84 organizaciones de todo el mundo y los clasificó con las siguientes dimensiones (Jobin *et al,* 2019): transparencia (usado por 73 organizaciones), justicia (68), no maleficencia (60), responsabilidad (60), privacidad (47), beneficencia (41), libertad y autonomía (34), confianza (28), sostenibilidad (14), dignidad (13) y solidaridad (6).

Además, la organización sin ánimo de lucro *Algorithm Watch man* tiene un inventario global (*Algorithm Watch*, 2018) de transparencia con 83 organizaciones que han publicado principios éticos de IA entre 2018 y 2019.

De estos estudios se deriva que la mayoría de los principios incluyen temas como transparencia, igualdad, no discriminación, rendición de cuentas y seguridad; algunos mencionan los derechos humanos y beneficencia. De su reflexión, se desprende que aún son pocas las compañías con estos importantes principios.

Por otro lado, en su Libro blanco de la Inteligencia Artificial publicado el 19 de febrero de 2020 (*European Comission*, 2020), la Comisión Europea identificó sectores y aplicaciones de alto riesgo, como, por ejemplo, salud o transporte, y aplicaciones con impacto significativo en las vidas de las personas (jurídico, vida o muerte, etc.). La regulación sería necesaria si se cumplen los dos criterios (sector y al mismo tiempo uso de alto riesgo). Propuso también definir dos listados exhaustivos (sector y uso) de alto riesgo, que se actualizarán periódicamente.

Con todas estas iniciativas de instituciones internacionales, organizaciones civiles, gobiernos nacionales e internacionales y empresas, se intenta regular

o autorregular el uso de la Inteligencia Artificial para que las máquinas tomen las decisiones dentro de un ámbito adecuado, que estas sean interpretables cuando impactan en la vida de las personas y que sus resultados no discriminen de una manera no deseada. A continuación se exponen, más concretamente, los principios éticos en IA.

VII. PRINCIPIOS ÉTICOS DE LA INTELIGENCIA ARTIFICIAL

Los principios éticos de la Inteligencia Artificial son una serie de criterios fundamentales que los miembros de una comunidad científica o profesional deben considerar para tomar decisiones en base a lo que se considera correcto e incorrecto sobre la utilización y el funcionamiento de esta tecnología en desarrollo (Salazar; Benjamins, 2021):

I. La Explicabilidad permite a las personas afectadas por el resultado de un sistema de IA acceder y entender el funcionamiento de esta tecnología con mayor facilidad. Esto implica proporcionarles información comprensible sobre los factores y lógica que condujeron a la obtención de los resultados y se puede lograr de diferentes maneras. Por ejemplo, según el contexto, los factores principales en la toma de decisiones, los factores determinantes, los datos, la lógica o el algoritmo detrás del resultado específico. Sin embargo, requerir de este principio también tiene desventajas ya que puede afectar negativamente la precisión, rendimiento, privacidad, costes, complejidad y seguridad de algunos sistemas de inteligencia artificial, porque implica la reducción de las variables de solución a un conjunto más pequeño para que los humanos puedan comprender y cuestionar los resultados respetando las obligaciones de protección de datos.

II. La Transparencia está sumamente relacionada con la explicabilidad y se basa en revelar cuándo se está utilizando la Inteligencia Artificial en diferentes acciones, como por ejemplo en las predicciones, toma de decisiones, recomendaciones, *bot* conversacionales, etc. Posibilita que las personas entiendan cómo se desarrolla, opera e implementa un sistema de inteligencia artificial en el dominio de una determinada aplicación, con capacidad de explicar por qué y qué información relevante proporciona

esta tecnología en la actualidad. Además, permite que los usuarios puedan tomar decisiones más informadas, incrementando la aceptación y confianza de los resultados obtenidos.

III. El ser humano como centro: la Inteligencia Artificial debe desarrollarse de acuerdo con los valores centrados en el ser humano, como la igualdad, libertad, equidad de estado de derecho, justicia social, protección de datos y privacidad, derechos del consumidor, equidad comercial, etc. Ya que algunas aplicaciones o usos de los sistemas de IA pueden implicar riesgos de que los derechos humanos y valores centrados en el ser humano puedan ser infringidos de forma intencionada o accidental. Se debe promover la alineación de valores en el diseño de esta tecnología junto a la intervención y supervisión humana de su funcionamiento, para garantizar que los comportamientos de los sistemas de inteligencia artificial protejan y prioricen al ser humano frente a los avances de esta tecnología. Todo eso fortalecerá la confianza de las personas y reducirá la discriminación u otros resultados sesgados.

IV. La toma de decisiones es un principio fundamental que se rige por implicaciones legales como la ética y moral para influenciar las acciones de los algoritmos programados por los seres humanos. Hoy en día, la Unión Europea exige a las entidades proporcionar explicación a las decisiones que fueron tomadas a través de la Inteligencia Artificial, ya que estás en cierta medida generan impacto e influyen en la sociedad. La máquina no asume ninguna responsabilidad porque esta no tiene derechos ni principios éticos. Por lo tanto, las personas son responsables de aplicar los marcos normativos en el desarrollo y diseño de los sistemas de esta tecnología para garantizar un correcto funcionamiento de la IA y así garantizar las razones por las que fueron tomadas determinadas decisiones.

V. No maleficencia en el uso de Inteligencia Artificial. Los sistemas de IA tienen la obligación de respetar y proteger a las personas a través de los derechos digitales. Estos consisten en el derecho a la privacidad en entornos digitales, protección a la integridad personal, *grooming*, ciberseguridad, derecho a la propia imagen, acceso a información veraz, etc. Es fundamental que los humanos tengan consentimiento sobre la utilización

correcta de los datos y capacidad de tomar decisiones. Es decir, tener conocimiento de la causa y finalidad de los resultados obtenidos por los sistemas de Inteligencia Artificial.

VI. Sesgos y discriminación. Evitar que los algoritmos tengan problemas de discriminación y buscar la manera de que los resultados obtenidos sean explicables. Los sistemas de Inteligencia Artificial nos ayudan en la selección de información y en la elaboración de nuevos artículos a partir del procesamiento de los datos, sin embargo, si hay sesgo del programador de la IA se pueden generar problemas de desconfianza en los resultados obtenidos. Es fundamental generar conciencia sobre el impacto que tienen las personas que trabajan con algoritmos para así evitar, detectar y mitigar sesgos de discriminación de género, raciales, etc., porque estos generan injusticias, prejuicios y concepciones de la realidad con capacidad de influir en la toma de decisiones de los seres humanos.

A pesar de estos principios éticos en IA compartidos por la mayor parte de usuarios y desarrolladores a nivel nacional e internacional, como se ha visto a lo largo de este capítulo, no son de obligado cumplimiento. Los casos de uso que ya han tenido lugar pronostican la clara necesidad de una regulación de obligado cumplimiento que ayude a impulsar la IA pero manteniendo la protección sobre los derechos humanos.

VIII. NORMATIVA MÁS ALLÁ DE LA ÉTICA

La Unión Europea lleva años trabajando en este ámbito y ha impulsado diferentes propuestas; en 2017 el Parlamento Europeo recomendó a la Comisión la regulación en materia civil para la tecnología robótica. En 2018 la Comisión publicó su primer comunicado sobre IA. En 2019 un grupo de expertos de la Comisión elaboró una guía para el desarrollo de una IA confiable. Y el 21 de abril de 2021 se ha presentado una propuesta de legislación.

La propuesta legislativa ha sido elaborada por la Comisión de la Unión Europea en respuesta a una proposición del Parlamento Europeo. El cuerpo de la ley ofrece límites para los casos más controvertidos en la aplicación de inteligencia artificial y centra su objetivo en conseguir la seguridad de los ciudadanos de los países miembros de la UE.

La propuesta sitúa el centro del interés en el desarrollo de una tecnología ética que garantice la seguridad y respete los derechos fundamentales. Para alcanzar este objetivo, la regulación diferencia cuatro niveles de riesgo:

— *Riesgo inaceptable:* los usos de IA que se encuentren en este nivel de la jerarquía quedarán prohibidos. Se refiere a las aplicaciones que suponen una amenaza para la seguridad, la vida o los derechos fundamentales. Un ejemplo de este nivel son los sistemas que incitan a la violencia o al odio, y los sistemas de puntuación social que podrían utilizar los Gobiernos para diferenciar y clasificar a los ciudadanos.

— *Riesgo alto:* el segundo nivel se reserva, principalmente, para el ámbito de la salud y de la educación. Aunque también incluye servicios públicos y privados, control de fronteras y administración de justicia. Las aplicaciones que quieran desarrollar prestaciones en estos campos deberán contar de forma obligatoria con: un estudio de evaluación y control de riesgos, una calidad alta en las bases de datos que alimentan el sistema, un registro de la actividad para mostrar en qué se basan los resultados obtenidos, una oferta de información clara y detallada a los usuarios y una tecnología sólida, segura y precisa.

En este nivel se encuentran los sistemas biométricos de reconocimiento facial, que quedan prohibidos en espacios públicos con las siguientes excepciones: desaparición de menores, alerta de ataque terrorista, localización de autores de delitos graves. Las excepciones deben contar, para poder implementarse, con una autorización judicial en la que además se delimite el tiempo y la zona geográfica en la que se aplicará.

— *Riesgo limitado:* en este nivel se encuentran los *chatbots*. La regulación establece que los sitios web y las aplicaciones que utilicen estos servicios de comunicación con el cliente deberán notificárselo. En los artículos dedicados a este nivel se hace hincapié en la transparencia, que además da lugar al nombre del título IV dónde se encuentran.

— *Riesgo mínimo:* para el nivel más bajo de la jerarquía no se contempla ninguna medida reguladora en la propuesta de ley. Esto es porque la Comisión no considera que las herramientas que pertenecen a este nivel supongan un riesgo

o potencial amenaza para la sociedad. Entre ellas se encuentra la aplicación de la IA a los videojuegos o los filtros para ordenar el correo electrónico.

Antes de la redacción de normas para inteligencia artificial y ante la posibilidad de su existencia, algunos profesionales del sector temen que la regulación frene el desarrollo de la tecnología. Para evitarlo, la UE está basando principalmente la regulación en base al riesgo. No regula la tecnología sino el uso que se hace de ella.

La UE espera conseguir con esta regulación impulsar la IA, basando este impulso en la confianza que generará la regulación, al ofrecer una protección adicional a los derechos fundamentales.

IX. MEDIDAS ADICIONALES DE CONTROL

Todas estas medidas de éticas y regulación en IA son necesarias, pero aún es necesario algo más que ayude a complementarlas. Entre ellas, podemos encontrar las siguientes:

I. Sello/Certificación. Revisión y análisis de los sistemas de IA empleados y desarrollados por empresas y organizaciones con el objetivo de ofrecer confianza a consumidores y usuarios.

II. Guías prácticas y realistas en el uso responsable de la IA (OdiseIA, 2021): su objetivo es establecer un ecosistema colaborativo donde empresas privadas de todos los sectores y sus homólogos organismos públicos compartan de manera aterrizada buenas prácticas en el uso ético de la IA. *Framework* con ejemplos concretos que integren de manera aterrizada una visión tecnológica, de gobierno, de negocio (particular para cada sector), ética y regulatoria en los sistemas de IA.

III. Cambio de política educativa:

- Integración del impacto de la IA en todos los grados académicos, y Educación secundaria (más allá de la técnica).

- Formación a cualquier grupo de edad y disciplina en el impacto de la IA.

- Fomento de la comprensión del uso de los datos privados.

- Talleres educativos, adaptados a cada necesidad, para discutir casos de impacto.

- Medidas para el refuerzo del espíritu crítico de la sociedad.

IV. Campañas de concienciación:

- Los casos en los que se prevea impacto, tanto públicos como privados deben estar precedidos de una campaña previa, mediática, de concienciación ciudadana con el objetivo de:

- Fomentar la transparencia de esa empresa u organización público/privada.

- Aclarar posibles dudas/sugerencias.

- Aumentar la confianza social.

X. BIBLIOGRAFÍA

– BRIEGA, L. R. E. (2017)*: Introducción a la Inteligencia Artificial. Matemáticas, Análisis de Datos y Python.* https://relopezbriega.github.io/blog/2017/06/05/introduccion-a-la-inteligencia-artificial/

– Carta de Derechos Digitales (2021): https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf

– CDE. (2020): *Inteligencia Artificial: amenazas y oportunidades.* Centro de Documentación Europea de la Universidad de Almería. https://www.cde.ual.es/inteligencia-artificial-amenazas-y-oportunidades/

– DARLINGTON, K. (2017): *Sistemas de IA explicables: comprender las decisiones de las máquinas.* OpenMind BBVA.

- elEconomista.es (2019): *El sector de la inteligencia artificial moverá 52.700 millones en 2025.* El Economista. https://www.eleconomista.es/gestion-empresarial/noticias/10231424/12/19/El-sector-de-la-inteligencia-artificial-movera-52700-millones-en-2025.html

- EUROPEAN COMISSION: *Regulation of the European Parliament and of the Council, Artificial Intelligence act and amending certain Union legislative acts.* (21.4.2021). Bruselas, p. 79.

- *Estrategia Nacional de Inteligencia Artificial* (2020): https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/ENIAResumen2B.pdf

- European Comission Regulation*: Excellent and trust in artificial intelligence* (2021): https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_en.

- FJELD, J.; NAGY, A. (2020): *Principled Artificial Intelligence: MAPPING CONSENSUS IN ETHICAL AND RIGHTS-BASED APPROACHES TO PRINCIPLES FOR AI.* https://cyber.harvard.edu/publication/2020/principled-ai

- GÓMEZ-DE-ÁGREDA, Á.; FEIJÓO, C.; SALAZAR-GARCÍA, I. (2021): "Una nueva taxonomía del uso de la imagen en la conformación interesada del relato digital. Deep fakes e inteligencia artificial". *Profesional de la información,* v. 30, n. 2, e300216.

- IEEE (2020): "Ehically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems". https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

- ODISEIA (2021): *Guía de buenas prácticas para un uso responsable de la IA:* https://www.pwc.es/es/publicaciones/tecnologia/assets/guia-buenas-practicas-uso-inteligencia-artificial-pwc-odiseia.pdf

- RAPOSO, S. (2020): *En un accidente con coche autónomo, ¿quién tiene la culpa? ¿O no hay culpables?.* Computer Hoy. https://computerhoy.com/noticias/motor/accidente-coche-autonomo-quien-tiene-culpa-no-hay-culpables-777005.

– SEBIO, M. (2020): *Inteligencia Artificial y ética.* Universidad Pontificia de Comillas. https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/37008/IA%20y%20etica_Sebio%20Martin,%20Margarita.pdf?sequence=-1

– O´NEIL, C. (2016): *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Random House LCC US

– Parlamento Europeo. (2020): *Inteligencia artificial: oportunidades y desafíos.* Parlamento Europeo. https://www.europarl.europa.eu/news/es/headlines/society/20200918STO87404/inteligencia-artificial-oportunidades-y-desafios

– Propuesta de Reglamento del Parlamento Europeo y del Consejo (2022): https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

– SALAZAR, I.; BENJAMINS, R. (2020): *Towards a framework for understanding societal and ethical implications of Artificial Intelligence.* Cornell University. arXiv:2001.09750v1.

- (2020): *El mito del algoritmo: cuentos y cuentas de la Inteligencia Artificial* (ANAYA) ISBN-10: 8441542805.

– SALAZAR, I. (2019): *La Revolución de los Robots. Cómo la Inteligencia Artificial y la robótica afectan a nuestro futuro.* (TREA) ISSN: 1 978-8417767-34-1

- (2018): "Los robots y la Inteligencia Artificial: nuevos retos del Periodismo." *Doxa Comunicación.* Vol: 27. ISSN: 1696-019X.

- (2019): *La Revolución de los Robots.* Ed. Trea.

– SALAZAR, I.; GÓMEZ DE AGREDA, Á. (2019): "Sesgos y perspectiva cultural en el entrenamiento de los algoritmos de inteligencia artificial". *Revista Privacidad y Derecho Digital.* Nº 15. ISSN: 2444-5762.

– TURING, A. M.: *Computing machinery and intelligence,* Mind, Volume LIX, Issue 236, 1 October 1950, Pages 433–460, https://doi.org/10.1093/mind/LIX.236.433.

– UNESCO (2021): "Recommendation on the Ethics of Artificial Intelligence". https://unesdoc.unesco.org/ark:/48223/pf0000381137

– WYLIE, C. (2020): *Mindf*ck: Cambridge Analytica. La trama para desestabilizar el mundo.* Roca Editorial.