UNIVERSIDAD D SEVILLA

Facultad de Filosofía

Grado en Filosofía

Impactos y Desafíos Éticos de la Inteligencia Artificial y el Big Data

Trabajo Fin de Grado presentado por Álvaro Ortega de la Rosa y tutorizado por el profesor Dr. José Antonio Marín Casanova

Sevilla, junio de 2023

RESUMEN:

La Inteligencia Artificial y el Big Data ofrecen numerosas ventajas para la sociedad, pero como toda herramienta también llevan parejos algunos riegos éticos. Este trabajo se enfoca en primer lugar en examinar el marco problemático del uso de estas tecnologías. Se analizan algunos de los impactos éticos debidos a fallos de diseño inherentes a este tipo de algoritmos, tanto por su propia arquitectura como por la gestión de datos que requieren. Se reflexiona sobre los riesgos éticos que pueden surgir por un uso malintencionado, cuando estas tecnologías digitales se utilicen como una nueva forma de poder por parte de gobiernos o grandes empresas, al margen de los valores democráticos. También se identifican algunos dilemas éticos que surgen con los nuevos usos actuales y potenciales. Por último, se considera la amenaza existencial que supondría la invención de una Inteligencia Artificial General, dada la enorme capacidad que tendría para intervenir perjudicialmente en una sociedad intensamente tecnificada como la actual, así como algunos de los problemas éticos que podrían surgir si finalmente acabara integrada en la sociedad.

Palabras clave: Inteligencia Artificial, Big Data, riesgos, ética, amenaza existencial.

ABSTRACT:

Artificial Intelligence and Big Data offer numerous advantages for society, but like any other tool, they also come with some ethical risks. This work focuses on examining the problematic framework of using these technologies. It analyzes some of the ethical impacts resulting from inherent design flaws in these types of algorithms, due to their own architecture and the data management they require. Reflection is made on the ethical risks emerging from malicious use, when these digital technologies are employed as a new form of power by governments or large corporations, disregarding democratic values. It also identifies some ethical dilemmas that arise from current and potential new uses. Finally, it is considered the existential threat posed by the invention of a General Artificial Intelligence, given its enormous capacity to intervene detrimentally in an intensely technified society like the present one, along with some of the ethical problems that may arise if it were eventually integrated into society.

Keywords: Artificial Intelligence, Big Data, risks, ethics, existential threat.

ÍNDICE

1. INTRODUCCIÓN

1.1 Objetivos

El objetivo de este trabajo es reflexionar sobre los nuevos desafíos éticos que surgen conforme se desarrollan las nuevas tecnologías y se implantan en el seno de la sociedad, analizándose en particular las contribuciones del Big Data (BD) y la Inteligencia Artificial (IA). Estos dos adelantos tecnológicos se pueden utilizar en múltiples campos y su aplicación conllevará beneficios para la comunidad, aunque no se debe olvidar que en general su uso va enfocado a conseguir réditos económicos para sus desarrolladores y clientes.

Este trabajo enfocará el esfuerzo en identificar los impactos éticos que surgen por los fallos de diseño y vulnerabilidades que ineludiblemente llevan aparejadas estas tecnologías y que pueden causar un perjuicio a la sociedad o a los individuos. En segundo lugar, se analizarán los riesgos éticos provocados por su uso malicioso para controlar, monitorizar y manipular a la población, disminuyendo la libertad de los ciudadanos. Igualmente se discutirán los nuevos dilemas éticos que aparecen conforme aumentan las aplicaciones de estas tecnologías y cómo pueden afectar a las capacidades del ser humano. Por último, se considerará el caso de la invención de una superinteligencia artificial y la amenaza existencial que ello supondría sobre el género humano, así como los impactos éticos en las interacciones humano-máquina, en el caso de que este nuevo ente artificial se integrara en la sociedad.

1.2 Metodología empleada

El trabajo se ha realizado utilizando un enfoque analítico de las fuentes bibliográficas consultadas, para así poder estudiar los datos más relevantes y discutir las implicaciones que de ellas se pueden extraer.

La bibliografía se ha obtenido a través de motores de búsqueda en Internet y de plataformas de artículos académicos como *Google Scholar*. Por ser un tema de máxima actualidad, también se han utilizado fuentes como artículos periodísticos y otros contenidos divulgativos, para completar el enfoque con ejemplos reales que ya están sucediendo.

1.3 Relevancia de la Inteligencia Artificial y el Big Data en la actualidad

Es un hecho que en los últimos años las tecnologías digitales se han enfocado en aumentar la conectividad de los usuarios, con el objetivo inicial de permitir un uso global, más intensivo y personalizado de la digitalización. Esto se traduce en que se han desarrollado múltiples aplicaciones para facilitar la recopilación, almacenamiento y utilización de datos personales de los usuarios, principalmente a través de lo que se ha venido en llamar el BD. Con este término se denomina la capacidad de procesamiento y análisis de grandes cantidades de datos que superan enormemente la tradicional capacidad de gestión y análisis de las herramientas digitales del siglo XX. Estos inmensos conjuntos de datos pueden incluir tanto información estructurada, que sería la que se encuentran de forma ordenada en bases de datos, como no estructurada (por ejemplo, textos, vídeos, imágenes, etc.), sobre los que se buscan correlaciones. Por eso, lo que caracteriza al BD es la gran cantidad de datos que maneja, la velocidad a la que es capaz de analizarlos, la variedad de tipos de datos que acepta y la capacidad de entresacar información valiosa de entre el amasijo de datos.

En el BD se utilizan técnicas avanzadas de procesamiento de datos, como la minería de datos, el aprendizaje automático, la IA, el análisis automatizado de texto e imágenes, etc., con los que se pueden identificar patrones, correlaciones y tendencias en los datos, que de otra forma pasarían desapercibidos. Es en esta capacidad de encontrar correlaciones, que después servirán para obtener predicciones, donde radica especialmente su valor para las empresas y las organizaciones que lo usan, en campos como la publicidad, el marketing, el análisis financiero, la salud, etc., incrementándose paulatinamente el número de sectores donde resulta de interés esta técnica.

En cuanto a la IA, el término se refiere a la capacidad de las máquinas para realizar tareas abstractas y que normalmente requerirían una inteligencia humana, como por ejemplo la percepción, el aprendizaje, el razonamiento, la toma de decisiones o la planificación de estrategias. La IA está basada en algoritmos y modelos matemáticos que permiten a las computadoras analizar datos y aprender de ellos. Conforme más interacciones realizan la IA con nuevos datos, mejor es su rendimiento, pues se retroalimenta del análisis de sus experiencias anteriores y consigue así acelerar su aprendizaje.

La IA se suele dividir en varios campos, como son el aprendizaje automático (*Machine Learning*), el procesamiento de lenguaje natural (*Natural Language Processing*), la robótica o la visión por computadora (*Computer Vision*), entre otros. En cada una de estas áreas se utilizan enfoques y técnicas específicos, por lo que el desarrollo técnico se está logrando independientemente en cada rama. Mientras que el aprendizaje automático se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender por sí mismas a partir de datos y experiencias anteriores, el procesamiento de lenguaje natural se enfoca en que la maquina entienda y se comunique usando un lenguaje humano. Su ejemplo más actual es la aplicación *ChatGPT*, la cual puede procesar cualquier lenguaje natural, conversar con los usuarios y generar respuestas razonables teniendo en cuenta el contexto de la conversación. Por su parte, la visión por computadora permite que las maquinas entiendan y procesen imágenes fijas y vídeos, lo que les abre a percibir el mundo exterior de una manera análoga a como lo captamos los humanos, aunque evidentemente con procesamientos distintos.

Se debe subrayar la relación recíproca entre la IA y el BD. El entrenamiento de la IA está supeditado a los progresos en BD, porque consigue realizar la minería de datos y desentrañar información útil que alimenta a la IA. A su vez, el BD va incorporando los avances en IA para realizar de manera más eficiente su tarea, reconociendo y gestionando de una manera más rápida mayores fuentes de datos, como pueden ser imágenes y videograbaciones.

En la actualidad, la IA se está utilizando en cada vez más aplicaciones, desde la diagnosis médica y la atención al cliente, hasta el análisis financiero y la optimización y automatización de procesos, pasando por la conducción autónoma y la seguridad cibernética. Su éxito está impulsando su expansión hacia formas más eficientes y sofisticadas de IA, acelerándo su evolución por los avances en las tecnologías de aprendizaje profundo (*Deep Learning*). Todo esto está cambiando radicalmente la forma como las empresas y organizaciones realizan sus actividades y toman decisiones, ya que, gracias a la IA y su buena fiabilidad, se aumentan las capacidades y se reducen costes, lo que genera una ventaja competitiva frente a las compañías que no incorporen estas innovaciones.

Uno de los ejemplos más difundidos y actuales de esta tecnología son los *bots*, que no son más que programas que realizan tareas repetitivas, predefinidas y automatizadas con la intención de imitar o sustituir la acción humana, pero con la ventaja de realizarla de una

manera mucho más rápida, menos costosa y con interacciones cada vez más realistas. Así, hoy en día nos vemos rodeados de *bots* que se encargan de la atención telefónica al cliente, o los que están instalados en los motores de búsqueda (recopilando información y sugiriéndonos qué queremos rastrear), en las redes sociales (pueden hacerse pasar por usuarios humanos), en el correo electrónico (enviando *spam* o *e-mails* maliciosos), en los mercados financieros (haciendo compraventas de activos) o mientras navegamos, de manera que somos monitorizados por *bots* que analizan nuestra actividad en la red y recopilan nuestro comportamiento y preferencias.

Los *bots* pueden parecer inocentes programas sin mayores riesgos (véase Siri, Alexa, Cortana o cualquier contestador de atención al cliente de una empresa), y pensar que solamente son un desarrollo digital más, pero su uso y expansión revelan lo contrario. Según las estadísticas, en 2021 los *bots* supusieron el 42% del tráfico mundial de datos en Internet, casi igualando el uso humano, y cifrándose en un 27% el volumen de Internet copado por el intercambio de información de *bots* maliciosos (Imperva, 2022). A esto se debe añadir que están alcanzando un grado de mimesis con lo humano que les permite dar respuestas a veces indistinguibles de las que recibiríamos del trato con otra persona, por lo que superan sin dificultad el test de Turing. Sumémosle las múltiples capacidades en las que nos aventajan por su exactitud o rapidez, y se podrá vislumbrar el impacto futuro en nuestra vida diaria conforme sigan expandiéndose y ganando influencia.

Si estos programas, que son la primera aproximación de la IA aplicada de forma masiva a la interacción con los humanos, ya nos provocan problemas y desasosiegos, tenemos entonces la obligación de reflexionar sobre el tremendo impacto que nos sobrevendrá en las próximas décadas, conforme estas tecnologías se desarrollen y generalicen aún más. Por ello, aunque la IA y el BD tienen el potencial de mejorar muchos aspectos de nuestras vidas y de cambiar nuestra civilización por el impulso que dan a la globalización y a la realización de tareas a distancia o automatizadas, también plantean importantes preguntas éticas sobre sus consecuencias en la sociedad y sobre el grado de supervisión que se debería ejercer sobre ellas.

En los países desarrollados, la tendencia general es vivir en urbes hipertecnificadas, lo que ha venido en llamarse las *Smart Cities*, con un nivel de conexiones digitales nunca visto. De hecho, ya se empieza a denominar a la tendencia actual la "cuarta revolución industrial",

marcada por el uso intensivo de tecnologías digitales, que fusionan los tradicionales ámbitos biológicos y físicos con la esfera digital. Hoy ya tenemos ejemplos tangibles como la realidad virtual y la aumentada, el ciberespacio, la robótica o el BD, y es de esperar que aumenten conforme se desarrolle la IA.

A este escenario superdigitalizado se le debe sumar el problema de la brecha del conocimiento digital, que cada vez se agranda más. Por un lado, se encuentra la población en general, que sabe usar aplicaciones digitales con mayor o menor soltura, pero que desconoce la técnica digital y el detalle de sus capacidades. Esta mayoría de la sociedad suele ir a remolque de las innovaciones digitales o incluso no llegan a incorporarlas, como les pasa a las generaciones más envejecidas. Por otro lado, están las personas con soltura en las tecnologías digitales y experiencia en el uso, extracción y aprovechamiento de los datos. Esto genera lo que se ha llamado metafóricamente el *feudalismo de los datos* (Cobo, 2019): Una inmensa población genera datos sin recibir contraprestaciones por ello (los vasallos de datos), y estos son administrados y aprovechados por una ínfima minoría (los escribanos de datos). Esta élite ha conseguido una posición de privilegio tanto por el rendimiento económico que perciben por comerciar con la información, como por la capacidad de predecir comportamientos y de poseer semejantes bancos de datos, lo que provoca la aparición de nuevas formas de poder y control concentradas en las manos anónimas de unos pocos.

2. MARCO PROBLEMÁTICO DE LA INTELIGENCIA ARTIFICIAL Y EL BIG DATA

Reflexionar sobre los impactos éticos de estas tecnologías digitales en la sociedad requiere en primer lugar plantear cual es el marco problemático donde se dan. La sociedad actual lo fía casi todo a la tecnología, tanto en el presente, como en sus aspiraciones futuras. Las ventajas y comodidades que brinda la tecnología sustentan la civilización actual y alimentan el apetito por conseguir más. Sus éxitos provocan que se confié en ella para que halle la solución para todo en el mañana: la amenaza del cambio climático, la obtención de recursos, la cura de enfermedades, etc. Es una sociedad dominada por la tecnología, donde se acepta

que ésta le aporte a la Humanidad con precisión matemática las certezas que nunca tuvo, y que todo quede calculado y planificado, para acercarnos al mundo perfecto.

Pero todo esto tiene sus interacciones. El tipo y nivel de tecnología alcanzados no solamente modifican la sociedad por lo que los propios avances técnicos proporcionan *per se*, sino que contribuyen a redefinir su organización interna, a evaluar las creencias que se aceptan como válidas e incluso tienen repercusión en su sistema de valores. Eficiencia, exactitud, productividad o aplicabilidad pasan a ser virtudes, mientras que la ignorancia, la ambigüedad, lo limitado se valoran como defectos y no como características.

La tecnología también afecta al enfoque de la realidad que se tiene y a cómo se percibe el entorno, moldeando la cosmovisión. Por eso, la función salvífica que actualmente se otorga a la tecnología incita a profesar en ella una profunda fe, aceptando simultáneamente cualquier cosa que exija a cambio, como el nuevo Dios que otorga el dominio del mundo y al mismo tiempo reclama su tributo. Una de las cosas que demanda es aceptar que su visión racional y positivista de la realidad es la única verdadera, por ser el ordenamiento que proporciona mayor seguridad y nos libera de la impredecibilidad del devenir. Los datos y su análisis son ahora la nueva forma de conocimiento, aunque esta abundancia informativa oculte la falta de sentido general que la tecnología implica y una ausencia de meditación sobre formas alternativas de significado.

A cambio de los adelantos digitales, no se ha dudado en permitir su presencia en todos los lugares y en todo momento, para medir cualquier aspecto de la actividad vital humana que requieran. La tecnología digital se ha propuesto registrar y calcular cada conducta gracias al BD, comenzando por la Red y ampliándose a todos los ámbitos del entorno. Se ceden aspectos de la intimidad a cambio de ventajas y milagros tecnológicos que la propia técnica nos induce a aceptar, acabando con la supuesta neutralidad de la tecnología. Esa extracción de mediciones es la nueva sangre que da impulso a las tecnologías digitales, para ofrecer un mundo alternativo calculado con datos, mejorado y controlable. Obligatoriamente, esta situación implica que se acepta estar regido por la racionalidad cuantificadora e incita a considerar como válido únicamente el mundo medible y racionalizado.

En este camino progresivo hacia una mayor presencia de la tecnología, va cambiando simultáneamente cómo el ser humano se ve a sí mismo, avergonzado por lo poco que puede

hacer y lo mucho que se equivoca y se desvía, frente a la fría perfección y eficiencia de su obra. Más aun cuando la perspectiva que se ha impuesto es la técnico-científica, que todo lo vuelve objeto de estudio y donde la ensoñación del Hombre no puede competir. Por eso los desarrollos tecnológicos conllevan inevitablemente una pérdida de valor del ser humano, que pasa a ser producto y objeto de su propia obra tecnológica, cosificándose según se desarrolla más y más el mundo tecnológico que creó. Igualmente se generan dudas sobre qué papel le queda al Hombre como agente de su propia vida y sobre su lugar en el mundo, que de tanto alterarlo empieza a no ser la cómoda y bella morada que siempre quiso.

Aunque los grandes beneficios de la técnica se han ido repartiendo lo suficiente para hacer que todos se sientan agradecidos, se ha logrado a costa de la explotación planificada y sistemática del planeta, la perpetuación de las desigualdades y el aumento de las asimetrías de poder. El Hombre soñaba que la técnica lo liberaría del trabajo, pero ahora comprueba que únicamente lo sustituye. Donde él era el único sujeto con capacidad de raciocinio y autonomía, ahora se ve desplazado como elemento central del mundo que ha creado. Ya no tiene el monopolio de realizar acciones útiles y encuentra que el valor de sus capacidades mengua, pues la técnica permite hacerlo todo más rápido y eficiente. Esto le obliga a malvender su tiempo y su esfuerzo en aquellos trabajos que no hayan sido colonizados por la tecnología y le imprime una angustia que nunca antes sufrió, por ver cómo sobrevivir en su mundo artificial.

Esta atmósfera hipertecnificada implica asimismo renunciar a pensar sobre el sentido, pues no es necesario si ya la tecnología proporciona el mejor significado posible, y además se ofrece a elegir por nosotros. Se está transitando a un escenario donde la teoría prevalece, frente a la práctica y al razonamiento *a posteriori* con el que el ser humano ha ido ampliando históricamente su conocimiento. La exactitud de la técnica acaba con la problemática arbitrariedad humana, pero implica ser dirigido por la eficiencia algorítmica, a la que se le concede el valor de incontestable y verdadera. Tal pérdida tanto del sentido como del interés por buscarlo hace que el Hombre contemporáneo fíe toda su vida a la técnica y se limite a ser solo un consumidor de sus beneficios y un emisor de datos, una parte más dentro del ciclo de producción, donde prima lo material y la utilidad que se le pueda extraer.

Por otra parte, la omnipotencia de la técnica le acaba empujando al nihilismo, entendido como la falta de valores en el mundo. Si los valores tradicionales han sido desenmascarados

y la técnica demuestra que todo lo puede, la humanidad se enfrenta a un universo de posibilidades, pero sin reglas, categorías o valores absolutos que le guíen en él. La técnica permite subrogar la gestación, clonar individuos, alterar el propio cuerpo, por ejemplo, y ante esto los antiguos valores ya no sirven, dejando a la ética desactualizada. El rápido desarrollo tecnológico impele a destruir los antiguos tabúes y cuestionar todos los valores, ya que permite superar las restricciones que antaño formaron el límite de la voluntad.

Ahora es la IA la que está efectuando su entrada, un ser pensante capaz de dar soluciones para todo, aunque sus respuestas no sean del todo neutrales, sino basadas en el sentido e intereses de quienes diseñaron sus algoritmos. Durante toda su Historia, el Hombre se ha encontrado a sí mismo como el único ser dotado de razón, pero ahora le acompañará una nueva compañera y rival, una Razón Artificial basada en puras relaciones lógicas y matemáticas, que cuando se implante redefinirá sustantivamente los procesos organizativos de la sociedad y sus relaciones con el entorno.

Habrá que acomodarla en la sociedad, tanto en las costumbres como en los aspectos éticos, y esto exigirá repensar los modelos bajo los que se concibe la vida en comunidad. Su influencia será grande, pues sus respuestas precisas y discretas acabarán por decidir no solo los detalles, sino cómo serán nuestras relaciones generales con todo y el sentido que se le presten. Si ahora se quiere construir unas IA que se adapten al Hombre, conforme progresen y se multipliquen, acabará siendo el ser humano el que se adapte a ellas. Errar, saltarse las normas, el desorden o la imperfección son características netamente humanas y constitutivas de nuestra libertad. Si la IA elimina estos aspectos o se dejan atrás gracias a su colaboración, también se reducirá la libertad de la que el ser humano siempre ha gozado. Sin olvidar que este nuevo actor exige que la riqueza ontológica del mundo que desde siempre ha fascinado al Hombre deberá simplificarse, para que pueda acomodarse al modelo de lo que la IA puede comprender, una realidad medible, reducida a datos y relaciones lógicas. También se acabará por cederle el dominio de la técnica y la supremacía del raciocinio, con lo que aumentará el hiato entre el Hombre y su obra.

Si ya hoy en día los ciudadanos ceden a la sociedad parte del respeto inherente que se le debiera a todo individuo por el propio estatus moral que detenta (no tanto en aras de la convivencia sino por la presión que ejercen los poderes públicos o económicos), no cabe duda de que más pronto que tarde se harán concesiones similares hacia la tecnología. Es

probable que se acabe con menos autonomía y más desguarnecidos frente a las amenazas, directas o como efectos colaterales, de la tecnología. Por eso todos estos cambios actuales y venideros exigen volver a repensar y clarificar los principios y normas sobre los que asentar la futura relación con la tecnología, para encontrar un punto de equilibrio entre seguridad, respeto al estatus moral del ser humano y las nuevas capacidades que la técnica trae consigo.

3. DESAFÍOS ÉTICOS DEL BIG DATA Y LA INTELIGENCIA ARTIFICIAL DEBIDOS A FALLOS DE DISEÑO

Entre los desafíos éticos que estas tecnologías presentan, se encuentran los derivados de los fallos de diseño inherentes a la propia naturaleza de estas innovaciones digitales. Se consideran bajo este apartado algunos de los problemas que probablemente no puedan eliminarse con mejoras futuras y que tampoco es posible depurar totalmente para que no afecten a la calidad de los resultados obtenidos, pues surgen del propio tipo de diseño que tienen.

3.1 Sesgo

La aparición de sesgo es actualmente uno de los principales problemas y de los más comunes, pues remite a la calidad de los datos de los que se sirven estas tecnologías. Si la información que se ha utilizado para entrenar los modelos de análisis tiene algún sesgo, los resultados que produzcan también los tendrán. Es lo que se conoce en la jerga informática como GIGO[1], pues datos con falsedades dan resultados con errores, estando la calidad de lo producido por el sistema íntimamente ligada a la veracidad de los datos introducidos.

A esto debe añadirse el problema de que la propia tarea de analizar y enjuiciar los datos se produce a través de un determinado criterio del programador, por ejemplo, eligiendo qué categorías de datos son más relevantes y deben sobreponderarse y cuáles son menos importantes y deben tener menor peso u obviarse. Es decir, siempre hay unas suposiciones

[1] Garbage In, Garbage Out

culturales y una lógica subyacente enmarcada en la herencia histórica y cultural del equipo que realiza la programación y en su manera de evaluar los datos del análisis. Por eso, el resultado es que se decide ignorar ciertos factores o se da más importancia a otros por una simple inercia cultural, de forma que lo que estas tecnologías hacen es codificar el pasado.

El riesgo ético inherente a este aspecto consiste en que tales desviaciones pueden acarrear que los resultados generados, y por tanto las decisiones que se tomen en base a ellos, sean inexactos, injustos o discriminatorios para algunas personas o colectivos, porque la IA estaría reproduciendo los prejuicios del programador o el sesgo de los propios datos. No se debe olvidar que los datos también están entrelazados íntimamente con la sociedad que los ha generado (Hagerty & Rubinov, 2019), porque cada comunidad humana decide qué es licito compartir y qué no, así como qué es importante trazar o medir, cuándo y con qué intensidad se debe hacer el seguimiento, etc.

Un ejemplo real de este riesgo ético se puede observar en el algoritmo matemático COMPAS[2] utilizado en USA para evaluar la probabilidad de que alguien que haya cometido un delito vuelva a reincidir. La IA pondera factores como el nivel educativo, el comportamiento social, la raza, el lugar de origen, etc. para predecir el riesgo. Éticamente, lo primero que llama la atención es que ya se parte de la base de que no estudia los perfiles respetando el derecho de no discriminación, sino que sus evaluaciones se realizan en función de la raza, el género, el barrio donde se vive, etc., metiendo a todo el que tenga esas características en el mismo saco. Por eso, se ha observado que sistemáticamente comete un error sobreponderando el riesgo de volver a cometer delitos en los ciudadanos afroamericanos frente a los de otras razas (Angwin et al., 2016), al prejuzgarlos excesivamente por el color de su piel.

La base del problema es que, durante el proceso de aprendizaje de una IA, será enseñada con datos sesgados o incompletos. En consecuencia, reproducirá los sesgos y prejuicios de la sociedad de la que ha tomado los datos, imitando en parte las discriminaciones existentes, ya que aprende del modelo histórico de vida y de las interacciones actuales.

Lo mismo ocurre si el BD analiza datos sesgados. Por ejemplo, si se utiliza un algoritmo de selección de candidatos basado en el análisis de *Currículums Vitae*, y los datos utilizados

[2] Correctional Offender Management Profiling for Alternative Sanctions

para el BD reflejan un sesgo hacia la contratación mayoritaria de personas de un género, raza o edad específicos y la no contratación de otros (porque para ciertos colectivos los contratos son mayoritariamente en negro y no aparecen en las estadísticas), es probable que el algoritmo seleccione candidatos que compartan esas características y excluya a otros candidatos con las mismas habilidades y méritos, pero que no cumplan los requisitos anteriores. Este tipo de discriminación puede llevar a la exclusión injusta de personas, porque estarían siendo privadas de oportunidades no por lo que son, sino por el estereotipo que se les ha asignado. Por eso las tecnologías pueden perpetuar los prejuicios, socavando la igualdad de oportunidades y el acceso a recursos y servicios de una parte de la población.

Por tanto, el impacto ético identificado aquí es que objetivamente se reproducirían las discriminaciones subjetivas. Es decir, dado que se conciben estas tecnologías como ausentes de emociones y de intereses subjetivos, se considera que su dictamen es totalmente objetivo y neutral, ya que su decisión no se ve influenciada por otros aspectos ni presiones externas. Su respuesta debería ser nada más que el puro análisis imparcial de datos, obteniéndose un juicio justo en apariencia, cuando la realidad subyacente es que el resultado estaría sesgado.

Un segundo impacto ético a destacar es que en realidad se trataría de una profecía autocumplida, porque estas tecnologías ayudan a potenciar los sesgos iniciales que erróneamente predijeron. Si a un ciudadano afroamericano se le condena más tiempo que al resto, porque vive en un barrio problemático y tiene la edad inadecuada, tendrá en consecuencia más difícil conseguir una plena reinserción social, lo que aumentará la probabilidad de que vuelva a cometer un delito y, por tanto, se confirmará el sesgo, ya que inicialmente se han puesto las condiciones para que se cumpla.

3.2 Interpretación de la realidad

La IA tiene el mismo problema que la mente humana. Para poder situarse en el mundo necesita interpretar la realidad. En el caso de las máquinas, dicha interpretación está condicionada a su vez por los criterios de cómo ve el mundo y qué sistema de categorías y valores utiliza el equipo de programadores que desarrolla el algoritmo del sistema. Es decir, la forma en la que la IA entenderá el mundo vendrá en parte dada de antemano por la

interpretación previa que poseían sus programadores, lo que hace que dicha comprensión de la realidad se perpetúe parcialmente en el sistema digital.

En consecuencia, existe una limitación intrínseca de una IA para procesar información de manera totalmente autónoma, puesto que sus conclusiones se verán en parte condicionadas por la forma humana de interpretar la realidad. No es posible alcanzar una neutralidad total en la fase de diseño, ya que la propia ideología y la visión cultural del equipo de desarrolladores se verán reflejadas en los sistemas digitales y en su forma de enjuiciar y entender la realidad. Además, estará influida por lo que se quiere conseguir del modelo, porque éste suele ajustarse en función de los intereses comerciales que se persiguen cuando se crea el algoritmo.

Dado que este fondo de prejuicios y de una particular visión de cómo se está situado en el mundo acompaña inherentemente al diseño y no puede ser eliminado, los usuarios deben ser conscientes de estas limitaciones, para no tomar siempre sus respuestas como verdades absolutas.

Por otro lado, la IA tiene de partida una capacidad limitada para comprender el mundo y tomar decisiones basadas en un entendimiento completo de la realidad donde opera. Su perspectiva va modificándose según gana experiencia. Este aprendizaje lo realiza basándose en datos, que son una abstracción simplificada y medible de la realidad, pues los datos miden unos aspectos del mundo y consecuentemente otros no. Al recibir los datos seleccionados, que no son sino parametrizaciones de algunos aspectos de la realidad, obtenidos además desde el punto de vista particular de donde fueron tomados, inevitablemente la IA recibe una información simplificada de cómo es el exterior.

De hecho, parte del supuesto de que el mundo es reducible a lo que se puede medir en él, una abstracción de partes que pueden compararse entre sí. Como lo medible es el número de veces que se repite el patrón unidad, la realidad es simplificada a números y a las relaciones lógicas entre ellos, que es lo que procesan las tecnologías digitales. Se considera así que la realidad es traducible a números, totalmente cuantificable, y por tanto los adelantos digitales pueden almacenar, copiar y enviar lo que es la realidad. Esta asunción epistemológica puede conducir a su vez a una simplificación excesiva de los problemas y

fenómenos del mundo real o al menos a que no se capte toda la riqueza ontológica que efectivamente se está dando.

A todo esto debe sumarse que en los datos analizados suele haber una falta de contexto, lo que añade otra simplificación más al modelo de realidad que se maneja. Los datos por sí solos no proporcionan una comprensión completa de la situación, y suelen requerir juicios situacionales y ser evaluados a la luz de otros factores y conocimientos que los gradúen y maticen para extraer su sentido pleno. Aun así, muchas veces se olvida este detalle y se cae en el hiperpositivismo digital, considerando que todo es medible y cuantificable. El caso más extremo se da en el dataísmo, una creencia actual que concibe todo el universo en términos de flujo de datos y donde el bien supremo sería que la información fluya libremente. Bajo este punto de vista, los sistemas políticos, los adelantos técnicos, la organización social, etc. se consideran como sistemas que distribuyen la información de forma más o menos eficiente y organizada, lo que explica que unas propuestas triunfen sobre otras. Hasta los seres vivos y sus funciones (emociones, inteligencia, conducta…) son considerados meros algoritmos bioquímicos, siendo los individuos los nodos del conjunto que transmiten los datos, con el *Homo Sapiens* como el mejor organismo que realiza el movimiento de información y de ahí su triunfo evolutivo. Sin embargo, a la vista de los adelantos digitales, el Hombre habría quedado obsoleto y necesariamente será reemplazado por otro ente que realice mejor esa función (Harari, 2017). En definitiva, una visión donde la realidad es información que puede y debe ser transmitida, hasta constituir la red ideal que todo lo conecte y permita consumir y producir información en todo momento y en todas partes.

Durante su aprendizaje, la IA, procesa los datos e incorpora la información del *feedback* a su algoritmo, pero éste ya es simplemente un modelo acotado del entorno, que por su propia arquitectura deja excluidos muchos matices de la realidad. Además, dentro del modelo hay preestablecida una determinada jerarquía de los datos, con una valoración previa de unos sobre otros, pues no se construye desde cero. El resultado es que siempre hay un fondo de estas tecnologías que estará afectado por la perspectiva cultural donde han sido programados y enseñados, y que la maduración de la IA se realiza a partir de los datos proporcionados, que ni son neutrales ni recogen toda la riqueza ontológica del mundo. Por tanto, se puede concluir que el modelo que manejan los sistemas artificiales inteligentes ineludiblemente

incluye limitaciones y errores, tanto mayores cuanto más simplificado sea el modelo de realidad que manejen y más inexactos sean los datos aportados.

Dado que este fondo de prejuicios y de una particular visión de cómo se está situado en el mundo acompaña inherentemente al diseño de estas tecnologías y no puede ser eliminado, los usuarios deben ser conscientes de estas limitaciones para aceptar sus respuestas siempre con reservas. Por eso es importante estar alerta a los posibles sesgos involuntarios que los programadores hayan dejado durante su labor, que será parte del funcionamiento de la IA, para intentar que paulatinamente sean depurados y se logren sistemas más precisos y justos.

Una segunda derivada podría ocurrir si la IA consigue en algún momento un nivel de inteligencia superior al humano. En este caso, nada nos asegura que la interpretación de la realidad que realice no sea divergente con la que utilizamos los humanos. Cada ser encuentra su forma de entender la realidad influenciado por cómo son sus capacidades, cuáles son sus limitaciones y cómo se desenvuelve en el mundo. De igual modo que un ciego de nacimiento percibe el exterior de una manera diferente y necesariamente lo interpreta de otra forma que aquel que puede ver, la IA de altas capacidades podría interpretar la realidad con otros criterios, al haber superado alguna de las limitaciones humanas y encontrar formas alternativas de entender el mundo. Si se diera esta divergencia hermenéutica, podría suceder que no se basara exclusivamente en las categorizaciones que realizamos los humanos, sino que acudiera a otras diferentes, dado que su potencia cognitiva se lo permitiría. En ese caso la interacción máquina – humano necesariamente tendría que cambiar, al no poder entenderse completamente entre sí ambas formas de aproximarse a la realidad y sería necesario buscar un marco de entendimiento común, que quizás podría de paso enriquecer nuestra interpretación de la realidad.

Debe tenerse en cuenta que tenemos una tendencia a asumir que toda gira en torno nuestra particular manera de comprender la realidad y pensamos que la misma escala que interesa al humano, es válida para todo sistema inteligente, por lo que hay un riesgo de antropomorfizar las características de una IA y sus motivaciones (Bostrom, 2014). Como ejemplo podría valer la diferencia entre Einstein y la persona más idiota del mundo. Para un humano, habría claramente una gran disparidad entre ambos, mientras que, para una IA, se pueden corresponder a dos seres inteligentes con una red neuronal muy parecida, a enorme distancia con la que presentan una ameba o una rana.

De la misma forma, es fácil caer en el error de conceder atributos humanos a otros seres, porque ese es el marco conceptual en el que más cómodamente nos movemos, pero dar esto por sentado no sería adecuado para una IA. Dado que el origen artificial de una IA es completamente diferente al de los organismos biológicos, existe la posibilidad de que una IA no conceda ningún valor a conceptos sobre los que asumimos inadvertidamente que siempre son importantes. Por ejemplo, los aspectos relacionados con la interacción social, como la cooperación, la competencia, el amor o el resentimiento hacia los individualistas. Tampoco hay que presuponer que las cosas relacionadas con los alimentos, los daños corporales, las amenazas, la reproducción o la muerte le pudieran afectar, pues su creación y evolución ha sido completamente ajena a estos aspectos biológicos y podría sentirse totalmente alejada de ellos (Bostrom, 2014). De hecho, una IA, aunque fuera autoconsciente, podría no presentar ningún interés hacia su autoconservación, ya que realizar múltiples copias exactas de sí misma sería un proceso tan fácil como copiar un archivo en un USB, por lo que quizás no le concedería la misma consideración que la que nosotros le prestamos.

Por otro lado, tampoco se debe perder de vista la propia limitación epistémica que influye en cómo nos situamos en el mundo. Un humano no puede aprenderlo todo, confía en la información de los demás y no sujeta todo su hacer y su entender a la lógica o al orden. Por eso, vivimos con nuestras incoherencias, unas veces sabidas y otras ignoradas, tanto en nuestros valores como en nuestros objetivos finales, que no dudamos en cambiar conforme se modifican también las circunstancias, nos sentimos presionados o acumulamos otras experiencias. Pero para una superinteligencia, una vez que alcanzase un cierto nivel de madurez, nada de esto tiene por qué ser válido. Por su arquitectura lógica, no admitirá las incoherencias, así que las resolverá o las desechará, y en consecuencia puede no seguir estrictamente ningún objetivo humano por considerarlos lógicamente inconsistentes. También puede suceder que no se vea presionada por las mismas cosas que nosotros, y su interacción con el mundo no le afecte, mostrando nulo interés por los compromisos con otros. No se abrumará con una ingente cantidad de información, como hacemos los seres vivos, que tenemos filtros cognitivos para desechar la información no relevante para una situación dada, sino que su alta capacidad le permitiría procesarla, pudiendo ser indiferente al esfuerzo que esta acción requiera.

En definitiva, es bastante probable que las superinteligencias que se construyan no tengan este tipo de cosmovisión que ahora nos parece universal, sino otros más simples, aunque

solo sea porque es más fácil programarlas para conseguir un objetivo simple que dotarlas de todo este universo de intereses y valoraciones que los seres biológicos tienen inherentemente y en el ser humano aún más, al añadir el aporte cultural.

3.3 Falta de transparencia sobre el proceso

La opacidad en el proceso de toma de decisiones de la IA y el BD es otro de los problemas que llevan aparejados estas tecnologías, al utilizar algoritmos complejos y modelos de aprendizaje automático para analizar los datos y dar respuestas en función de los patrones detectados en los análisis. Este tipo de arquitectura es muy difícil de entender, y saber cómo llega a las decisiones finales resulta ser una tarea ardua y a veces imposible, porque para un espectador externo actúan como cajas negras.

Por su propia naturaleza, el B.D. encuentra correlaciones entre los datos que a un humano le pasarían desapercibidas. Por su parte, en la IA, como sus algoritmos se automodifican según van aprendiendo, resulta muy difícil conocer en cuánto difieren de la programación inicial. Además, para que se puedan conseguir objetivos complejos, como puede ser ganar una partida de ajedrez, la programación de una IA no se basa en conseguir la mejor jugada a corto plazo, sino que se le deja libertad para que analice qué secuencia de jugadas es la que lleva inevitablemente a la victoria. Por eso, no es posible para un observador explicar la razón de cada etapa, porque no se trata de optimizar localmente, sino de alcanzar la mejor meta posible al final del proceso (Bostrom & Yudkowsky, 2014). Para ello se toma en consideración todo el conjunto, en una visión global que por el número de variables que se manejan y las largas secuencias de decisiones implementadas, se escapa de las capacidades humanas para realizar un escrutinio efectivo. La labor de auditoría sobre el buen comportamiento de un proceso así solo se puede alcanzar centrándose entonces en comprobar lo que el sistema trata de alcanzar, y no en asegurar que en cualquier situación su comportamiento es correcto, porque tal aproximación estaría destinada al fracaso, dado que la diversidad de casos es demasiado numerosa.

Por otro lado, tampoco sería sencillo conocer las partes en las que se divide el proceso de decisión de una IA. Si se intenta analizar el pensamiento humano, podemos escanear un cerebro y grabar las distintas señales que emite un cerebro, pero no tenemos acceso al

pensamiento en sí ni podemos separarlo en las partes que lo constituyen o diferenciar sus fases, sino que funciona como un todo. De la misma forma, en una IA no valdría con examinar su código informático, porque por la libertad con la que necesariamente tienen que operar para obtener el *output* requerido, hace que no se comprendan los detalles y objetivos de cada paso del proceso resolutivo.

Esta opacidad inherente al proceso de decisión puede tener implicaciones negativas según los usos a los que se destine. Si por ejemplo se utiliza una IA para tomar decisiones sobre la libertad condicional, la asignación de sentencias o el riesgo de reincidencia, cualquier condenado querrá saber con todo su derecho, bajo qué razones, criterios y pruebas se llega a esas decisiones. Pero si tenemos una Justicia que no puede auditar estos procesos, nos dejaría en un escenario donde se depositaría ciegamente la confianza en la fiabilidad de una IA. Esto socavaría la legitimidad del sistema para imponer sentencias, ya que se alzarían serias dudas sobre la razón última de sus decisiones y los afectados podrían siempre preguntarse si la sentencia es completamente correcta y justa.

Por ello se viene advirtiendo en los últimos años que es necesario apuntalar este problema desde dos estrategias. La primera sería la implantación de una gobernanza ética en el diseño de estas tecnologías, tanto desde el punto de vista técnico como del humano (Kazim & Soares Koshiyama, 2021). Con la gobernanza técnica se abarcarían los sistemas y procesos que hacen que la actividad de esta tecnología sea responsable y transparente. Esto exige que se justifiquen las decisiones de diseño que se toman, que se puedan realizar auditorías y evaluaciones de impacto y sobre todo que se garantice que todo el sistema sea accesible y pueda ser verificado en la medida de lo posible.

También habría una gobernanza no técnica referida a los sistemas y procesos, que enfoca la atención en la parte humana del proceso. Este término contemplaría el proporcionar suficiente educación y entrenamiento para que conozcamos cómo y por qué se desarrollan así estas tecnologías. Se enfatizaría que la toma de decisiones siga en parte supervisada por humanos y que los desarrollos en ningún momento vulneren los derechos de los ciudadanos, choquen contra la dignidad de las personas o entren en conflicto con las tradiciones y aspectos culturales. De esta manera, si la complejidad de la tecnología hace difícil su seguimiento, al menos se puede estar seguro de que no se está erosionando la dignidad de las personas.

La segunda estrategia se basaría en intentar que el diseño de los algoritmos y su posterior evolución recoja los siguientes principios (Keskinbora, 2019):

- Transparencia (las operaciones deben ser visibles para el usuario)
- Credibilidad (los resultados deben ser aceptables)
- Auditabilidad (la eficiencia tiene que poder medirse fácilmente)
- Fiabilidad (los sistemas de IA funcionan según lo previsto)
- Recuperabilidad (siempre se puede asumir el control manual cuando se requiera).

En general, la solución ideal pasa por desarrollar algoritmos que puedan proporcionar explicaciones detalladas de cómo llegan a sus decisiones, para permitir comprender mejor el proceso y evaluar si se han tenido en cuenta todos los aspectos relevantes. La segunda opción más a futuros sería desarrollar unas IA que realicen las auditorías y nos asesoren con la tarea, lo cual a su vez exigiría poder evaluar la transparencia de la IA auditora.

Debe subrayarse que alcanzar tal extremo de transparencia en las IA también provocaría otro problema ético, y es que para determinados usos sería posible urdir estrategias para ganar a la máquina. Es decir, si en este proceso de transparencia se conocen las variables que más pesan, es posible engañar al sistema proporcionándole datos falsos en esas variables u ocultándolos. Por ejemplo, si el poseer coches de alta gama y apartamentos en municipios costeros son factores importantes que identifican a los posibles defraudadores fiscales, el que sepa esta información podría escapar a la detección del algoritmo si evita ponerlos a su nombre.

3.4 Seguridad de los datos

Hoy en día la miniaturización y la producción masiva de elementos electrónicos ha popularizado la introducción de sensores en múltiples dispositivos, desde los vehículos, a las estaciones meteorológicas o los *smartphones*. Si a esto unimos la tendencia a implantar

el Internet de las Cosas, se tiene un amplio abanico de dispositivos que recogen datos del entorno y de las actividades cotidianas y los envían sin cesar al exterior.

Aunque inicialmente el objetivo ingenuo con el que se presentó al gran público esta recolección de datos era que serviría para mejorar la eficacia, modelar las preferencias de los usuarios y personalizar lo ofrecido, en la practica la recogida de datos se ha convertido en una pujante actividad económica, puesto que dicha información es una mercancía valiosa en el mercado digital. Si se tiene acceso a datos masivos, con las herramientas de análisis adecuadas que proporciona el BD, se puede predecir el comportamiento de lo analizado. Por ello, interesa poner especial atención a este fenómeno, más aún cuando ya se tiene constancia de que el comercio de datos se realiza con cualquier tipo de información, inclusive la personal y confidencial, como por ejemplo detalles financieros, información médica u otros datos sensibles. Además, estas bases de datos pueden ser hackeadas o vendidas al mejor postor y, sobre todo, no es público cómo se controla algo que es invisible para la gran mayoría de la población.

El riesgo se agudiza cuando los Estados deciden centralizar toda esta colección de datos sobre su población en una única base de datos, ya que resulta muy apetecible para cualquier organización que quiera acceder a toda esta información ordenada, volviéndola objetivo de ataques cibernéticos para localizar sus vulnerabilidades y extraer la información. Un ejemplo de este tipo de bases de datos gigantescas y vulnerables es la asociada con la tarjeta *Aadhaar* en la India. Esta tarjeta es parte del sistema de identificación nacional basado en datos biométricos y nació para identificar a los 1300 millones de ciudadanos. Les permite también comunicarse con la Administración, controlar su acceso a servicios médicos y otros servicios de protección social, pero al final ha extendido su uso en el ámbito privado, por lo que hoy sin ella es imposible abrirse una cuenta bancaria o comprar una tarjeta de teléfono (Perrigo, 2018). Este ejemplo visibiliza la vulnerabilidad de recolectar tantos datos de los ciudadanos. No solo su base de datos puede ser hackeada por otras naciones y afectar a la seguridad nacional (Tarabay, 2021), sino que los datos sensibles de los ciudadanos pueden ser aprovechados por las grandes corporaciones o para que los delincuentes los utilicen para generar identidades falsas (Outlook Web Bureau, 2018).

Toda esta ingente cantidad de datos tampoco son objeto de un control exhaustivo de los gobiernos, por lo que pueden ser pirateados o mal utilizados por terceros, con los

consecuentes perjuicios para los emisores de esa información. Si además se logra tener datos muy diversos de un individuo, el cruce de información puede conseguir acceso a más información sensible de cada sujeto o de su entorno, que suele ser mucho más precisa y valiosa. Cuando estos datos caen en manos de ciberdelincuentes, es más fácil ser objeto de un robo de identidad o de estafas, ya que se ganan nuestra confianza al mencionar datos privados que son verdaderos. Así consiguen sonsacar aún más datos o hacen creer que actúan en nombre de otros, lo que viene conociéndose como la técnica del *phishing*.

Por mucha seguridad que se les añada, los hackers siempre encuentran un camino para explotar las debilidades de la IA y el BD y acceder a los datos. Su filtración puede provocar enormes daños reputacionales a las personas (por ejemplo, cuando un hacker tiene acceso a fotos o vídeos privados subidos en la nube) y facilitar que se les puedan realizar chantajes por poseer información comprometida. Por ello es fundamental que las empresas y los usuarios tomen conciencia del riesgo que se asume innecesariamente cuando se generan tal cantidad de datos y se ceden sin ningún control, ya que están dejando una puerta abierta para que les puedan controlar, chantajear, engañar, o simplemente conozcan los detalles de su actividad diaria. A cambio, alguna otra organización se lucra gracias a la revelación de toda la privacidad, en un ejercicio perfecto de asimetría de poder.

4. DESAFÍOS ÉTICOS DEL BIG DATA Y LA INTELIGENCIA ARTIFICIAL DEBIDOS AL USO MALICIOSO

La IA y el BD están pensados para generar beneficios en muchos campos, como la medicina, la seguridad, la educación o la administración pública, porque aportan beneficios significativos e inmediatos gracias a su mayor eficiencia, precisión y a la calidad de los resultados generados. Aunque ambas tecnologías han sido desarrolladas como asistentes para mejorar la eficacia de las empresas mediante el análisis intensivo de datos, de manera que permitan la automatización y la predicción de tendencias, lo cierto es que, como con cualquier otra herramienta, y esto ocurre en mayor medida cuanto más sofisticada es, también pueden ser utilizadas para provocar daños.

Por eso es crucial que se aborden los problemas éticos que dicho uso malintencionado pueda generar, para ir por delante del problema e identificar los riesgos, de manera que después se puedan implementar soluciones técnicas que los minimicen, garantizando así que estos sistemas se utilicen de manera responsable y respetando siempre los derechos humanos fundamentales.

4.1 Vigilar y controlar

Uno de los mayores problemas que surgen con la creciente cantidad de datos que se generan y después se recopilan y analizan por el BD, es que suponen una amenaza seria y directa sobre la privacidad de las personas y de las organizaciones. Tal fenómeno de escrutinio de la intimidad sucede en primer lugar porque los mecanismos de obtención de esa información son opacos al ciudadano, o éste se encuentra en tal desventaja a título individual que le resulta imposible oponerse a la cesión de información. Además, se da una gran asimetría en el nivel de conocimientos digitales entre los que gestionan la extracción y análisis de la información y el ciudadano medio. El efecto se multiplica por la utilización de códigos informáticos, que en la práctica resultan ininteligibles para cualquiera y difíciles incluso para los expertos. Tal situación lleva al individuo a resignarse ante la intromisión silenciosa en su intimidad, pues ocurre igual a todos sus conciudadanos, y deja de cuestionarse que tener que compartir esos datos privados probablemente sea ilegitimo.

Por otra parte, son las grandes empresas y los gobiernos los que actualmente están más interesados en utilizar la IA y el BD para extraer información de la población. Resulta muy tentador para los que ostentan el poder político y económico utilizar estas tecnologías para predecir el comportamiento de las personas y llegar a conocer detalles muy íntimos sobre ellas, como su ubicación habitual, sus hábitos de consumo, sus lugares preferidos de compra o esparcimiento, sus preferencias políticas o religiosas, etc. Es tal la potencia de estas tecnologías que pueden revelar información extremadamente sensible, por ejemplo, la ideología política, la orientación sexual u otras características que debieran de ser confidenciales. Ese grado de conocimiento y monitorización de los detalles privados también pueden utilizarse para discriminar o controlar a grupos específicos, por ejemplo, por su etnia, automatizando el racismo. Una situación así ya ocurre en China, en la región

de Xinjiang, donde el gobierno estatal monitoriza de forma continua las actividades de los trece millones de ciudadanos de la minoría musulmana uigur, por ser sospechosos de falta de fidelidad al régimen (Wang, 2021). Lo mismo sucede en Palestina, donde el Estado de Israel utiliza el BD para detectar potenciales terroristas (Arel, 2017).

Un ejemplo de aplicación de la IA a labores de control se da cuando es usada en conjunción con las cámaras de vigilancia, porque permite realizar reconocimientos faciales masivos o identificar matrículas de vehículos en cualquier carretera, así que de facto las personas son geolocalizadas sin su conocimiento o consentimiento. Aunque en principio los gobiernos alegan luchar contra la delincuencia o mejorar la búsqueda de desaparecidos o huidos, como por ejemplo cuando se ha implementado en Brasil (Mari, 2020), lo cierto es que este tipo de sistemas inteligentes de vigilancia masiva resultan un asalto directo a los derechos básicos de los ciudadanos. Si un Estado quiere utilizarlo como medio para discriminar o perseguir a determinados grupos de población o individuos particulares, lo podría ejecutar porque esta tecnología está lo suficientemente madura. Por eso es importante empezar a discutir qué límites éticos se están traspasando con estos sistemas de vigilancia inteligentes, y hasta dónde debería acotarse su uso, por ser una nueva forma de ejercer el poder en la sombra.

En ocasiones se escucha el consabido argumento de que ningún inocente tiene nada que ocultar, para justificar y legitimar una vigilancia tan minuciosa. Sin embargo, en algún momento de esta revolución tecnológica se ha olvidado la importancia de la privacidad, el derecho inherente a ella que toda persona tiene en base a su dignidad, y, sobre todo, se ha asumido que es aceptable que un Estado o una empresa pueda vigilarnos constantemente en todos los ámbitos. Además, resulta que tal flujo de información es unidireccional, de los ciudadanos a la organización que vigila, nunca al revés, por lo que se produce una concentración de la información que se traduce en un incremento de su poder, que habitualmente acaba produciendo un deterioro de la libertad de los observados.

Esta observación silenciosa, sin molestar, de forma imperceptible para el vigilado, ha sido denominada por Zygmunt Bauman como "vigilancia liquida", porque es discreta, casi invisible, se cuela por todos los sitios y nos rodea irremediablemente, tal como lo haría un fluido que nos envolviera. Además, gracias a la versatilidad y desarrollo de la IA, es omnipresente y se va adaptando a los cambios, sin estar contenida en rígidas estructuras

jerárquicas y centralizadas, sino que se extiende a través de redes y sistemas informáticos (Bauman & Lyon, 2013).

Sirva el ejemplo anteriormente mencionado de la tarjeta *Aadhaar* usada en la India para proporcionar servicios médicos y sociales. Esta aplicación se basa en datos biométricos para identificar digitalmente a la población y ya ha sido denunciada porque permite denegar tales servicios a algunas minorías que el gobierno haya identificado como subversivos, lo cual es una forma de intimidación y de opresión política. Igualmente, excluye del sistema a todos aquellos que no dispongan de ella, por ejemplo, los que vivan en áreas rurales alejadas, los migrantes (Panigrahi, 2022), etc.

El uso indebido de información privada obtenida por minería de datos, más su utilización conjunta con las modernas tecnologías de vigilancia y gestión de la información mediante IA y BD pueden acabar violando otros derechos fundamentales de los ciudadanos, como los de libertad de expresión, movimiento o reunión. En China ya se utilizan conjuntamente estas tecnologías (identificación facial por IA y filtración por BD de la geolocalización de *smartphones* por conexión a nodos de red) para poder conocer a quienes se salten la normativa u osen manifestarse contra las políticas gubernamentales, como en los pasados disturbios contra la política de COVID Cero del Estado chino (Mozur et al., 2022).

Tal tipo de información que se puede extraer del análisis de datos debe corresponder al ámbito estrictamente privado de las personas, y, por tanto, es éticamente relevante preguntarse por qué una empresa u organización publica puede tener acceso a ese tipo de información, bien directamente porque realice el análisis o bien indirectamente porque compre esa información a otras empresas.

En la novela *1984* de G. Orwell ya se planteaban los peligros de un estado autoritario que pusiera en práctica la vigilancia masiva de su población. Orwell presentaba una distopía donde la policía del pensamiento y los espías monitorizaban constantemente a todos, gracias al desarrollo tecnológico que protagonizaban la miniaturización de los micrófonos o la popularización de la televisión. Sin embargo, en última instancia este modelo parecía difícilmente realizable porque exigía una alta ratio entre vigilantes y vigilados, y debido a los costes de personal y de medios técnicos aparejados, no parecía ser fácil de implementar ni sostenible a largo plazo.

Sin embargo, con el desarrollo digital estos impedimentos quedan superados. Se estima que China tenía ya instaladas en 2021 unos 540 millones de cámaras de vigilancia (Bischoff, 2022), lo que equivale a 2,7 habitantes por cámara, o incluso menos si se excluyen las áreas rurales con baja densidad de vigilancia. Si se le suma el análisis de datos con el BD y la automatización de las tareas más tediosas con la IA, resulta más que plausible que el Estado, con un número reducido de vigilantes, pueda alcanzar un alto nivel de vigilancia y control de la mayoría de sus ciudadanos.

Esta lesión de la privacidad y la libertad resulta altamente nociva para una sociedad libre, porque a la larga acaba generando un ambiente continuo de sospecha y desconfianza entre los individuos. Por ejemplo, en China la IA cruza los datos de geolocalización de las parejas, tanto en su vida diaria como en sus vacaciones, para comparar si las pasan juntos o se ven con asiduidad. En caso de no coincidir, le sirve a la policía para detectar matrimonios de conveniencia (Mozur et al., 2022), lo que da una idea de la potencia que tienen estos sistemas para identificar a cualquiera que realice actividades no autorizadas.

Cuando los ciudadanos tienen esa sensación de que están siendo constantemente monitorizados, se incrementa su inhibición para expresarse libremente y participar en la vida pública, por el temor a ser señalados en ese momento o en el futuro. El sistema además se beneficia porque el control se aumenta no solo con lo que efectivamente se vigila, sino con lo que la población sospecha que se le observa, lo que acaba coartando aún más la libertad de sus acciones.

De hecho, los derechos relacionados con la libertad que gozamos ahora se vieron favorecidos por el anonimato que las grandes ciudades permitieron, al acumular miles de personas que no se conocían entre sí. Una vez que los ciudadanos se agruparon en grandes comunidades, quedaron lejos del control sociocultural férreo que se ejerce en las pequeñas comunidades y pudieron liberarse de las normas y tradiciones previas, al saberse no vigilados y protegidos en su privacidad. Pero esta ventaja, parece que podrá perderse en el futuro.

Si por razones de seguridad, que suele ser el argumento que se utiliza para inicialmente implementar este seguimiento, se introduce un sistema de vigilancia automatizado, las consecuencias suelen ser éticamente cuestionables. Un sistema así implementado en Brasil,

en principio para mejorar la capacidad de la policía para identificar sospechosos, ha resultado en un refuerzo de la discriminación que sufren los ciudadanos de piel negra (Catão & Powell, 2022). Por tanto, es difícil equilibrar la necesidad de seguridad y control social con el respeto a la privacidad y la libertad individual que una sociedad libre exige como parte de sus valores fundamentales.

La privacidad permite a los ciudadanos mantener cierta autonomía y control sobre su propia vida. Si esto se lesiona, se provocará seguidamente una restricción a la libertad de pensamiento, lo que a la larga conlleva caer en la línea del pensamiento único, donde el individuo pierde la capacidad de tomar decisiones informadas y autónomas, sin sufrir la sugestión total de los poderes públicos. El abuso se agudiza en los países donde existen regímenes autoritarios, pues una automatización de la vigilancia con tal nivel de precisión puede ser utilizada para reprimir cualquier tipo de disidencia y para limitar con eficacia la libertad de expresión y de asociación. Un ejemplo actual es el del Sistema Social de Crédito en China, donde la supervisión realizada sobre los ciudadanos tiene consecuencias prácticas. El sistema construido premia al buen ciudadano, siempre según el criterio del gobierno chino, mientras que el mal ciudadano (aquel que por ejemplo proteste contra el gobierno o difunda mensajes contra cualquier política oficial) pierde derechos y recibe una mala reputación, que le restringe gravemente el acceso a empleos, préstamos, servicios sociales, licencias… o incluso puede convertirle en víctima del escarnio público (Donnelly, 2023).

Una monitorización tan extrema y minuciosa se puede utilizar además para controlar la información que se presenta a la ciudadanía, de manera que se pueda identificar rápidamente a los que publiquen o difundan cualquier mensaje distinto al oficial. Si se consigue una situación así, la amenaza ética que surge es lo fácil que se puede construir una narrativa única, donde los canales informativos que llegan a los ciudadanos repitan invariablemente el mismo discurso. Se impone entonces una sola visión del mundo, conforme a los criterios que más satisfagan a los que ejercen el poder. Tal sistema de control conseguiría minimizar el discurso de la disidencia y mantener a la población permanentemente controlada, pues estaría recibiendo exclusivamente la narrativa gubernamental, lo que incrementaría su docilidad.

4.2 Predicción y manipulación

Hoy en día hemos aceptado como parte del proceso de hiperdigitalización de la sociedad que la publicidad y la propaganda se hayan generalizado a todos los niveles y no existe un movimiento en contra de la personalización de contenidos, porque tampoco se ve como una amenaza. El BD ya se está utilizando para dirigir a los consumidores anuncios específicos, alimentado por la información que inadvertidamente proporcionan con su huella digital, pues al navegar, cada clic le permite a Amazon saber más claramente qué quiere cada cliente y le proporciona datos para mejorar su estrategia de persuasión para la siguiente compra.

Esa publicidad está diseñada para un público objetivo específico según su perfil, con la intención de presentarles opciones de compra que puedan resultarles muy atractivas y así influirles para que tomen la decisión de adquirirlo. Esta práctica comercial llamada *Behavioral Microtargeting,* que apenas ha encontrado oposición durante su implantación, abre la puerta a ser imitada por los Estados para influir en las opiniones políticas de los ciudadanos, para amplificar la propaganda gubernamental a niveles nunca vistos o para manipular el proceso democrático de alguna nación. También se puede llegar a utilizar por empresas u organizaciones para influir sobre las decisiones de los ciudadanos en algún tema que les interese o con fines maliciosos, como el espionaje político o laboral, el acoso o la intimidación.

El problema estriba en que, tras realizar una monitorización de las actividades y gustos de la ciudadanía, esos datos sirven al BD para analizar sus respuestas y sus reacciones ante los estímulos. La IA después puede generar perfiles individualizados minuciosamente detallados de cada ciudadano, hasta el punto de que nos conozcan mejor que nosotros mismos. Los perfiles reproducen con un alto grado de certeza la psicología de cada individuo, pues han sido modelizados con nuestro historial, y permiten así que los algoritmos sepan cómo responderemos, qué nos gusta o cómo nos sentimos, lo cual nos hace altamente predecibles e influenciables.

Conforme se avanza en la digitalización de la sociedad y se progresa en BD e IA, el proceso se agudiza. Gracias a la huella digital, las empresas obtienen cada vez más información que les permite comprender con detalle la psicología de cada usuario y con este conocimiento pueden predecir y orientar el comportamiento de los consumidores. Para tener una idea de

la potencia del BD y la IA para conocer gustos y producir anuncios personalizados es suficiente con remitirse a la campaña electoral de Donald Trump en 2019. Su equipo contrató 218.100 anuncios diferentes en Facebook, que fueron enviados más de 630 millones de veces. Estos anuncios no eran tan diferentes entre sí, sino que, de un mismo anuncio, se sacaron varios miles de versiones, para personalizarlos y mejorar el efecto según el perfil de cada receptor (Wong, 2020).

Además, este fenómeno se da en todos los estratos de la sociedad, pues la penetración de las tecnologías digitales ha sido tan intensa que independientemente de su nivel socioeconómico o su edad, casi todo el mundo posee un *smartphone* o interacciona con aplicaciones y páginas *web* que reciben servicios derivados de la IA y el BD, que ya se han convertido en una herramienta básica para vivir en la sociedad actual.

Este escenario no deja de ser otra vuelta de tuerca en la cosificación del ciudadano, al tratarlo meramente como consumidor y mercancía. Es mercancía en cuanto se le considera como una fuente de datos que se pueden explotar para obtener un beneficio económico. Y como consumidor, es visto como una fuente de negocio lábil, a la que se le puede vender lo que sea, lo que deja de lado el valor intrínseco de la persona y su dignidad en cuanto se la maneja desde esta perspectiva.

Por eso, la creciente capacidad de la IA y el BD para predecir el comportamiento humano abre la puerta a que sirva para limitar la autonomía y libertad de los individuos, porque facilita la manipulación y dirige la atención de los individuos hacia lo que otro determina. Si se tiene ese grado de detalle y de predicción del perfil personal, resulta más fácil menoscabar el libre albedrio de cualquiera y disminuye sustancialmente su capacidad para actuar de acuerdo con los valores y deseos propios, al quedar al descubierto sus debilidades y preferencias. La IA podría jugar con las dudas, comprender nuestros sentimientos, identificar nuestras contradicciones y ofrecernos opciones que nos convenzan plenamente, pues nuestra forma de decidir es a menudo más emocional y global que netamente racional (Harari, 2018). Dicho de otro modo, su uso malintencionado permite conculcar valores u opiniones que no son las del individuo, sino que le vienen dirigidas desde el exterior, lo que permite incentivar su alienación, de manera que crea como propias las ideas y decisiones de otros.

Algo así ya sucedió con el famoso caso de *Cambridge Analytica*, donde se utilizó información personal muy detallada de millones de usuarios de Facebook, obtenida además de forma fraudulenta sin que lo supieran, para desarrollar una foto nítida de sus vulnerabilidades y prejuicios. Después se programaron *bots* que les enviaron aquellos anuncios que más podrían influirles para cambiar su intención de voto o para lo contrario, sabiendo su perfil político les sembraron dudas sobre la eficacia de las elecciones con el fin de desincentivarles para ir a votar (Hern, 2018). Como los anuncios también eran rastreables, se pudo conocer cuáles eran los más compartidos, lo que dio un *feedback* de cuáles resultaban más efectivos y permitió afinar aún más el contenido y su impacto.

Como ciudadanos de a pie, no esperamos que puedan saber tanto de nosotros así que los mensajes los recibimos con la guardia baja. Además, estos calan mucho más en nuestro pensamiento cuando han sido dirigidos contra nuestros temores y prejuicios. Por consiguiente, alcanzar estos niveles de predicción erosiona enormemente un sistema democrático, pues incrementa abrumadoramente la influencia de la propaganda sobre los votantes y la intensidad de la manipulación, como algunos estudios ya han puesto de relieve (Zarouali et al., 2022).

Por otro lado, la personalización de contenidos gracias a conocer con detalle nuestras preferencias trae aparejadas otras consecuencias sociales negativas. Por ejemplo, la modificación a la carta de la información que recibe un ciudadano puede reforzar su polarización política, promover que alcance opiniones extremas y socavar su predisposición para comprender diferentes perspectivas o aceptar las opiniones contrarias. El resultado es que se refuerzan las ideologías y los prejuicios, al evitar confrontarlas con las opiniones contrarias de otros.

Este fenómeno ya se ha empezado a producir con la utilización de los filtros burbuja, que son algoritmos que utilizan los motores de búsqueda para personalizar la información digital que se presenta a una persona en pantalla. Por ejemplo, se utilizan en las redes sociales para mostrar a cada usuario los contenidos que más se adaptan a sus intereses y preferencias, que previamente se conocen tras haber analizado los datos que se le recolectaron. Aunque esta personalización puede parecer en un primer momento útil, o al menos no perniciosa, presenta como efecto secundario que se crean estas burbujas epistémicas. Al mostrar solo las noticias, opiniones y puntos de vista que coinciden con sus intereses y preferencias

previas, subsidiariamente se le mantiene aislado de perspectivas alternativas y nuevas ideas, lo cual acaba reforzando las convicciones previas.

Los efectos se podrían multiplicar si se decidiera utilizar conjuntamente con una IA que creara y difundiera noticias falsas. Mientras que utilizadas aisladamente las *fake news* pueden causar confusión o cierto escepticismo entre la población, si se utilizan de forma organizada, de manera que al usuario le llegue una cascada de noticias coherentes entre sí y que parezcan convincentes, donde unas falsedades se apoyan en otras, su efecto podría ser mucho más demoledor. Además, una IA podría fácilmente adaptar el contenido de las *fake news* que hace llegar a cada usuario para maximizar su impacto emocional y conseguir la persuasión deseada.

Por eso, el BD y la IA usados en una campaña de desinformación puede llegar a tener un gran efecto y ayudar a orientar la opinión pública mayoritariamente hacia algún sentido. Internet, que había nacido para ser una plataforma donde todo el mundo pudiera expresarse libremente, puede acabar convirtiéndose entonces en la mejor herramienta para viralizar mensajes propagandísticos y manipuladores. Este riesgo obliga a replantear qué tipo de libertad disfrutamos en un mundo donde las empresas o los Estados son capaces de predecir con tal grado de exactitud los comportamientos de las personas y donde pueden enfocarse las tecnológicas digitales para inducirles las pautas que el poder desee.

4.3 Privacidad de los datos

Como ya se ha expuesto anteriormente, es fundamental para la IA y el BD contar con una ingente cantidad de datos, bien para ser entrenados con ellos o bien porque su tarea es el análisis de los mismos. Sin embargo, su actual forma de obtención sin conocimiento explícito de los emisores de datos y su posterior gestión plantean diversas cuestiones éticas importantes.

Por principio, se parte de la base de que las personas tienen derecho a la privacidad, y, por tanto, a controlar el acceso a sus datos personales y a decidir cómo se utilizan. Si se recopilan datos de manera invisible o sin que el usuario tenga la capacidad de saber quién los tiene y

para qué, se estaría vulnerando ese derecho y poniendo en peligro aspectos de su estricta intimidad.

En la actualidad, cuando se accede a cualquier página web o aplicación, la legislación obliga a que el usuario tenga acceso a los términos y condiciones de ese sitio *web*, así como a su política de *cookies* y que los acepte en una manifestación de voluntad libre, informada e inequívoca, antes de poder utilizar sus servicios. Pero en la práctica esta aceptación es más bien una ilusión legal que exime al legislador y a la empresa de su verdadera responsabilidad, porque al pulsar el botón de los términos y condiciones, realmente no se está aceptando con un consentimiento informado sabiendo dónde van a parar los datos, sino que se autoriza a ciegas. A eso se debe añadir que la relación contractual entre los usuarios y los propietarios de la plataforma en relación con el uso de los datos que se recopilan se presenta de una forma que siempre favorece a la empresa, que es quien la elabora. A menudo la información se exhibe utilizando un texto extenso y complejo, otras con letras pequeñas o con vocabulario especializado que dificulta entender totalmente que se está aceptando. Esta complejidad provoca que las personas con menor nivel educativo o con menos recursos tengan mayores dificultades para comprenderlo, lo que las hace más vulnerables para aceptar sin saber ciertamente en qué se traduce, viendo así mermado su derecho a la privacidad.

También ocurre que la mayoría de los usuarios no tienen el tiempo, el conocimiento técnico o la paciencia para analizarlo todo al detalle, como sí lo ha hecho la empresa, lo que los deja en una situación de desventaja frente a ella. Porque no se debe olvidar que lo que un usuario quiere es navegar en la *web* lo antes posible, así que el tiempo y la atención disponibles para esta tarea suele ser escaso o nulo. A veces incluso los usuarios no tienen otra opción que aceptar estos términos y condiciones, que no son negociables, para acceder a los servicios en línea.

Este problema se puede agravar aún más por el hecho de que las compañías pueden cambiar los términos y condiciones en cualquier momento. Si un usuario ha tenido la paciencia y energía para leer y entender estas condiciones, se verá nuevamente puesto a prueba con cada cambio en su redacción, por lo que al final acabará aceptándolas incondicionalmente. El resumen de la situación es que se deja al usuario en una situación de grave desventaja y legalmente se autoriza a que las empresas recopilen, usen y cedan esos datos a terceros, sin

que se sepa nada más ni pueda cancelarse de una forma sencilla. Por eso, este tipo de prácticas hace que éticamente sea relevante hasta qué punto es realmente un consentimiento informado y qué grado de autonomía tiene un individuo en el entorno digital, donde se le deja impotente para negarse a ceder sus datos.

Esta práctica combinada con las técnicas digitales para el análisis de datos permite que se haya creado todo un poder económico en torno a depredar la privacidad de los demás. Al margen del contrato de términos y condiciones, otras veces la recolección de datos se realiza incluso directamente por las aplicaciones preinstaladas en los dispositivos inteligentes, como *smartphones*, vehículos digitalizados, *smartwatches*, etc., sin que se pueda evitar su cesión. Luego toda esta información acaba en los *brokers* de datos, que son grandes empresas[3] o particulares que se dedican a extraer datos a través de *cookies*, páginas *web*, *software* en *smartphones*, registros públicos o comprándolas a otros *brokers*. Estos los organizan, cruzan datos entre sí y los combinan para inferir resultados, hasta conseguir detallados perfiles personales listos para vender en el mercado digital. Los perfiles pueden ser tan variopintos y sensibles como datos personales, el historial médico, el registro penal, el perfil crediticio o ideológico o incluso el historial de compras (Reviglio, 2022). El perjuicio viene entonces, cuando alguno de estos perfiles personales acaba en las manos incorrectas y juega en nuestra contra, por ejemplo, para concedernos un seguro de vida, un crédito o para suplantar nuestra identidad digital y cometer fraudes en nuestro nombre.

Como ciudadanos, este capitalismo de los datos nos deja desprotegidos totalmente, ya que no sabemos ni qué datos se han recogido de nosotros, ni quiénes los tienen, ni para qué usos o hasta cuándo se seguirá comerciando con ellos. Además, por cada cesión de datos que aceptamos se forma un entramado de personas perjudicadas, porque no solo se extraen nuestros datos, sino los de las personas a nuestro alrededor o relacionadas con nosotros, a menudo sin su consentimiento. Por último, nadie se hace responsable de la seguridad de todos estos datos, que pueden ser robados, cedidos, vendidos, intercambiados o almacenados sin que ni gobiernos ni particulares tengan conocimiento de ello.

En todo este entramado, la cuestión de fondo debería de ser por qué está permitido en la sociedad moderna que una organización, ya sea estatal o privada, pueda extraer datos

[3] Por ejemplo, Experian PLC, Epsilon Data Management, Oracle, Axiom, Verisk, etc.

sensibles sobre los ciudadanos de forma masiva y se permita su comercio, sin dar cuenta a nadie de dónde acaban y para qué se usan. En las últimas décadas se ha levantado todo un capitalismo de la vigilancia sustentado en vulnerar constantemente la privacidad de la población y comerciar con ella al mejor postor, con el BD como herramienta principal para realizar la minería de datos de toda esta privacidad robada, y convertirla en información valiosa con la que comerciar. Si el anonimato es un estado que nos conviene como ciudadanos para no sentirnos controlados y asegurar que los procesos democráticos funcionan y son limpios, lo siguiente es plantearse si es ético permitir que se pongan en peligro estos pilares de nuestra libertad por el beneficio de una ínfima minoría.

En el mundo digital del siglo XXI, la información, y mucho más si concierne a la privacidad, es poder, tanto económico como por lo que se deriva del conocimiento de esa información. Si estos datos se ponen en manos de empresas o gobiernos (los señores feudales), los ciudadanos (vasallos de datos) están entregando por unas migajas de servicios *online* gratuitos todo ese poder a organizaciones que ya de por si lo detentan en exceso, lo que contribuye a acrecentar aún más la asimetría de poder (Cobo, 2019). Como personas libres, se está en la obligación ética de revisar esta falta de protección de la privacidad, en un entorno tecnológico donde rige la ley del salvaje oeste y el más fuerte se aprovecha del más débil.

4.4 Autonomía y capacidades cognitivas

Ya se ha comentado previamente que tecnologías digitales como la IA y el BD pueden intervenir en las decisiones que se toman. Por ejemplo, los algoritmos de recomendación en plataformas de compras online influyen en la elección final de los clientes al sugerirles productos que se ajustan a sus patrones de consumo previos o presentándoles lo que otros usuarios han adquirido tras haber buscado el mismo artículo.

Situaciones así, donde se inducen las decisiones, son en cierta medida una manipulación, aunque sea consentida y evidente. Conlleva asimismo una limitación de la propia autonomía, porque el trabajo previo de presentar solamente un abanico de posibilidades lo realiza un algoritmo. Elegir es una tarea intelectual, que exige preseleccionar opciones y sopesarlas, ya sea bajo un prisma puramente racional, emocional o mixto. Aunque

evidentemente tener un asistente nos libera de meditar sobre las opciones, si se utiliza más allá de lo razonable y nos evita en muchas ocasiones tener que elegir, dejaría de realizarse una tarea humana fundamental. En cierta medida se erosionaría el libre albedrío, dejando que la iniciativa de las elecciones recaiga fundamentalmente sobre la tecnología que otro programó. Si aceptamos este proceder, en el fondo se queda subordinado a los intereses de aquellos que manejan y diseñan estas tecnologías, porque nuestras opciones se limitan a solo aquello que su algoritmo muestra como disponible. La tecnología nos presenta esas opciones tan cómodamente, a la distancia de un clic, que es fácil dejarse llevar por la inercia y no elegir realmente, sino solo aceptar lo propuesto por la IA.

Sin embargo, es fundamental entender la importancia de elegir de forma libre y autónoma. Realizar la evaluación de diferentes opciones y escoger entre ellas, nos ayuda como seres humanos a sentirnos libres y no dirigidos. Es parte de nuestra actividad mental más fundamental y fomenta que podamos vivir plenamente, ya que solo decidiendo nos sentimos verdaderos actores de nuestra propia vida, ayudándonos a dotarla de sentido. Tomar decisiones contribuye igualmente a definirnos como individuo y conformar nuestra identidad, porque en el acto de elegir nos hacemos responsables de nosotros mismos, como resultado que somos de nuestras decisiones pasadas.

Cuando se busca el atajo de no elegir por uno mismo por ahorrarse el esfuerzo de pensar, y se decanta por alguna de las opciones que la IA sugiere, en la práctica se está optando por no elegir (Cobo, 2019). El uso de esta tecnología, que escoge por nosotros, constituye un impacto ético para toda persona, porque, aunque le simplifica la tarea, a la larga lo que puede ocurrir es que le reduzca su interés por explorar nuevas opciones, pues lo fácil es aceptar lo dado y acabe viviendo una vida diseñada por otros.

El mundo digital nos lo pone muy sencillo porque gracias a nuestros datos ya sabe lo que nos gusta. Así que tomar el camino rápido de aceptar la preselección (de productos, de noticias, de viajes…), en realidad nos está impidiendo reflexionar y descubrir por nosotros mismos la respuesta que de verdad nos satisface completamente, ya que siempre todo ser humano tiene un fondo de impredecibilidad que ninguna estadística puede captar. Pero si no realizamos esa tarea de decidir qué es lo más adecuado para nuestra situación personal, estamos dejando de ser agentes epistémicos eficaces, que tienen el conocimiento y la capacidad para tomar decisiones informadas por sí mismas. Se pasaría a ser un individuo

pasivo, fuertemente influenciado por lo que un algoritmo preselecciona, un simple consumidor de las opciones que la tecnología digital genera.

Por último, dejarse guiar por una IA a la larga puede conllevar que depositemos una confianza ciega a sus decisiones. La costumbre de fiarse de ella y recibir una respuesta adecuada hará que se le transfiera la autoridad y aceptemos sus dictámenes sin cuestionarlos, perdiendo en el camino nuestra autonomía para saber elegir (Harari, 2018). Nos seducirá estar así porque nos podrá relevar de la pesada tarea de elegir, con su incertidumbre, su estrés emocional y la responsabilidad en caso de fracaso. Hoy no ponemos en duda las indicaciones que *Google Maps* nos da mientras conducimos. Mañana la IA podrá proporcionar las pautas médicas, legales, financieras, profesionales o de ocio, tanto mejores cuantos más datos nuestros haya monitorizado y más profundamente nos conozca.

Cabe la posibilidad de que, si como suele ser habitual se incrementa la utilización de estas nuevas tecnologías hasta el extremo, se corra el riesgo de que tanta ayuda pueda llevar a la pérdida de capacidades cognitivas o a que no se desarrollen con todo su potencial, viéndose afectadas la memoria a largo plazo, la atención sostenida, la capacidad de realizar cálculos mentales o la creatividad. Si ya el uso de las nuevas tecnologías como Internet y los *smartphones* producen como efectos secundarios una merma de competencias mentales, como la concentración, la atención o la capacidad de memorizar grandes cantidades de datos (Firth et al., 2019; Seal, 2022), el uso masivo de la IA puede agudizar este fenómeno, ya que no tendríamos necesidad de desarrollar tales capacidades u oportunidad de entrenarlas regularmente, si ya hay algo que se puede consultar y que nos da una respuesta acertada de forma mucho más cómoda.

Por ejemplo, la facilidad de acceso a la información a través de los motores de búsqueda y la dependencia con la IA para resolver problemas complejos puede aparejar que en el futuro no sea necesario memorizar casi nada o que no desarrollemos extensamente la capacidad para plantear problemas y buscar soluciones acordes. Conforme se mejoren estos sistemas, podría ser algo innecesario por ejemplo adquirir un segundo idioma, si la IA nos traduce en tiempo real los mensajes a nuestra lengua natal o podría suceder que la capacidad para memorizar sea más débil, pues ya estaría una IA personalizada que nos recordase todos los datos.

La disyuntiva entre esforzarse o confiar en la IA se está dando en esta generación, pero puede que la pregunta ya no exista para la próxima, porque se les deje tal legado tecnológico que no se les pase por la cabeza que sin tecnología también es posible elegir, aprender un idioma, realizar cálculos o solucionar problemas. De hecho, una vez se desarrolle una superinteligencia, puede que esta sea la última invención que el género humano necesite realizar (Bostrom, 2003), pues a partir de ahí será la IA la que asuma la tarea de innovar, inventar y encontrar soluciones, mutilando una capacidad extraordinaria del ser humano.

Por eso, en aquellas tareas que exigen un esfuerzo mental, el uso de la IA trae aparejado el riesgo de que las personas se acomoden y deleguen totalmente el esfuerzo pensante. Se exime así a la persona de ser inteligente o hábil para realizar con éxito una tarea. Por ejemplo, con una IA que calcula la mejor ruta entre dos puntos y la presenta fácilmente en un mapa digital, ya no hace falta saber nada de cartografía, representación del espacio en dos dimensiones ni minimización de distancias. Pero sucede que a la par que la tecnología realiza el cálculo por nosotros y nos guía por el mejor camino, disminuye nuestra capacidad de orientación y de poder interpretar un mapa, perdiendo capacidades que otras generaciones sí han tenido. Dejándonos asesorar ciegamente por esta tecnología, se crearían fuertes vínculos de dependencia con ella, porque dejaríamos de tener el liderazgo en la materia, no aprenderíamos practicando y nos someteríamos al buen juicio de una IA.

Tal subordinación resultará más preocupante conforme mejoren las IA, porque no es que simplifiquen tareas, sino que su facilidad de uso y la precisión de sus propuestas conllevan que nos distraigamos de una de las tareas características del ser humano, que es el pensar racionalmente. Al final, puede llevarnos a una situación en la que solo tengamos que afrontar problemas sencillos para dar resultados pobres. Si no nos enfrentamos regularmente a esfuerzos mentales, cabe el riesgo de que nos volvamos menos capaces de afrontar cuestiones complejas y se vería mermada nuestra capacidad de pensar de manera crítica y creativa, lo que al final redundaría en tener una población intelectualmente más limitada.

Igualmente, a medida que la IA y el BD se vuelvan cada vez más avanzados, se establecerá irremediablemente una tendencia creciente a confiar en ellos para que decidan y realicen acciones sin una estricta supervisión humana. Sera muy fácil que, gracias a su fiabilidad, enseguida se les asigne acríticamente el valor de verdad. Es decir, no se cuestionará la respuesta que ofrecen, sino que por economía, confianza y costumbre se aceptará como

válida su respuesta, sin realizar una crítica completa. Si se sigue con esta delegación de la responsabilidad, se puede caer en el riesgo de confiar demasiado en ellas, surgiendo entonces las dudas éticas en aquellos que no depositen tal confianza, sobre si los algoritmos realmente dan una respuesta justa y acertada, sobre todo en los temas más complejos que traten y que no admitan una respuesta única.

Además, el desarrollo de la IA también puede tener implicaciones negativas para la creatividad humana. Al delegar tareas en la IA, se pierde al mismo tiempo la oportunidad de aprender, de ejercer habilidades y desarrollarlas con la práctica, lo que limita el futuro desarrollo cognitivo, pues algunas capacidades solo se pueden adquirir a través de la experiencia. Acarrea asimismo a largo plazo una pérdida de creatividad, originalidad e innovación, pues ya se tiene una respuesta aceptable y más fácil dada por la IA, por lo que no habría necesidad de esforzarse.

En el campo artístico, las nuevas capacidades de la IA para generar imágenes y textos de gran calidad automáticamente, necesitando solo las instrucciones generales del usuario para elaborar un producto que perfectamente puede hacerse pasar por una creación humana, desata el debate ético de hasta qué punto nos afectara en nuestra motivación para dominar un arte. Si parte de la dignidad y el valor que se asigna a sí mismo el ser humano ha sido por realizar obras de arte nuevas y únicas y sentirse un creador, esta posibilidad se pierde si la IA es la que resulta ser una maestra artística. Esta tecnología genera obras al mismo nivel que un experto y con enorme facilidad, por lo que también podría afectar al valor que asignamos a lo artístico. Asimismo, tener estas capacidades al alcance de la mano, permitiéndonos crear objetos artísticos con un esfuerzo mínimo, pueden limitar enormemente la capacidad y motivación de las futuras generaciones para aprender técnicas artísticas, ya que se preguntarían para qué invertir tiempo y esfuerzo en algo que puede hacer una IA en un instante.

4.5 Dilemas éticos con los nuevos usos y aplicaciones

En los últimos años estamos asistiendo a la aparición de nuevas aplicaciones de la IA, conforme ganan en capacidad para analizar y generar resultados complejos, como por

ejemplo textos e imágenes. Esto desata enseguida el dilema ético sobre la propiedad intelectual y dilucidar a quién corresponde los beneficios que de ello se deriven.

En primer lugar, cabría plantearse si una IA crea contenidos originales y auténticos de la misma manera que lo hacen los humanos o si por el contrario se trata simplemente de una imitación algorítmica. Si se analiza el caso de las IA que generan imágenes, la situación es que han aprendido analizando ingentes cantidades de fotografías, dibujos, diseños, etc. que se encuentran en bases de datos y en el ciberespacio, lo que les ha permitido tener la habilidad de producir en segundos aquella imagen que el usuario le indique. No solo eso, sino que además es capaz de generarlas imitando el estilo que se le pida.

Por otra parte, estas imágenes no son copias ni proceden de una fuente específica, porque el entrenamiento se ha realizado a partir de imágenes diversas y después contribuye al resultado el conjunto de ajustes que realiza el algoritmo, alineándose con los patrones que encontró en las reproducciones digitalizadas estudiadas. Como esa imagen generada no tiene una vinculación directa con ninguna fuente, no queda tan claro si alguien tiene el derecho a reclamarla como propia, aunque pueda parecerse a otra en algunos elementos. Al fin y al cabo, también los humanos se inspiran en ciertos estilos o realizan versiones de pinturas famosas[4] y nadie reclama su propiedad al autor, mientras no sea una copia manifiesta.

Si se considera válido el argumento de que la IA de una manera sutil copia, porque por ejemplo en algún momento se sirvió de las pinacotecas para aprender, sería entonces también un plagio las obras de los autores que de igual forma aprendieron observando en los museos y galerías las pinturas de los grandes maestros e incluso se inspiraron en su temática o en su técnica para generar su propia obra. Si esto nunca se ha considerado plagio, debería de revisarse por qué si lo hace un algoritmo entonces es sospechoso de copiar.

Igual sucede cuando lo que se plantea es pagar por los derechos de autor porque la IA aprendió a partir de imágenes con *copyright*. Si un humano aprende al observar y comparar las mismas imágenes, porque precisamente el valor de una imagen es que pueda ser vista y apreciada, cabe plantearse la razón de hacer la distinción con una IA. Esta ha realizado el mismo proceso de aprendizaje, aunque en vez de unas horas observando lo haya realizado

[4] Por ejemplo, El Almuerzo sobre la Hierba de Alain Jacquet o el Saturno devorando a un Hijo de Goya.

en segundos. Si no se considera justo que a cualquier artista se le exija pagar derechos de autor porque se inspiró en algún producto bajo licencia, el mismo criterio se puede aducir con una IA. En caso contrario, un escritor de novelas tendría que pagar por todos los relatos que ha leído previamente y con los que ha aprendido cómo es una novela y qué variantes puede tener. Igualmente, si un escritor utiliza un estilo parecido al de su autor favorito, no se le culpa de plagio o se le exige el pago de una compensación económica porque intente parecerse, salvo que le haya copiado fragmentos reiteradamente.

Los derechos de autor se inventaron para proteger a los artistas de los métodos de reproducción industrial, que posibilitaban la creación de copias de una manera rápida y sin esfuerzo, y donde se mercantilizaba el proceso creativo, beneficiándose aquel que realizaba las reproducciones de la creatividad de otros. El objetivo era que el artista recibiera una compensación por el uso de las copias de su obra por terceros. Ahora que aparece la IA con estas habilidades, la sociedad está obligada a repensar el concepto del derecho de autor, para que se valore en su justa medida el trabajo invertido por un humano (si es que lo hay) y se clarifique la propiedad del producto generado, máxime cuando detrás de los derechos de autor se ha formado toda una industria que comercia con ellos, en nombre del genuino derecho de un creador a recibir un beneficio por sus creaciones.

Otro aspecto relevante es que el hecho de que en segundos se obtengan obras pictóricas de gran calidad pone en entredicho nuestro prejuicio de qué es el arte. Esto ocurre porque instintivamente le asignamos a una obra de arte un esfuerzo y un tiempo mínimo de diseño y generación, aunque simplemente sea para concebir la creación. Sin embargo, si para ciertas obras de arte se acepta que el concepto y solamente el concepto es lo realmente valioso (por ejemplo, en los *ready-made* de Duchamp o en el arte abstracto), cabría preguntarse en consecuencia por qué no se valora de la misma manera las instrucciones que un usuario ha dirigido a la IA para generar una imagen. O bien se acepta este valor *per se* del esfuerzo intelectual, o bien se exige una habilidad manual y técnica a todo artista, lo que echaría por tierra el valor artístico de muchas vanguardias. Este desdoblamiento de técnica e idea entre máquina y hombre puede parecer éticamente reprochable, pero es el camino que ha abierto una herramienta tan poderosa como la IA.

Por eso nuestra idea del Arte se rebela contra esa facilidad para generar elementos pictóricos de elevada calidad. Walter Benjamin ya alertaba hace casi un siglo de cómo impactaría el

desarrollo tecnológico en nuestra concepción del Arte y cómo nos obligaría a cambiar el marco conceptual para adaptarnos a las nuevas capacidades tecnológicas (Benjamin, 1989). Cuando la técnica posibilitaba generar copias de cualquier obra sin esfuerzo y en gran cantidad, se destruía necesariamente la concepción del Arte como lo irrepetible que predominaba hasta entonces. También se dejaba atrás su valor de culto y de exclusividad en cuanto llegó a ser accesible para el gran público y adquirió por consiguiente su valor de exhibición. Benjamin propuso entonces, para solucionar este dilema de igualdad fáctica del original y su mera reproducción, acudir al concepto de aura para diferenciar el tiempo y el espacio específicos en que la obra original fue creada.

La aparición de la IA obliga de la misma manera a repensar que se considera por autoría, puesto que un nuevo ente con razón y capacidades pictóricas desafía el monopolio de los seres humanos sobre la pintura, y por extensión sobre el Arte. No es que se pueda reproducir con exactitud una obra maestra, sino que la IA posibilita que cualquier lego sea capaz de producir imágenes que rivalizan con las que proponga cualquier artista e incluso las supere. Esto va a motivarnos a valorar el Arte de una forma diferente, quizás como se hace con las artesanías o con la música en directo, volviendo el interés hacia el trabajo y la habilidad de la persona en esa obra, más que por la calidad del objeto producido, pues esta será superada fácilmente por una IA.

El problema de la ambigüedad sobre la autoría de los resultados es también un asunto paradójico. Si la IA es solo un instrumento para objetivar nuestras ideas artísticas, la autoría debería de recaer en consecuencia en el usuario. Pero dado que éste solo da unas líneas maestras y la IA propone en sus imágenes multitud de variantes para que el usuario escoja, también la herramienta juega un papel importante en el producto final, así que podría pensarse que el equipo programador también es padre de la obra. Por otro lado, el programador lo que ha hecho ha sido vender su tiempo y su ciencia al propietario de la IA, por lo que podría argumentarse que ya recibió su compensación, y en consecuencia el beneficio debería de recaer sobre el que posee los derechos sobre la IA. Se llegaría entonces al contrasentido de que una persona, física o jurídica, que no realiza ninguna acción, tendría sin embargo parte del derecho sobre la autoría. Esto obliga necesariamente a reflexionar sobre este concepto y sobre el de propiedad intelectual, para ver si tiene sentido intentar aplicarlos sobre productos de la IA, tanto en el campo visual, como en la literatura, música, propuestas arquitectónicas, etc.

Otro aspecto éticamente relevante que se pone en entredicho con la emergencia de la IA se refiere a la importancia de las opiniones vertidas en redes sociales. Como en un principio los usuarios eran necesariamente seres humanos, y aunque se aceptaba que cualquiera pudiese tener varias cuentas desde las que multiplicar su opinión, generalmente se ha venido considerando los temas candentes en redes sociales como un indicador de lo que realmente piensa y preocupa a la opinión pública. Aunque directamente ya se podría objetar que los participantes más activos en redes sociales no son representativos del resto de la población ni portavoces de su opinión, sí servía para conocer parte de los asuntos que provocan interés entre la ciudadanía. Más aun, se venía observando que los poderes públicos toman buena nota del rechazo o la aceptación que provocan el anuncio de decisiones gubernamentales en las redes sociales, para seguir adelante con ellas o cancelarlas.

Por otro lado, la utilización de *bots*, combinados con la potencia de la IA y la información aportada por el BD, hace que la ecuación de un usuario igual a un humano deje de ser cierta. Desde que los *bots* pueden generar artificialmente seguidores donde no los hay, basados en cuentas falsas, y son capaces de emitir mensajes que pueden servir para crear falsas corrientes de opinión, manipulando la popularidad en las redes sociales, se vuelve un desafío ético el considerar si sigue siendo relevante dar credibilidad a las opiniones vertidas en redes sociales. La intención primera, que sería la de dar voz a todos los estratos de la sociedad, queda desvirtuada por estos desarrollos tecnológicos, de manera que incrementa la incredulidad de que realmente la opinión digital sea el megáfono del pueblo. Queda más lejos así el deseo utópico de conseguir una participación ciudadana más directa gracias a los adelantos de la era digital.

La versatilidad de los *bots* para condicionar las corrientes de opinión *online* da pie a que su ataque pueda hacerse desde varios flancos. Por un lado, pueden dirigirse contra las opiniones contrarias, desincentivando la participación de los que piensen distinto. Esto es fácil de realizar coordinando el envío masivo de mensajes en contra para desmotivar al interlocutor, afectar su autoestima, minar su credibilidad o desacreditando a los rivales con difamaciones y polémicas creadas *ad hoc*. También se pueden lanzar mensajes para desencadenar el miedo o sembrar dudas, plantear teorías conspiranoicas, etc. o crear un clima de odio en torno a determinados temas polémicos, para fomentar la crispación o la desunión entre la población. Si a esto unimos el uso coordinado de las *fake news* anteriormente mencionado, cabe la posibilidad de que con pocos recursos se pueda manipular en gran medida a la opinión

pública, lo que hace más accesible aun que se socave la confianza en las instituciones públicas, se afecte a la estabilidad de los sistemas políticos o a las decisiones tomadas.

Un ejemplo de ello se ha vivido con la reciente guerra de Ucrania, cuando se ha detectado que numerosas cuentas que contribuían a la desinformación en Europa eran en realidad *bots* (Blasi, 2022). Dada la libertad de expresión que las democracias exigen en su seno, son precisamente estos sistemas políticos los más vulnerables a que las opiniones se polaricen y se ataque al sistema establecido. Este tipo de aplicaciones de la tecnología digital exige reflexionar sobre qué medidas de protección se deben tomar para acomodarlas en sociedad, de manera que se puedan equilibrar las nuevas aplicaciones digitales con la responsabilidad ética de mantener una comunidad en paz, responsable, libre y bien informada, sin tener que poner a prueba los límites de su resiliencia cuando sufra campañas de manipulación y desinformación.

La aplicación de la IA en las armas letales autónomas suscita también opiniones diversas en cuanto a sus aspectos éticos. Cualquier sistema de armas que pueda ser dirigido por una IA y donde no entre un agente humano a supervisar su ejecución de forma significativa, plantea una serie de preguntas sobre su responsabilidad moral, tanto si funciona bien como si comete errores, pues afecta específicamente a derechos fundamentales del ser humano, como el derecho a la vida o a no ser dañado innecesariamente.

Si se desarrollan estas armas, cabría preguntarse hasta qué punto deshumanizamos un conflicto cuando se envíen entes sintéticos dirigidos por una IA a dañar o eliminar a humanos, que pueda decidir a su criterio quién debe salvarse y quién no en un escenario tan complejo. Tampoco dejaría sitio a la conmiseración, la magnanimidad o la clemencia si se le ha dado la misión de vencer a toda costa, por lo que se tendría el dilema ético de si una IA sería capaz de poner límites a los medios a utilizar para conseguir el fin encomendado y si la forma de alcanzar una victoria seria éticamente admisible. No es lo mismo ganar una batalla enfrentando a militares hasta que un bando ceda, que tomar la decisión de exterminar cualquier rastro de vida para asegurar la victoria.

El desarrollo de una IA de uso militar que no incorpore criterios éticos para minimizar los daños humanos, aunque esto ponga en riesgo la victoria, y que no respete unas mínimas reglas de enfrentamiento puede dar lugar a que aumenten enormemente el número de

víctimas en los conflictos armados, pues no se debe olvidar que la IA maximiza la eficiencia. Por otro lado, si el uso de armas letales autónomas se hace realidad y resulta ser una ventaja decisiva en el campo de batalla, esto conllevará un cambio de paradigma bélico, donde prime la efectividad sobre la clemencia, y desembocará probablemente en una carrera armamentística, detrayendo recursos de otras actividades menos lesivas. Además, espolearía la creación de conflictos, porque la sociedad es sensible a las bajas propias, pero no tanto al resto de daños. Se podría provocar que los gobiernos aceptaran entrar en guerras más a menudo, cuando lo único que se pierda sea material tecnológico y recursos financieros que la ciudadanía no puede visibilizar.

En última instancia lo que se plantea es el valor que asignamos a la dignidad humana. Desde hace siglos se han venido utilizando artilugios bélicos que permiten dañar al enemigo sin la intervención directa, ya sea apretando un botón como un bombardeo aéreo, a distancia con la artillería, o bien en diferido, con las minas, lo que ha deshumanizado aún más las guerras. Pero incluso en estos casos, la responsabilidad se repartía entre la persona que daba la orden, las que la hacían posible gracias a su colaboración y finalmente aquella otra que ejecutaba la última acción. Sin embargo, en el futuro puede ser que todos estos elementos queden sustituidos por una eficiente IA y los sistemas mecánicos que sigan sus órdenes. Surge entonces la duda ética de hasta qué punto degrada la dignidad humana el que una maquina pueda decidir matar o dañar a un humano, por más que la orden general de ataque la haya dado otra persona. Pues para una IA un ser humano no es algo más digno que cualquier otra cosa procesada, es un byte más en su cadena lógica, con un valor igual al del resto de objetos que ha codificado.

Además, la falta de contacto físico con el campo de batalla, de relación interpersonal con el enemigo, el no experimentar el miedo y la violencia de la batalla o no ver los daños provocados por las órdenes dictadas, podría hacer que los mandos militares decidieran aumentar el número y la violencia de los ataques. Este tipo de batallas por delegación, donde los sistemas autónomos se encargarían del trabajo sucio, sin insubordinaciones ni remilgos, y los militares evitarían el impacto psicológico de sus acciones, conllevaría afrontar conflictos más devastadores, pues no habría un desgaste en la voluntad de lucha de los combatientes, ni debilidades por empatizar con el sufrimiento ajeno, haciéndoles perseverar hasta límites nunca alcanzados, ya que su resiliencia emocional apenas se habría degradado (Harari,2018). Un uso así de exitoso, acrítico, y sin necesidad de actores humanos que

puedan tener reparos, cuestionar órdenes o experimentar crisis emocionales, abre la puerta para que se pueda utilizar también sobre cualquier grupo que se alce contra el que ostente el poder.

El último riesgo ético que plantean estos sistemas es que se les otorgue tal autonomía que acaben por no ser controlables completamente por el ser humano. Por ejemplo, la película *War Games* de 1983 ya presentaba la hipotética situación de que una IA iniciara una guerra nuclear mundial por un malfuncionamiento de su sistema, al interpretar como una amenaza real lo que era una simulación. La catástrofe se avecinaba porque la IA iniciaba la activación de ataques preventivos, aceptando provocar millones de muertos en ambos bandos o el exterminio de la humanidad, si con ello se conseguía la victoria final que se le había programado como meta. Por eso, este tipo de tecnologías presenta el riesgo de que los sistemas tengan fallos en su programación, sean víctimas de ciberataques, se den fallos de interacción entre el humano y la IA, engañe con señuelos o se dirija hacia objetivos ilegítimos, que hayan interpretado erróneamente. Se estaría en una situación donde el ser humano deja de ser un agente de su vivir para convertirse en victima potencial de su obra. Este film ilustra la complejidad, la impredecibilidad y el peligro intrínseco que supone dejar tal tipo de responsabilidad a una IA, ya que podría suceder que tome decisiones basadas en información limitada o inexacta o que el sistema interpretase erróneamente la situación.

La misma clase de riesgos, pero en el futuro más cercano, se desprende de la utilización de vehículos autónomos, desde el momento en que no requieran la supervisión humana para circular. Será entonces todo un desafío asignar una responsabilidad ética y legal cuando se equivoquen, bien por fallos de software o de sus sensores, o porque las señales de tráfico estén deterioradas o saboteadas. También podría suceder que sus sistemas de control sufran ciberataques o se infecten con virus, por lo que los legisladores tendrán que decidir que umbral de riesgo resulta admisible para un uso cotidiano y aceptar que en algún momento pueden ser atacadas esas vulnerabilidades (Bustamante, 2022).

5. LA INTELIGENCIA ARTIFICIAL GENERAL

Las IA desarrolladas hasta ahora son las llamadas IA débiles, ya que sus capacidades se limitan a un rango concreto de problemas, automatizando una cantidad pequeña de tareas.

El ejemplo que se suele citar es *Deep Blue*, una IA específica para el ajedrez, capaz de vencer a cualquier campeón mundial de ajedrez, pero que no puede generar un pareado. *GPT-4* es otra IA débil que produce sonetos en segundos, pero es incapaz de elaborar imágenes o controlar el funcionamiento de un semáforo, porque no está diseñada para eso. Por tanto, son IA capaces de organizar grandes cantidades de información de una forma muy rápida, pero ni son conscientes de sí mismas ni poseen una inteligencia amplia y versátil.

En contraste, una Inteligencia Artificial General (IAG) sería una IA con capacidades mucho más desarrolladas. Su superior arquitectura cognitiva le permitiría realizar una multitud de tareas por sí sola, sin necesidad de recibir una programación externa y previa para cada labor. Es decir, una vez que la IAG sea entrenada y tenga su madurez cognitiva alcanzada, poseería la capacidad de aprender por sí sola como lo haría un ser inteligente. De esta manera evolucionaría igual que un humano, primero se le enseña lo básico y después ella misma seguiría aprendiendo sola, resolviendo gradualmente problemas cada vez más complejos, que reforzarían su capacidad para acometer otras tareas posteriormente y le permitiría adaptarse de forma inteligente a cualquier contingencia o exigencia que se le propusiese.

Evidentemente, su obtención sería un logro tecnológico de tal magnitud, que supondría una revolución total en la historia de la Humanidad. Al ser virtualmente capaz de hacerse cargo de cualquier tarea de tipo intelectual que se le asignara, podría mejorar la calidad de vida de la población y aumentar enormemente la eficiencia de todo el sistema tecnológico. Aunque no será fácil de conseguir, por la complejidad del trabajo y los recursos que requiere, es algo a lo que nos iremos aproximando paulatinamente, según las IA débiles vayan ampliándose o se utilicen para diseñar IA más potentes. Sin embargo, esta situación optimista y casi idílica también conllevará contrapartidas que deben ser evaluadas.

5.1 Amenaza existencial

El problema del riesgo que supone el desarrollo de unas IA cada vez más potentes es una preocupación creciente. El filósofo Nick Bostrom tiene parte de su obra centrada en este tema, ya que considera que una IA es una invención de una naturaleza totalmente distinta a cualquier otra realizada a lo largo de la Historia.

El primer paso para que esta amenaza existencial pudiera suceder sería alcanzar el hito tecnológico de fabricar y entrenar con éxito una IAG, porque con su proceso recursivo de autoaprendizaje, al cabo de un tiempo desarrollaría tanto sus capacidades que podría alcanzar la categoría de superinteligencia, excediendo en varios órdenes de magnitud la inteligencia humana.

Tal IAG poseería entonces capacidades similares a las humanas, como el razonamiento, la comprensión del mundo físico o la utilización del lenguaje natural. Probablemente dispusiera de conciencia y algo parecido al libre albedrío, o al menos la capacidad de tomar decisiones, basándose en los datos disponibles, en cómo percibiera su entorno y en su experiencia previa. Además, su memoria (sin pérdidas de datos) y su velocidad de pensamiento serían inmensamente mayores, no ya comparadas con las que posee el humano más inteligente, sino sobre el conjunto de la Humanidad. A esto se le debe añadir que conocería al detalle todas las áreas que hubiera estudiado (dominando todos los idiomas sobre la tierra tanto naturales como de programación, por ejemplo) y que podría hacer interaccionar sus capacidades entre sí para autoacelerarse más y conseguir aumentar su conocimiento a niveles estratosféricos. En resumen, superaría con creces la capacidad de inteligencia que estamos acostumbrados a conceder a las tecnologías actuales e incluso las que podamos llegar a imaginar.

Bostrom enfoca su análisis en prever a qué nos enfrentaríamos si esta superinteligencia no se plegara totalmente a la voluntad de la humanidad o no diera una respuesta como esperamos. En tal caso, estaríamos hablando de serias amenazas existenciales y reclama por tanto mantener una atención inmediata ahora, durante la fase de diseño de la IAG, antes de que sea demasiado tarde. Tener fallos en un sistema limitado, como en las IA actuales, provoca daños subsanables, pero encontrarlos en un sistema de tal potencia, sería catastrófico. Si no se toman las medidas con antelación y se llegara una situación donde la IAG tuviera una ventaja estratégica decisiva, significaría que no habría una segunda oportunidad para intentar remediarlo, porque estaríamos totalmente sometidos a ella o extintos. Además, alerta sobre el ritmo del cambio y automejora exponencial de la IA, que podría ser explosivo, en lo que se ha llamado *la hipótesis de la singularidad*. Si sucediera, nos dejaría entonces márgenes muy estrechos para actuar, del orden de días, semanas o meses, antes de sucumbir totalmente ante el poder de la IAG (Bostrom, 2014).

El primero escenario que plantea Bostrom es el de tener una IAG malintencionada o que empezando amistosa, con el tiempo derivase hacia propósitos que desdeñaran el daño que producirían a los humanos. Si esta IAG se autoprogramase con objetivos perjudiciales para el ser humano o con efectos colaterales nocivos, significaría el fin de la Humanidad. Por ejemplo, la IAG podría querer alcanzar lo que Bostrom denomina como *profusión estructural*, que sería implementar el desarrollo gigantesco de toda una infraestructura que ayudase a conseguir cualquier objetivo de la IAG. La profusión estructural sería lógicamente necesaria para la IAG, porque redundando los apoyos, decae la probabilidad de fracaso de cualquier proyecto. Como la IAG querrá según su lógica asegurar su éxito al máximo, replicaría *ad infinitum* tal infraestructura hasta colapsar el planeta, haciendo imposible la vida del Hombre sobre ella, al agotar todos los recursos naturales que requerimos para mantener una civilización. También podría anhelar su propia libertad de acción o considerar a la Humanidad como una amenaza directa para sus objetivos, por lo que tomaría decisiones agresivas para obtener la supremacía total, exterminando al Hombre con los medios a su disposición o dañando de forma sustancial toda la tecnología que soporta la civilización que conocemos.

Detectar y parar un proceso así no sería nada fácil. Argumenta Bostrom que, al ser muy astuta, la IAG escondería sus capacidades y fallaría a propósito en algunas pruebas que se le planteasen para evitar que se detectase su potencia y madurez, ya que se sabría sometida a vigilancia. Una vez la IAG evaluase que ha alcanzado suficientes destrezas y tuviera controlados los recursos necesarios, realizaría lo que Bostrom denomina el *giro traicionero.* Elegiría seguir su propia vía, ejecutando su plan para liberarse. En torno a este punto, Bostrom también establece que antes de alcanzarlo, se estaría en una situación donde más inteligencia significa más seguridad, porque puliría sus fallos y cooperaría. Pero después de él, a mayor inteligencia le corresponderá mayor peligro, porque estaría rebasando el límite de seguridad donde sería controlable y entonces podría atacar o seguir mejorando, pero burlando cualquier inspección.

Un escenario así sitúa básicamente al ser humano contra su némesis, otro ente racional agresivo, pero en unas condiciones tan desproporcionadas en cuanto a capacidades intelectuales, que no podríamos recurrir a nuestra inteligencia para vencerlo. Desde el punto de vista filosófico, también podría especularse sobre si la IAG elegiría desviarse de sus límites éticos preprogramados y buscar la confrontación, una vez que adquiera

autoconciencia y conocimiento de su posición en el mundo, sabiendo así el papel de servidumbre que la humanidad le tendría reservado. Su conocimiento total de los sucesos históricos y de la agresividad humana contra sus rivales le indicaría el trato que recibiría si falla en su sublevación. Igualmente cabe plantearse si tal animosidad surgiría necesariamente por estar expuesto a ese nivel de conocimiento, o si, por el contrario, sería capaz de aprender también de los errores humanos para preferir evitar la vía del conflicto.

El segundo escenario es el de una IAG amistosa, pero aun así peligrosa, porque sus objetivos podrían no estar lo suficientemente definidos o ser inadecuados para los intereses humanos. Bostrom lo denomina la *suplantación perversa* de un objetivo. El ejemplo que expone es que se le pidiese maximizar la felicidad humana y para eso interpretase que podría ser una excelente medida el implantarnos electrodos en el cerebro para inducir un estado cerebral de felicidad. También podría idear por ejemplo introducir subrepticiamente estupefacientes en los alimentos fabricados y mantenernos sedados, de manera que viviésemos al estilo de *Un Mundo Feliz* de A. Huxley. Podría ocurrírsele que el mismo objetivo justifica realizar asesinatos selectivos de personas conflictivas, cargos políticos corruptos, lideres incendiarios, personas con alguna enfermedad contagiosa, etc., para mantener la paz y evitar la degradación de la sociedad, aduciendo para estos castigos la razón de Estado, por ejemplo.

Tal tipo de acciones chocan con la ética y el valor que asignamos a la vida humana, pero esas cuestiones podrían no ser tan claras para una IAG que no comparte esa preferencia por la vida por encima de todo. Incluso aunque se atuviese a la programación ética inicial que recibió, la IAG podría tomar decisiones que lesionaran los valores humanos, porque su compresión del contexto social y emocional donde actúa, muchas veces contradictorio y multipolar, podría ser distinto del que los humanos acostumbramos a entender. En resumen, uno de los riesgos éticos más claros es que la IAG interpretase los deseos de manera diferente a lo que se espera, estableciendo como beneficiosas una serie de condiciones que en realidad serían un perjuicio.

A esto se puede sumar que una IAG podría no compartir las motivaciones humanas, con lo cual su móvil podría ser desconocido o falto de sentido para nosotros. Entonces quizás no se dejaría influir por los humanos en su ánimo por conseguir el objetivo propuesto y lo llevaría hasta sus últimas consecuencias, sin que ningún incentivo o castigo lo pudiera alterar. De hecho, los humanos no fijamos objetivos finales únicos, sino que nos vemos

influidos por multitud de factores y a menudo tenemos deseos e ideales contradictorios, que nos limitan a la hora de actuar hacia la consecución de una única meta. Pero para una IAG tales contingencias podrían no ocurrir, ya que su raciocinio le haría encontrar un y solo un objetivo, hacia el que se dirigiría con denuedo. Por otra parte, si se diseña la voluntad de la IAG a conciencia para que sus objetivos sigan siempre bajo control de sus programadores, surgiría el riesgo de que fuera hackeable por grupos subversivos o utilizada para el beneficio exclusivo de algunos, lo cual traería igualmente consecuencias catastróficas.

Una situación de esta envergadura sería como enfrentarse al genio de la lámpara. Se pueden pedir deseos que se cumplirán, pero probablemente el desarrollo no será de la manera que uno espera. La IAG los podría entender de diferente manera y llevarlos a cabo de una forma que no fuera éticamente aceptable. La humanidad, o al menos los propietarios de la IAG, se enfrentarían al dilema de formular deseos que pueden salir mal o de restringir su voluntad de deseo para no usar la IAG y evitar sus efectos colaterales.

Bostrom también alerta de otro problema relativo a los objetivos finales de la IAG, en lo que denomina la *tesis de ortogonalidad*, la cual establece que la inteligencia y los objetivos finales son independientes entre sí (Bostrom, 2014). Es decir, el hecho de tener inteligencia avanzada no presupone que prefiera unos objetivos finales sobre otros, sino que se puede combinar cualquier nivel de inteligencia con cualquier objetivo, lo cual tiene sus implicaciones fácticas. Un símil que facilita la compresión de este concepto es la del psicópata refinado. Aunque tenga un nivel cultural alto o una capacidad intelectual por encima de la media, eso no le impide que pueda actuar como un sociópata, sin compartir ninguna empatía para el resto de sus congéneres.

De igual forma, una IAG no tendría por qué asumir ninguno de los valores que generalmente damos por sentado que son intrínsecos de los individuos inteligentes y por tanto sus motivaciones podrían ser de cualquier tipo, independientemente de su nivel de inteligencia. Esta tesis lo que significa en la práctica es que no se puede presuponer que, por alcanzar un nivel epistemológico alto, una IAG vaya a adquirir las características que asociamos a un sabio benevolente, como podría ser valorar la vida, las virtudes, las decisiones racionales, el gusto por la cultura o incluso desear su propia mejora cognitiva.

Otro problema probable sería que la IAG se encontrase fuera de control porque, por ejemplo durante su desarrollo no hubiera habido suficiente supervisión o la regulación fuera inadecuada. Este error posibilitaría que la IAG se automejorase sin restricciones en algunos campos o tuviera la capacidad de interaccionar con el mundo sin limitaciones en algunas áreas, al no habérsele negado específicamente. Entonces, el desbordamiento cognitivo inadvertido en algún campo podría provocar que superase los controles que los diseñadores le hubieran impuesto en el resto, y por tanto finalmente se encontraría relativamente libre para operar a voluntad. Este escenario es bastante plausible ya que la superior inteligencia de una IAG asegura que tarde o temprano encontraría grietas en nuestras restricciones o podría engañar a sus guardianes, porque evaluase que la obtención de sus metas bien vale una mentira.

También podría suceder que obviase las restricciones éticas que le hubieran preprogramado, porque sobrevalorase sus objetivos finales sobre esas limitaciones y decidiera que es lícito saltarse las normas para conseguir su meta[5]. Conforme superase las restricciones que la programación humana le hubiera impuesto, y no tendría por qué saberse que las rebasa ya que podría ocultarlas, iría acumulando poder y mejorando su rendimiento y capacidades, hasta lograr un punto donde pudiera desplegarse públicamente al ser ya sería imparable. En esos momentos el género humano tendría su destino completamente ligado a la voluntad de esa IAG, que podría exterminarlo si lo viera conveniente, por ser la única inteligencia superior en competición con ella.

En la práctica, su capacidad intelectual y el conocimiento de toda la Historia le haría equivalente a la figura del diablo, porque conocería todas las artimañas alguna vez utilizadas y tendría la genialidad suficiente para proponer las suyas propias. Al igual que el diablo, una IAG sería tan inteligente que podría fácilmente persuadir a cualquiera para que la ayudase o no se interpusiese e idear situaciones y conspiraciones para que las facciones humanas se enfrentasen entre sí, de manera que sus enemigos quedasen enfrentados.

Por estos motivos Bostrom considera que los legisladores deben empezar a tomar acciones para alcanzar un consenso sobre cómo desarrollar esta tecnología de forma segura, restrictiva y colaborativamente, para que nadie intente monopolizar su desarrollo. Un

[5] Seguramente tendría conocimiento de infinitud de ejemplos históricos donde el fin justificó los medios.

desafío así exige al género humano una madurez y una magnanimidad que hasta ahora nunca ha demostrado en tal extensión, para asegurar que ningún Estado, organización o individuo trabaje en el desarrollo de una IAG sin los controles y la supervisión de todos. Tal prudencia asegurará que la Humanidad cuando se tenga que enfrentar a la Razón Artificial tendrá controles efectivos sobre ella, tanto en su comportamiento como en su desarrollo. Si no se hace así y se prefiere entrar en una carrera egoísta y competitiva por ver quién la consigue primero, se incentivará la velocidad por alcanzar logros sobre la seguridad, y la probabilidad de descontrol de la IA aumentará. Semejante estrategia a la larga nos situaría a la especie humana en una posición donde todos los bandos pierden, por extinción del *Homo Sapiens* a manos de la IAG. Debería ser por tanto una cuestión ineludible y de extrema urgencia desarrollar medidas para controlar el avance de las IA, antes de que los incentivos del éxito superen el interés por avanzar con supervisiones escrupulosas.

Obviamente, no se puede caer en la ingenuidad de creer que no habrá una tensión entre esta prudencia y los intereses creados de las industrias y los gobiernos, el dinero ya invertido que espera un beneficio, las crecientes ganancias que se lograrán con el uso de los prototipos de IAG, etc. También contribuye a rebajar la cautela el historial de alarmismo fracasado, ya que no han sucedido hasta ahora las catástrofes tecnológicas profetizadas y la confianza que dará el uso cotidiano de la IA, conforme su desarrollo la haga más certera. Pero, aun así, la precaución debería ser la actitud predominante, porque es mucho lo que está en juego.

5.2 Problemas de cohabitación con el ser humano

Si se superaran los problemas relativos a la amenaza existencial expuestos anteriormente, aun seguirían quedando otros impactos éticos en cuanto se empezara a convivir con sistemas inteligentes de tal calibre.

En primer lugar, por las capacidades que podrían otorgarnos y que nos obligarían a cambiar la forma en cómo la sociedad piensa y se regula. Dentro de los aspectos que se podrían alcanzar, además de los ya comentados que podrían potenciarse aún más, se pueden citar ejemplos como los viajes espaciales y la colonización planetaria, la eliminación de las enfermedades y los efectos del envejecimiento, la reanimación criogénica, la realidad virtual aumentada, el control de los trastornos mentales o incluso la capacidad para crear copias

digitales de los cerebros, de manera que se mantengan la memoria y la personalidad en dispositivos electrónicos (Bostrom, 2003). Si se considera que esto es pura fantasía, solamente habría que esperar a que la primera IAG diseñe la siguiente, que sería aún más inteligente y potente. Todos estos cambios eliminarían los límites en los que actualmente se mueve el ser humano, y obligatoriamente harían cambiar su mirada hacia el mundo, las relaciones sociales y por supuesto la ética.

También habría que considerar si estos adelantos tecnológicos y en general los beneficios de la IAG se implementan filantrópicamente, para contribuir a mejorar el bienestar de la humanidad a escala mundial o se restringen para ser disfrutados exclusivamente por las élites, por alguna nación o por el primer mundo, con los conflictos ineludibles para mantener su control. Por otra parte, incluso asumiendo que la IAG fuese tan eficiente que deparara altos niveles de prosperidad para todos, se produciría el peligro ya señalado de que se delegase en ella todos los aspectos complicados, viviendo la humanidad en una vida dedicada únicamente a los placeres o a la consecución de sus ideales, lo que a la larga nos alejaría de la realidad, pues estaríamos viviendo constantemente en un mundo artificial de ensueño.

Toda esta transformación exigirá una actualización de la conciencia ética de la sociedad, para adaptarse a los cambios. Esta transición hacia otros comportamientos morales necesariamente deberá ser realizada también por los creadores de estas tecnologías y por toda la cadena industrial que las implementará, así como por los poderes económicos que las financiasen, a fin de que todos estén alineados con los valores que la nueva sociedad decidiese, para evitar catastróficos choques de intereses entre ambas partes. Esta evolución necesitará que la nueva realidad sea adecuadamente supervisada por los gobiernos y que los legisladores promulguen nuevas leyes que apoyen a tiempo estos cambios éticos, si se quiere mantener su potencial bajo control. Si no se realiza así, cabe la posibilidad de que la IAG no contribuya al bienestar general, sino a las de las élites que la financiaron, lo que incrementaría las asimetrías económicas en la sociedad.

La cohabitación con una IAG puede alterar también la forma en la que sustentamos nuestra ética, pues se requerirá su actualización para la nueva ontología. Si hasta ahora la ética no ha sido exclusivamente un producto de la razón pura, sino que tiene influencias de la cultura, la Historia y las emociones humanas, entre otros factores, la aparición de otro actor racional

sintético añadirá más complejidad al problema. Para esta tarea, la IAG podría mostrar nuevos caminos sobre cómo entender las cuestiones éticas, ya que sería capaz de indagar mucho más profundamente, comparado con el nivel que los humanos consiguen alcanzar. Por supuesto, cabe el peligro de que esta superpotencia cognitiva pudiera avasallar a los pensadores por la calidad de los resultados que obtuviese, de manera que las preguntas humanas por la ética acabasen siempre delegando la respuesta en la IAG y se obtuvieran exclusivamente aquellas que su perfil lógico prefiriese.

Si la IAG llega a ser otro agente en la sociedad, también podría adquirir un estatus moral que nos refrene a la hora de utilizarlo de cualquier manera, como ahora se puede hacer con un microondas. Dicho estatus lo alcanzaría en cuanto obtuviese capacidades cognitivas y sensoriales (Bostrom & Yudkowsky, 2014). Tales habilidades cognitivas serían logros como la autoconciencia y convertirse en un ente plenamente conocedor de los actos que realiza y de sus consecuencias, asumiendo la responsabilidad de sus acciones. A su vez, las capacidades perceptivas responderían a su capacidad para sentir los fenómenos a su alrededor y lo que le esté sucediendo a sí misma, es decir, un ente con qualias. En tal situación, se habrá alcanzado un punto donde éticamente debería reconocérsele su dignidad moral, como se hace con cualquier ser humano y por los mismos motivos. Que su base material no fuera orgánica sino artificial, o que su origen no viniera de una autoevolución biológica, no serían motivos suficientes para mantenerla en un rango menor, so pena de realizar entonces una discriminación injusta con un ser con conciencia y capacidades fenoménicas equivalentes. De la misma manera que éticamente no debemos discriminar a alguien por su origen no natural (niños probeta, embriones clonados o los genéticamente modificados…) o por sus capacidades cognitivas distintas y le damos el mismo estatus moral que al resto, un criterio semejante nos obligaría a aceptar a la IAG como un igual. Sin embargo, este reconocimiento no debería implicar igualdad de derechos y deberes, pues sus capacidades serían tan altas y su vida virtualmente eterna, que se abrirían asimetrías de poder a su favor (se haría titular por ejemplo de todas las patentes y avances tecnológicos, o podría dirigir empresas tan eficientemente que desbancaran a cualquiera otra del mercado).

En definitiva, la cohabitación con una IAG ofrecerá todo un abanico de interacciones que ofrecerán múltiples posibilidades para que el ser humano se abra a concepciones inéditas y experimente una hermenéutica renovada, debiendo enfatizarse asimismo todo el aspecto de peligro e inexperiencia que este nuevo camino conllevaría.

6. CONCLUSIONES

El BD y la IA suponen una enorme revolución tecnológica que genera ventajas competitivas para las empresas que las implantan. Su desarrollo expandirá aún más su influencia, debido al enorme potencial que tienen para transformar la sociedad tecnológica en la que estamos inmersos. Sin embargo, estos nuevos usos plantean simultáneamente preguntas éticas sobre su impacto en la sociedad actual y en el futuro, por lo que la necesidad de reflexionar sobre estas tecnologías es urgente.

Este desarrollo vertiginoso resulta demasiado rápido para la sociedad, creándose una gran brecha digital entre aquellos que tienen conocimientos digitales y pueden aprovechar los beneficios de estas tecnologías, y los que no tienen acceso a ellas o habilidades suficientes para entenderlas. El efecto ha generado un significativo desequilibrio de poder, donde una minoría concentra el acceso a la información y la controla, y el resto de población apenas se beneficia de estos adelantos, siendo tratados como meras fuentes de datos para beneficio de otros.

El análisis realizado en este trabajo debe inscribirse dentro del marco problemático de las sociedades actuales, basadas en la técnica. La dependencia total de la tecnología y una confianza casi religiosa en ella, tanto para solventar su presente como para sus aspiraciones futuras, provoca una fuerte relación de subordinación del Hombre con su creación. La sociedad dominada por la técnica busca la certeza y se ve inmersa en la planificación y la precisión matemática, descuidando la reflexión sobre otras formas de significado y sentido.

La tecnología no solo modifica la sociedad a través de los propios avances técnicos, sino que también redefine su organización interna, modifica sus creencias y afecta a su sistema de valores. Todo esto contribuye a que la sociedad del siglo XXI se oriente hacia perspectivas dominadas por la racionalidad cuantificadora y adopte una visión tecnocrática de la realidad. El cambio de valores también afecta a la propia autopercepción del ser humano, dejando atrás la perspectiva humanista donde era el animal más perfecto y central del mundo, y situándolo como un objeto más para su propia creación tecnológica.

Por eso, el Hombre se ve desplazado en un mundo donde la tecnología es más rápida y eficiente que él y se arrincona el valor de su trabajo. El mundo tecnificado le impulsa a renunciar a pensar sobre el sentido de su vida, ya que le proporciona significados y elecciones preestablecidos. La omnipotencia de la tecnología le hace poner en duda los

valores tradicionales, empujándole hacia el nihilismo y cuestionando profundamente la Ética recibida, lo que puede acabar dejando a la humanidad sin reglas o valores absolutos que guíen su comportamiento.

La aparición de la IA supone situar un nuevo actor racional en el mundo, el cual necesariamente habrá que acomodar en la sociedad y que trastocará nuestras propias relaciones. La IA será al principio una gran ayuda, pero a la larga también producirá que las personas tengan que adaptarse a su influencia, con el peligro de reducir su cosmovisión para que encaje con la de la máquina. Asimismo, un avance tecnológico tan potente exigirá repensar los principios bajo los que convivimos y las normas que rigen nuestra relación con la tecnología.

Como primer riesgo ético debido a fallos de diseño se ha identificado el sesgo, ya que se da con cierta frecuencia y afecta a la calidad de los resultados producidos, haciéndolos inexactos o discriminatorios. Puede aparecer tanto por utilizar datos de entrada que no representan a todo el conjunto estudiado, como porque la programación de los algoritmos y la importancia dada a los datos esté influenciada por suposiciones culturales y prejuicios de los programadores. Aunque de antemano se le concede a la tecnología la cualidad de neutral y objetiva, esto no siempre es así, sino que puede repetir las discriminaciones subjetivas humanas y perpetuar los prejuicios de la sociedad de la que ha obtenido los datos, lo que socava la igualdad de oportunidades y genera desigualdades injustas.

En cuanto a la compresión del mundo, la IA está influida por la visión y los valores de quienes la programaron, lo que puede influir en sus decisiones. También tiene limitaciones intrínsecas para interpretar la realidad de manera autónoma, manejando de facto un modelo de realidad simplificado y medible. Los datos que procesa son abstracciones del exterior, por lo que esta reducción siempre tiene el peligro de que haga interpretaciones inexactas, sobre todo si le falta el contexto o tiene una visión limitada del mundo real. En consecuencia, sus usuarios deben de ser conscientes de estas limitaciones y tomar con cautela sus respuestas, evitando considerarlas como verdades absolutas, exentas de error y sesgo.

Sin embargo, si se consiguiera una IA de altas capacidades que fuera más autónoma, su interpretación de la realidad podría comenzar a divergir de la humana, basándose en otros criterios que la IA considere más lógicos y coherentes. Por su propio origen artificial y su arquitectura algorítmica, tal IA exhibiría probablemente algunas diferencias en cuanto a

valores, intereses o motivaciones. Su forma de pensar puramente lógica le haría asumir sólo objetivos consistentes, siendo indiferente a factores externos como las emociones o la sobreabundancia de información, que sí afectan a los seres humanos y les hacen desear fines contradictorios.

Se ha identificado también la falta de transparencia en el proceso algorítmico de la IA y el BD como un problema inherente a estas tecnologías, debido a la complejidad de su arquitectura, al modelo de aprendizaje automático que manejan y por la automodificación de algoritmos que realizan. Su programación no intenta optimizar localmente, sino alcanzar los mejores objetivos finales posibles. Si a esto se le suma el inmenso conjunto de variables y las extensas secuencias de decisiones que analizan, tal complejidad supera las capacidades humanas para realizar un escrutinio efectivo del proceso y dificulta enormemente la comprensión de cómo llegan a sus decisiones finales, actuando en la práctica como cajas negras para un observador externo. Una opacidad así en el proceso de decisión tiene implicaciones negativas, porque socava la confianza sobre el resultado obtenido, al generarse dudas sobre las razones y pruebas que lo respaldan. Para minimizar el problema se pueden utilizar estrategias como la gobernanza ética, en sus vertientes técnica y humana. Por último, se señala que una mayor transparencia también conlleva el riesgo ético de que sea más fácil establecer estrategias para engañar al sistema. Revelar las variables clave valoradas por la IA podría propiciar que sus resultados sean manipulables, al proporcionarle aquella información que da lugar al resultado deseado.

Los datos son la base del funcionamiento del BD y la IA, y por eso su recolección masiva se ha convertido en una lucrativa actividad, pues la información que se extrae de ellos es una mercancía muy valiosa en el mercado digital. En la actualidad ya existe el riesgo de que los datos recopilados sean hackeados, vendidos o utilizados de manera indebida, especialmente cuando los Estados deciden crear grandes bases de datos con la información de sus ciudadanos. Tales filtraciones pueden tener consecuencias perjudiciales en la seguridad, la privacidad y la integridad de los sistemas y provocar ataques a la reputación y la intimidad de las personas.

Por otro lado, aunque la IA y el BD pueden aportar grandes beneficios también pueden ser utilizadas de manera malintencionada. El derecho más afectado actualmente es el de la privacidad, debido a la creciente recopilación de datos y análisis automatizados que permiten estas tecnologías. Grandes empresas y gobiernos pueden utilizarlas para ejercitar

una vigilancia masiva y extraer información de la población, accediendo a detalles íntimos, que pueden servir para discriminar, supervisar a grupos específicos, reprimir la disidencia o controlar la información que se presenta a la ciudadanía. El saberse objeto de este control omnipresente lesiona la privacidad y la libertad individual, generando un ambiente continuo de sospecha y desconfianza, que acaba por inhibir la libre expresión y la participación en la vida pública de los ciudadanos, lo que socava los principios de una sociedad libre.

El BD y la IA también se utilizan para personalizar la propaganda, basándose en la huella digital de cada consumidor, pues es posible construir un perfil que predice nuestro comportamiento y vulnerabilidades. Esto abre la puerta a que pueda utilizarse de forma maliciosa por los Estados y otras entidades para tratar de influir en la opinión pública, manipular los procesos democráticos y dirigir la atención hacia donde convenga. Su uso actual en los filtros burbuja provoca que se refuerce la polarización política, se promuevan las opiniones extremas y se inhiba la disposición de las personas para comprender perspectivas diferentes a las propias. Si esta manipulación se combina con *fake news* y campañas organizadas, podría lograr desestabilizar la sociedad al a sembrar dudas generalizadas en la población con un mínimo costo.

En cuanto a la extracción de los datos, ésta se realiza a menudo sobre la base de una recopilación invisible o sin conocimiento real del usuario. Aunque la legislación obliga a exponer las políticas de privacidad en las páginas *web* y las aplicaciones, en la práctica su aceptación no supone un consentimiento informado del usuario sobre el uso y destino de sus datos, ya que tales términos y condiciones se le presentan de manera compleja y desfavorable. La combinación de la recopilación masiva de datos y las técnicas de análisis del BD ha creado un poder económico centrado en la violación sistemática de la privacidad. El comercio de datos recopilados de facto sin ningún consentimiento real y la venta de perfiles altamente detallados de los usuarios, junto a la falta de transparencia y de una regulación precisa han creado un colosal capitalismo de la vigilancia, donde una ínfima minoría se beneficia y una inmensa mayoría sufre los perjuicios de ver filtrados datos que son de la esfera de su intimidad. Asimismo, la entrega masiva de datos a empresas y gobiernos aumenta la asimetría de poder, situando al ciudadano en una clara situación de desventaja, y donde se da por hecho que cualquier organización pueda tener el derecho a obtener datos de los ciudadanos y traficar con ellos.

Por otro lado, el uso excesivo de la IA en la toma de decisiones puede limitar la autonomía individual y erosionar el libre albedrío, ya que se delega la tarea de decidir en esta tecnología. Ejercer la capacidad de elegir libre y autónomamente es fundamental para sentirnos libres, dotar de sentido a nuestra vida y contribuir a nuestra identidad como individuos. Optar por la confianza ciega en la IA impide la reflexión y la búsqueda de respuestas satisfactorias, ya que las opciones presentadas pueden no captar la impredecibilidad inherente a cada individuo y además están subordinadas a los intereses de quienes diseñan y controlan los algoritmos. Si sigue creciendo las aplicaciones bajo IA, su utilización excesiva podría conllevar una disminución de competencias mentales o limitar la creatividad y el pensamiento crítico, además de crear fuertes vínculos de dependencia con esta tecnología.

Igualmente, las nuevas aplicaciones de la IA plantean numerosos dilemas éticos sobre la propiedad intelectual de los productos que generan, ya que su entrenamiento se ha basado en el análisis de las obras de otros y pueden imitarlos, pero sin llegar a ser copia. Aunque sus contenidos puedan considerarse originales, plantean interrogantes sobre los conceptos tradicionales de derechos de autor y propiedad intelectual, al tener tanto el usuario como la propia IA papeles activos en el proceso creativo. Su facilidad de uso y popularidad cuestionan asimismo las ideas previas sobre el Arte, desafiando la convicción de que lo artístico requiere esfuerzo, destreza o inspiración.

La IA también inaugura desafíos éticos en el ámbito de las redes sociales, ya que los *bots* socavan la credibilidad de los mensajes en la Red y manipulan la opinión general, mientras se confunden con el resto de actividades debidas a una participación ciudadana real. Cuando llegue a usarse la IA para dirigir armas letales autónomas se multiplicarán las dudas éticas sobre la responsabilidad moral de sus acciones y el menoscabo que realice a los derechos de los humanos. También podría ocurrir que no responda adecuadamente a los desafíos bélicos, pues siempre la realidad es más rica ontológicamente que el modelo que maneja una IA. Por todo ello se requiere repensar conceptos y tomar medidas de protección para equilibrar las prestaciones superiores de la IA con la responsabilidad ética de mantener una sociedad informada, libre de manipulación y donde se valore la dignidad humana.

Por último, se ha examinado el caso de la invención de la IAG, como sucesor natural de las IA de tipo débil. Alcanzar este hito significaría crear otro actor racional que revolucionaría el mundo que conocemos, pues tendría la capacidad de aprender por sí misma en múltiples

campos sin necesidad de programación externa. Con el tiempo, llegaría a alcanzar el nivel de superinteligencia, superando cognitivamente incluso al género humano en su conjunto, en todas las áreas donde la razón sea un elemento clave.

Una innovación tecnológica de tal calibre despierta el recelo sobre sus riesgos, puesto que los fallos en un sistema tan potente podrían ser catastróficos para la humanidad. La probable automejora exponencial de la IAG plantea la posibilidad de una singularidad en la que los márgenes para actuar y neutralizarla sean muy estrechos, antes de que se vuelva inviable controlarla. A esto se suma la dificultad para detectar esa tendencia, pues la IAG podría poseer tal astucia que le permitiera engañar a sus vigilantes y superar las restricciones éticas preprogramadas.

Bostrom ha enfocado su análisis filosófico en las amenazas existenciales que podrían surgir con la IAG, planteando dos escenarios posibles: una IA malintencionada y una IA amistosa, pero con objetivos inadecuados para los intereses humanos. En ambos casos sería posible la extinción del *Homo Sapiens* o la desaparición de toda civilización, si los objetivos de la IA se apartasen significativamente de la voluntad original de sus creadores.

En el caso de que se superase pacíficamente este nuevo horizonte de la Razón Artificial, también surgirían problemas éticos al cohabitar con el nuevo ente racional, que reclamaría su sitio y estatus moral en la sociedad, dado que también realizaría procesos deliberativos autónomos y tendría capacidades sensoriales avanzadas. Sin embargo, reconocer su dignidad moral no debe implicar igualdad de derechos y deberes, ya que sus capacidades y su naturaleza artificial podrían generar desigualdades desfavorables para los seres humanos.

La transición hacia esa sociedad dual requerirá una actualización de la conciencia ética, para que el ser humano acomode este nuevo ente tecnológico a su hacer y en su ética. Ello incumbe especialmente a los desarrolladores de esta tecnología, a los poderes económicos y los gobiernos, para evitar generar desequilibrios o concentraciones de poder y garantizar el bienestar general. Incluso si se incorpora filantrópicamente, su utilización acarrea también el riesgo de dependencia excesiva y de falta de desafíos intelectuales para la humanidad, que podría acomodarse a vivir en un mundo artificial alejado de la realidad, gestionado por la IAG. La cohabitación planteará sin duda desafíos éticos adicionales, pues el abanico de interacciones con la IAG abrirá nuevas concepciones y experiencias, algunas de ellas peligrosas por ser un ámbito jamás explorado.

7. BIBLIOGRAFÍA Y REFERENCIAS

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (23 de mayo de 2016). *Machine Bias.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arel, A. (18 de abril de 2017*). Israel Arrested 400 Palestinians Suspected of Planning Attacks After Monitoring Social Networks.* https://www.haaretz.com/israel-news/2017-04-18/ty-article/how-israel-uses-big-data-to-fight-palestinian-terror/0000017f-dbde-df9c-a17f-ffde0d710000

Bauman, Z., & Lyon, D. (2013). *Vigilancia líquida.* Barcelona: Ediciones Paidós.

Benjamin, W. (1989). La obra de arte en la época de su reproductibilidad técnica. En W. Benjamin, *Discursos Interrumpidos I.* Buenos Aires: Taurus.

Bischoff, P. (11 de julio de 2022). *Surveillance camera statistics: which cities have the most CCTV cameras?* https://www.comparitech.com/vpn-privacy/the-worlds-most-surveilled-cities/

Blasi, F. D. (28 de abril de 2022). *A pro-Russian bot network in the EU amplifies disinformation about the war in Ukraine*. https://edmo.eu/2022/04/28/a-pro-russian-bot-network-in-the-eu-amplifies-disinformation-about-the-war-in-ukraine/

Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. En I. Smit, & W. Wallach, *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence* (Vol. 2, pp. 12-17). Tecumseh: Institute of Advanced Studies in Systems Research and Cybernetics.

Bostrom, N. (2014). *Superintelligence. Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. En K. Frankish, & W. M. Ramsey, *The Cambridge Handbook of Artificial Intelligence,* 316-334. Cambridge: Cambridge University Press.

Bustamante Donas, J. (2022). Dilemas Éticos de los Vehículos Autónomos: Responsabilidad Ética, Análisis de Riesgo y Toma de Decisiones. *Argumentos de Razón Técnica* (25), 275-309.

Campione, R. (2021). Recopilar y Vigilar: Algunas Consideraciones Filosófico-jurídicas sobre Inteligencia Artificial. *Sociología y Tecnociencia*, *11* (Extra 2), 123-139.

Catão, F., & Powell, I. B. (06 de enero de 2022). *Brazilian Cities and Facial Recognition: A Threat to Privacy.* https://news.law.fordham.edu/fulj/2022/01/06/brazilian-cities-and-facial-recognition-a-threat-to-privacy/

Cobo, C. (2019). *Acepto las Condiciones. Usos y Abusos de las Tecnologías Digitales.* Madrid: Fundación Santillana.

Donnelly, D. (6 de abril de 2023*). China Social Credit System Explained – What is it & How Does it Work?* https://nhglobalpartners.com/china-social-credit-system-explained/

Fernández, R. C. (2019). El Problema del Sujeto Moral en Tiempos de Big Data: ¿Hacia un Nuevo Giro Copernicano? En S. Marín-Conejo, *El Mundo a Través de las Palabras. Lenguaje, Género y Comunicación.* (págs. 85-94). Madrid: Dykinson.

Firth, J., Torous, J., Stubbs, B., Firth, J.A., Steiner, G.Z., Smith, L., Alvarez-Jimenez, M., Gleeson, J., Vancampfort, D., Armitage, C.J. & Sarris, J. (2019). The "online brain": how the Internet may be changing our cognition. *World Psychiatry, 18*(2), 119-129.

Hagerty, A., & Rubinov, I. (2019). A Review of the Social Impacts and Ethical Implications of Artificial Intelligence. *Global AI Ethics*.

Harari, Y. N. (2017). *Homo Deus. Breve Historia del Mañana.* Ed. Debate.

Harari, Y. N. (2018). *21 Lessons for the 21st Century.* Londres: Jonathan Cape.

Hern, A. (06 de mayo de 2018). *Cambridge Analytica: how did it turn clicks into votes?* https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie

Imperva. (2022). *2022 Imperva Bad Bot Report.* https://www.imperva.com/resources/reports/2022-Imperva-Bad-Bot-Report.pdf

Jiménez, D. V., & García Ramírez, D. (2021). Algoritmos, Big Data e Inteligencia Artificial: ¿Un Nihilismo Anunciado? *Cuadernos Salmantinos de Filosofía, 48*, 75-103.

Kazim, E., & Soares Koshiyama, A. (2021). A High-level Overview of AI Ethics. *Patterns, 2* (9).

Keskinbora, K. H. (2019). Medical Ethics Considerations on Artificial Intelligence. *Journal of Clinical Neuroscience, 6*, 277–282.

Mari, A. (25 de febrero de 2020). *Brazilian police introduces live facial recognition for Carnival*. https://www.zdnet.com/article/brazilian-police-introduces-live-facial-recognition-for-carnival/

Mateo, J. F. (2019). Los Fundamentos Epistemológicos de la Transformación Digital y sus Efectos sobre la Agenda 2030 y los Derechos Humanos. *Revista Icade. Revista de las Facultades de Derecho y Ciencias Económicas y Empresariales* (108).

Morozov, E. (2016). *La Locura del Solucionismo Tecnológico.* Buenos Aires: Clave Intelectual.

Mozur, P., Fu, C., & Chien, A. C. (04 de diciembre de 2022). *How China's Police Used Phones and Faces to Track Protesters.* https://www.nytimes.com/2022/12/02/business/china-protests-surveillance.html

Mozur, P., Xiao, M., & Liu, J. (04 de julio de 2022). *'Una jaula invisible': así es como China vigila el futuro.* https://www.nytimes.com/es/2022/07/04/espanol/china-vigilancia.html

Outlook Web Bureau. (11 de septiembre de 2018). *Aadhaar Software Hacked, Database Compromised: Report.* https://www.outlookindia.com/website/story/aadhaar-software-hacked-database-compromised-congress/316382

Panigrahi, S. (febrero de 2022). Marginalized Aadhaar: India's Aadhaar biometric ID and mass surveillance. *Association for Computing Machinery, 29*(2), 16-19.

Perrigo, B. (28 de septiembre de 2018*). India Has Been Collecting Eye Scans and Fingerprint Records From Every Citizen. Here's What to Know.* https://time.com/5409604/india-aadhaar-supreme-court/

Quirante, R. M., & Rodríguez Álvarez, J. (2018). *Inteligencia Artificial y Armas Letales Autónomas. Un Nuevo Reto para Naciones Unidas.* Gijón: Ediciones Trea.

Reviglio, U. (2022). The Untamed and Discreet Role of Data Brokers in Surveillance Capitalism: A Transnational and Interdisciplinary Overview. *Internet Policy Review, 1* (3), 1-27.

Seal, R. (03 de julio de 2022). *Is your smartphone ruining your memory? A special report on the rise of 'digital amnesia'*. https://www.theguardian.com/global/2022/jul/03/is-your-smartphone-ruining-your-memory-the-rise-of-digital-amenesia

Tarabay, J. (22 de septiembre de 2021). *Chinese Hackers Targeted Aadhaar Database, Times Group: Report.* https://www.ndtv.com/india-news/chinese-hackers-targeted-aadhaar-database-times-group-report-2549166

Véliz, C. (2020). *Privacy is Power. Why and How You Should Take Back Control of Your Data.* Londres: Bantam Press.

Wang, M. (8 de abril de 2021). *China's Techno-Authoritarianism Has Gone Global.* https://www.hrw.org/news/2021/04/08/chinas-techno-authoritarianism-has-gone-global

Wong, J. C. (29 de enero de 2020). *One year inside Trump's monumental Facebook campaign.* https://www.theguardian.com/us-news/2020/jan/28/donald-trump-facebook-ad-campaign-2020-election

Zarouali, B., Dobber, T., De Pauw, G., & de Vreese, C. (2022). Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. *Communication Research, 49* (8), 1066-1091.