Facultad de Ciencias Económicas y Empresariales

ESTUDIO DE LA NECESIDAD DE UNA IMPLEMENTACIÓN ÉTICA EN LOS ALGORITMOS DE IA

Autor: Alejandro Chávez Macías

Director: Javier Fuertes Pérez

MADRID | Diciembre de 2023

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Alejandro Chávez Macías, estudiante de Administración de Empresas de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Estudio de la necesidad de una implementación ética en los algoritmos de IA" declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación [el alumno debe mantener solo aquellas en las que se ha usado ChatGPT o similares y borrar el resto. Si no se ha usado ninguna, borrar todas y escribir "no he usado ninguna"]:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
3. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 1/12/2023

Firma: Alejandro Chávez Macías

I. RESUMEN

El presente trabajo tiene como objetivo el análisis de la necesidad de un planteamiento ético aplicable en el diseño y uso de algoritmos de Inteligencia Artificial. Para conseguirlo, se hace uso de una metodología exploratoria, revisando tanto artículos éticos como científicos, con el fin de mostrar en el proyecto un enfoque global del problema.

Desde sus modestos inicios con el perceptrón hasta la sofisticada IA Generativa, se rastrea la evolución histórica de los algoritmos para proporcionar un contexto técnico que enriquezca la comprensión de los desafíos éticos actuales.

Posteriormente se analizan los beneficios que aportan los algoritmos de IA a la sociedad en la actualidad, y se abordan diversas problemáticas emergentes, como el dilema de la responsabilidad o el problema de la generación de información no veraz, a las que debe buscarse solución. Además, se exploran las distintas propuestas de distintas organizaciones en su búsqueda por regular a los algoritmos, destacando la complejidad y la urgencia de establecer marcos éticos en este campo en constante evolución.

La conclusión del trabajo propone la "algor-ética" como un campo de la ética dinámico sobre el cual construir propuestas de soluciones a los dilemas cambiantes de los algoritmos de IA. Desde la algor-ética, se construye una guía aplicable de buenas prácticas en el diseño y uso de algoritmos, inspirada en los principios de la Rome Call for AI Ethics. Esta guía integra propuestas técnicas, como librerías de Python que mitigan el sesgo algorítmico, junto con teorías éticas de la responsabilidad en los algoritmos, como el principio de responsabilidad de Hans Jonas.

Palabras clave: Algoritmos, Inteligencia Artificial, Ética, Transparencia, Algor-ética

ABSTRACT

The aim of this paper is to analyse the need for an ethical approach applicable to the design and use of Artificial Intelligence algorithms. To achieve this, an exploratory methodology is used, reviewing both ethical and scientific articles, in order to show a global approach to the problem in the project.

From modest beginnings with the perceptron to sophisticated Generative AI, the historical evolution of algorithms is traced to provide a technical context to enrich the understanding of today's ethical challenges.

This project then analyses the benefits that AI algorithms bring to society today and addresses several emerging issues that need to be addressed. In addition, it explores the various proposals from different organisations in their quest to regulate algorithms, highlighting the complexity and urgency of establishing ethical frameworks in this evolving field.

The paper concludes by proposing "algor-ethics" as a dynamic field of ethics on which to build proposed solutions to the evolving dilemmas of AI algorithms. From algor-ethics, an applicable guide to good practice in the design and use of algorithms is constructed, inspired by the principles of the Rome Call for AI Ethics. This guide integrates technical proposals, such as Python libraries that mitigate algorithmic bias, together with ethical theories of responsibility in algorithms, such as Hans Jonas's principle of responsibility.

Key words: algorithms, AI, ethics, transparency, algor-ethics

Índice

Índice de figuras

1. INTRODUCCIÓN

1.1. OBJETIVOS

El presente trabajo de investigación está orientado a explorar los algoritmos de Deep Learning y los dilemas sociales que suscitan en la actualidad. En concreto, a través de estas páginas, se tratará de buscar un planteamiento ético que aborde eficazmente estos dilemas y que pueda implementarse en la práctica. Se muestra a continuación un desglose de los diferentes objetivos de este trabajo:

1) **Definir** el concepto de algoritmo y su integración en la inteligencia artificial, y **explorar** la evolución de los algoritmos hasta la actualidad con el fin de **proporcionar un punto de vista técnico** que complemente al enfoque ético del trabajo.
2) **Analizar** el impacto social de los algoritmos de Deep Learning:
    a) **Evaluar** de forma crítica el **impacto de los algoritmos** en la actualidad, junto a sus problemáticas éticas.
3) **Examinar** las propuestas existentes para abordar los anteriores dilemas éticos:
    a) **Investigar** las propuestas éticas de diversas instituciones, abarcando tanto aspectos relacionados con políticas públicas como con el ámbito jurídico.
    b) **Explorar** y **analizar** las propuestas éticas existentes en el ámbito de las políticas públicas que ataquen a los problemas previos.
    c) **Proponer** un marco ético que pueda ser aplicable en el diseño y uso de algoritmos, integrando lo aprendido anteriormente.

1.2. METODOLOGÍA

Para cumplir los objetivos previos, se empleó una metodología **exploratoria** en la confección del trabajo, revisando bibliografía relevante para el tema y extrayendo información que pudiera ser significativa para la investigación. Preferentemente, para analizar los dilemas actuales, se ha optado por recopilar bibliografía de carácter novedoso entre el año 2016 y 2023. Se considera esta metodología la más adecuada para el trabajo, debido, por una parte, a que existe una **falta de consenso** en una ética enfocada en los algoritmos, por lo que es interesante conocer distintas posturas éticas que pudieran ser aplicadas en este ámbito. Por otra parte, la **naturaleza evolutiva** de la tecnología hace necesaria una investigación sobre los más recientes cambios en los algoritmos de IA, además de los nuevos dilemas éticos en los que éstos se encuentran inmersos. Finalmente,

esta investigación serviría para **formular interrogantes** que pudieran ser resueltos mediante un **análisis de la literatura** de la ética recopilada.

1.3. ESTADO DE LA CUESTIÓN

En los últimos años han emergido complejas cuestiones éticas asociadas a los algoritmos de Inteligencia Artificial. Esto, unido al rápido avance reciente de esta tecnología, ha hecho aún más complicada la formación de un consenso en la creación de un marco ético que pueda regular el diseño y uso de algoritmos. Formular ese consenso se vuelve una necesidad dado el impacto significativo que los algoritmos tienen en la sociedad actual. Como se verá, la alta versatilidad de los algoritmos actuales permite que estos sean aplicados en distintos ámbitos como la medicina y la educación. Si bien, esto puede facilitar y mejorar la vida humana, si no se solucionan previamente los dilemas éticos que poseen de por sí los algoritmos, estos podrían terminar por traer sus problemas a otras áreas donde serían potencialmente dañinos. Se hace necesario entonces encontrar un marco ético que pueda evolucionar junto a la tecnología, además de una propuesta de ética aplicada, que permita a los desarrolladores de algoritmos crear algoritmos éticos, con un enfoque antropocéntrico y sostenible.

1.4. PARTES PRINCIPALES DEL TFG

El presente trabajo se halla dividido en los siguientes cinco capítulos:

- **Introducción:** Muestra la razón de ser de este trabajo. Se especifican los objetivos con los que se crea este trabajo y la metodología empleada en su elaboración.
- **Historia del desarrollo de los algoritmos de IA:** Proporciona un contexto técnico al lector sobre los algoritmos, y se aborda su historia, desde que apareciera el primer algoritmo a mediados del siglo XX, hasta la actualidad, con los algoritmos de IA Generativa.
- **Impacto de los algoritmos de Inteligencia Artificial en la actualidad:** Se analiza el impacto social de los algoritmos de Deep Learning y se detalla en profundidad cuáles son los problemas que traen consigo los algoritmos

generativos. En el mismo capítulo, se debaten las distintas propuestas que han publicado recientemente diversas instituciones, tanto en el ámbito de las políticas públicas como en el ámbito jurídico, que buscarían afrontar los dilemas éticos de los algoritmos.

- **Ética aplicada como solución a los problemas de los algoritmos:** En este capítulo se exploran distintas teorías éticas recientes sobre la tecnología que pudieran ser aplicadas a un marco ético aplicable para el desarrollo y uso de futuros algoritmos de IA.
- **Conclusiones:** Se trata del **capítulo final** del trabajo. Se analiza cómo se han cumplido los objetivos del trabajo y se abordan futuras líneas de trabajo a partir de las cuáles se pudiera seguir construyendo.
- Finalmente, se muestra la **bibliografía** del trabajo, en la cual aparecen las fuentes revisadas que han dado forma a este proyecto.

2. HISTORIA DEL DESARROLLO DE LOS ALGORITMOS EN LA IA

En este capítulo, se recalca la importancia de conocer algunos de los algoritmos matemáticos que modelan la tecnología que existe detrás de las inteligencias artificiales de aprendizaje profundo (Deep Learning). Esto se debe a que gran parte de la bibliografía existente que aborda la ética en los algoritmos de IA, carece de los fundamentos técnicos necesarios que puedan explicar su funcionamiento, y viceversa, los artículos científicos que tratan los avances de los algoritmos evitan abordar estos avances desde una perspectiva ética.

Un acercamiento técnico pudiera servir para mostrar, desde otra perspectiva, por qué una falta de transparencia en la tecnología puede ser problemática. La transparencia, o la explicabilidad en el funcionamiento de los algoritmos, es algo crucial en la búsqueda de un planteamiento ético aplicable, puesto que permite tener la capacidad de encauzar estos sistemas a obtener resultados controlados que aboguen por generar soluciones éticas a los problemas que se le planteen. Como se verá en el capítulo, los algoritmos de IA, a lo largo de los años, han perdido su transparencia inicial con el fin de obtener mejores resultados, ser más rápidos y poder producir mayores beneficios económicos.

Por ello, se tratará en estos puntos de explicar la historia y desarrollo de la inteligencia artificial (IA), a partir de los algoritmos que rigen su funcionamiento, desde sus etapas más primitivas, hasta su estado actual. Con ello, se busca desmitificar la tecnología y dejar atrás la noción errónea de que un algoritmo de IA es tan solo una caja negra, de funcionamiento inexplicable, capaz de resolver problemas sin la supervisión humana. De esta manera, se busca exigir la responsabilidad tanto sobre su creación como sobre su uso.

Además, así, podrá aportarse un punto de vista ético, correctamente fundamentado en la tecnología existente, para así poder dar directrices basadas en la ética para futuras implementaciones de algoritmos. También, se busca que esta explicación sirva como punto de partida para filósofos y expertos en ética, para que así puedan estar involucrados en la creación de algoritmos desde el minuto cero. Después, podrá cuestionarse también, si es adecuado para el desarrollo humano dejar parte de la autonomía de la especie en manos de estos algoritmos, y si lo es, en qué condiciones deben someterse.

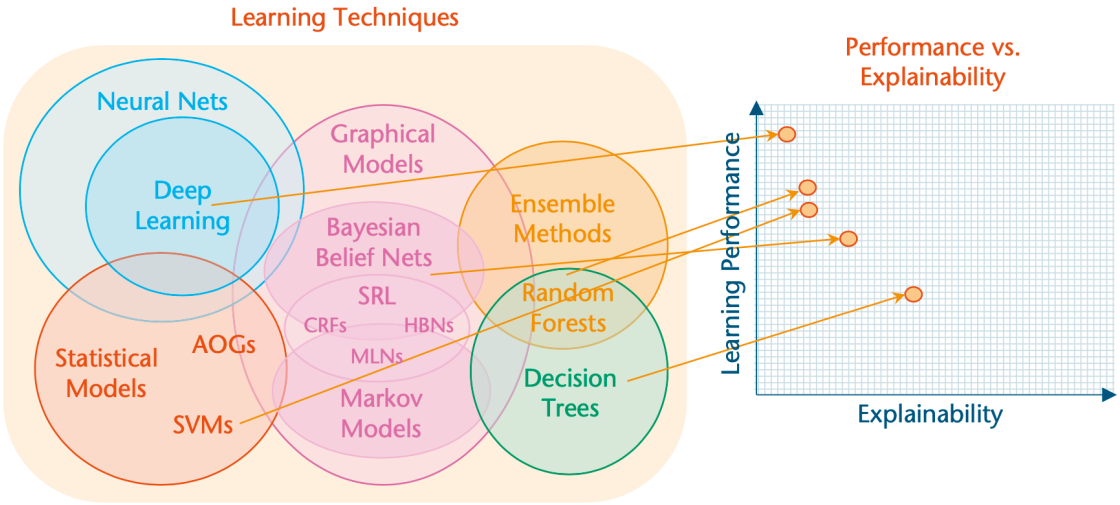

Figura 1. Relación de rendimiento y explicabilidad en técnicas de aprendizaje de algoritmos de Machine Learning. Recuperado de "Darpa's Explainable Artificial Intelligence Program", por Gunning, D., & Aha, D. (2019). AI magazine, 40(2), 44-58.

Para justificar la elección de qué algoritmos se profundizarán en el trabajo, se muestra la *Figura 1*. En esta figura puede comprobarse la existencia de una relación entre explicabilidad y precisión en los modelos algorítmicos. Aquí, se muestran distintos tipos de algoritmos aplicados a la Inteligencia Artificial, los cuales se listan a continuación por orden de menor a mayor rendimiento: árboles de decisión, redes bayesianas, modelos estadísticos, el algoritmo de randomForest y, por último, las redes de Deep Learning. A menudo, los modelos con mejor rendimiento son los menos explicables (algoritmos de Deep Learning), y los más explicables, son los menos precisos (árboles de decisión) (Gunning, D., & Aha, D., 2019).

Debido a las restricciones de extensión que posee este proyecto al tratarse de un Trabajo de Fin de Grado, se delimitará su alcance exclusivamente a los algoritmos de Deep Learning. Esta elección se fundamenta, por una parte, en la **baja explicabilidad** que poseen estos algoritmos, y por otra, **en la creciente presencia** de éstos en la sociedad debido a su buen funcionamiento. Dentro del campo del Deep Learning, se dedicará una sección para explorar los algoritmos de IA Generativa, ya que su naturaleza creativa impone nuevos dilemas éticos que debieran tratarse. Para comprender qué son los algoritmos de Deep Learning, antes se realizará un repaso por la historia de los algoritmos que han contribuido a su desarrollo en la actualidad.

Previo a abordar la historia del desarrollo de estos algoritmos, se realiza un repaso por la definición de "algoritmo". Esto permitirá sentar las bases conceptuales antes de adentrarnos en la evolución de estos algoritmos.

2.1. LA RELACIÓN ALGORITMO-IA

Previo a explicar qué algoritmos se encuentran detrás de las inteligencias artificiales, y con el fin de que sea comprensible para cualquier usuario conocer la tecnología subyacente, se debe comenzar desde lo básico. Es por ello por lo que en las siguientes líneas se definirá qué se entiende por Inteligencia Artificial dentro del contexto de la investigación de este trabajo.

2.1.1. Definición de IA

La definición de qué es la IA es difusa, y puede variar según el investigador al que se le pregunte, pues dependerá en parte del propósito con el que se cree. Decir que se espera que sea un sistema inteligente es una definición vaga, pues cada persona tiene su propia definición de inteligencia, y muchas veces, los sistemas autónomos movidos por IA no parecen poseer inteligencia, o no una inteligencia que se asemeje a la humana. Roger S. Schank (1987), en su paper, "*What is AI, anyway?",* se encontraría con este problema de definición. Parte de la base de que todo investigador de IA busca crear, como primer objetivo, una máquina inteligente, y, en una segunda instancia, descubrir qué es lo que origina la inteligencia. Finalmente, concluye que el objetivo de estas máquinas inteligentes es resolver problemas, por lo que la inteligencia que se espera de estos sistemas se resume en su capacidad para adaptarse a resolver un problema, aprender de experiencias previas y de actuar en consecuencia, mejorando su rendimiento de forma autónoma. Concluye que la IA es una tecnología que debería aplicarse en todos los campos de conocimiento y no solo en computación. Debería implantarse en ámbitos como la antropología, la medicina, o la política, pues es una tecnología que puede que tenga más que aportar en dichos campos que en la propia computación (Schank, R. S., 1987).

Schank concluye su análisis prediciendo un futuro en el que la IA pudiera ser aplicada a todas estas ciencias, una visión que empieza a materializarse en la actualidad. No obstante, tendrían que pasar varios años para que su predicción se hiciera realidad.

2.1.1.1. Motivos del avance de la IA

Esta tecnología ha avanzado por la historia a pasos cortos, con grandes pausas entre medio, en los comúnmente conocidos como "inviernos de la IA", periodos en los que se hacía patente la desconfianza a que esta tecnología tuviera utilidad o que simplemente fuese posible de desarrollar. Sin embargo, si la IA ha llegado a nuestros días, y no ha sido hasta ahora cuando estamos observando un "boom" alrededor de ésta, se debe a los siguientes factores:

- En primer lugar, debido a lo explicado en la aún vigente **Ley de Moore**, que propone que cada dos años el número de transistores en un procesador se duplica con respecto al anterior, según las estimaciones de Gordon Moore (1975), cofundador de Intel. Esto implica que el poder de computación actual supera significativamente al que hubiera en la aparición de los primeros algoritmos de aprendizaje automático. Además, las tecnologías Cloud ofrecen al usuario la capacidad computacional de un centro de datos en un teléfono móvil usando una conexión a Internet. Estas tecnologías aprovechan las capacidades de la computación remota, escalable tanto vertical como horizontalmente, reduciendo en gran medida los costes para el usuario.
- En segundo lugar, la **proliferación de una sociedad de la información**, o sociedad del dato, en la que cualquier movimiento, transacción, o incluso 'like' a una imagen por parte de un usuario, es transformado en información, susceptible de ser utilizada para entrenar a modelos de IA.
- Y, por último, y tal como interesa investigar, **debido a los avances en los algoritmos** (Ergen, M., 2019). Muchos algoritmos de IA están basados en modelos estadísticos que datan del siglo XIX, ya que, como se definió previamente, un algoritmo no tiene por qué estar vinculado a software para existir.

De hecho, el algoritmo que regiría el comportamiento de la primera IA aparecería definido, unos años antes de su creación, en forma matemática. Sin embargo, no sería hasta 1956, con la creación del modelo del perceptrón a manos del profesor Frank Rosenblatt, que la IA comenzaría a dar sus primeros pasos en el mundo de la computación.

Entrando en la otra parte de esta relación, se verá que la definición de qué es un algoritmo es igualmente ambigua.

2.1.2. Definición de algoritmo

La definición de algoritmo es amplia, y puede variar en función del campo de la ciencia que lo tenga como objeto de estudio. Un acercamiento puede encontrarse en la definición que aparece en el libro *Introducción a los Algoritmos*, en el cual, se define el concepto de algoritmo de la siguiente forma: "Un algoritmo es cualquier procedimiento matemático bien definido que toma un valor o conjunto de valores como entrada, y produce un valor o conjunto de valores como salida. Un algoritmo es pues, una secuencia de pasos computacionales que transforman a la entrada en la salida" (Cormen, T. H., et al., 2022).

Por el contrario, esta definición resulta muy restrictiva en comparación con cómo se utiliza el concepto de algoritmo en la literatura de la ética y gobernanza de la IA (Larsson, S., & Heintz, F., 2020). Como ejemplo, Larsson y Heintz se refieren a un artículo sobre la transparencia de los algoritmos, el cual lista siete puntos que debieran abordarse. De ellos, solo uno de esos puntos se enfoca específicamente en los algoritmos tal y como se ha definido previamente, mientras el resto se ocupan de cuestiones relacionadas con los datos, objetivos, resultados, cumplimiento, influencia y uso.

Aunque no forme parte de la definición inicial, es crucial mencionar la relevancia del dato, el cual es uno de los actores principales en el diseño de los algoritmos. Éste es información digitalizada que puede utilizarse para entrenar al algoritmo, cuyo rendimiento dependerá directamente de la calidad de la información con la que ha sido entrenado. Es común que aquellos que diseñan algoritmos pueden tomar decisiones arbitrarias en su diseño, y concretamente en la cadena por la cual se procesa el dato, ya que los algoritmos reflejan los sesgos de sus creadores, pueden reforzar pensamientos sobre la sociedad, y favorecer a unos frente a otros (Kemper, J., & Kolkman, D., 2019). Por consiguiente, debe ponerse énfasis en que la ética regule el uso de la información en la tecnología. Este concepto será recurrente en los puntos siguientes.

Por tanto, para la línea de investigación de este trabajo, cuando se haga referencia al término algoritmo, se incluirán todos los factores que influyen en su diseño.

En el siguiente punto se ahondará en cómo han avanzado estos algoritmos hasta llegar al estado en el que nos encontramos actualmente.

2.2. EVOLUCIÓN DE LOS ALGORITMOS DE INTELIGENCIA ARTIFICIAL

2.2.1. El perceptrón, el primer algoritmo de la IA (Década de 1950)

Podría decirse que el nacimiento de la Inteligencia Artificial moderna tuvo lugar con la invención del modelo del perceptrón, el primer algoritmo aplicado a la inteligencia artificial, el cual sería precursor y sentaría las bases de lo que es la IA actual tal y como la conocemos. Sin embargo, el perceptrón nacería basándose en un algoritmo que se crearía en 1943 y cuya finalidad no estaría orientada a originar el campo de la IA. Ni siquiera era un algoritmo computacional, sino más bien, un modelo matemático.

Este modelo estaría fundamentado por los hallazgos del psicólogo Warren McCulloch y el filósofo Walter Pitts. McCulloch, influenciado por las teorías de la cibernética que encontrarían su auge a principios del siglo pasado, junto a los Principia Mathematica, escritos por los filósofos Bertrand Russel y Alfred Whitehead, estaría convencido de que el cerebro humano, como cualquier otro sistema dependiente de la lógica en el universo, debía ser susceptible de poder ser traducido a un modelo matemático (Galaviz, J., 2016).

Para ello, basa su modelo en la estructura biológica de una neurona corriente. La neurona humana está compuesta por un cuerpo celular y conexiones nerviosas; estas conexiones se dividen en dos tipos: las dendritas, las cuales reciben los impulsos nerviosos, y el axón, que se encarga de enviar respuestas a partir de la información de las dendritas mediante impulsos a otras células.

En el caso de la neurona de McCulloch, puede traducirse dendritas por inputs y axón por output, y el mecanismo, sin tener en cuenta los procesos químicos de la neurona, sería prácticamente el mismo. Esta neurona ficticia recibiría una serie de inputs binarios (1/0), realizaría una suma ponderada de estos, en función de unos pesos asignados a cada input, es decir, importancia, y se estimaría así un resultado que llegaría al output, o el axón de esta neurona (McCulloch, W. S. y W. Pitts, 1943).

En base al hallazgo de esta neurona artificial, el profesor Rosenblatt, unos años más tarde, desarrollaría el algoritmo del perceptrón y lo publicaría en 1958 (Rosenblatt, 1958). Este sistema, que es considerado como una versión mejorada de la neurona de McCulloch, es capaz de aprender, como lo haría una neurona real. Esto lo conseguiría realizando un reajuste automático de los pesos que se asignarían a cada input de la neurona. Puede verse el esquema matemático de como funcionaría este modelo en la Ilustración 2. Aunque este mecanismo pueda parecer simple, es el modelo en el que se fundamentan en la actualidad los algoritmos de redes neuronales de deep learning, e inteligencias artificiales generativas, solo que en su estado más fundamental.

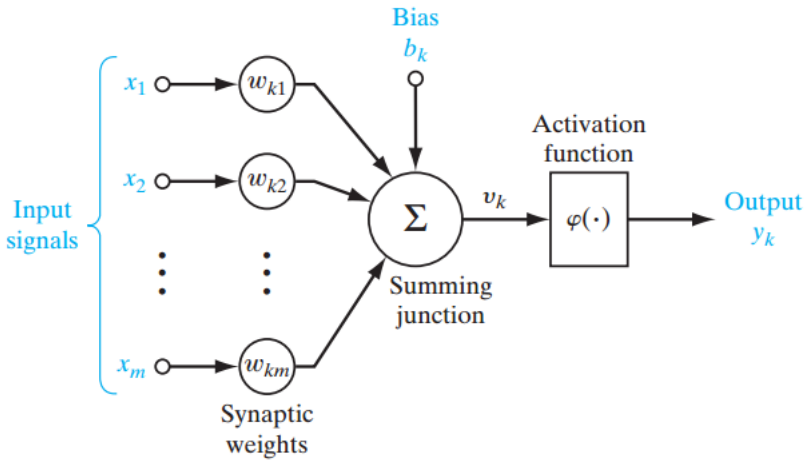

Figura 2. Esquema de un modelo matemático de perceptrón. Recuperado de "Predicción de agentes patógenos en plantas ornamentales utilizando redes neuronales", por Escobar, Emmanuel & García-Díaz, Noel & Verduzco, Jesus & Andrade, Juan. (2017).

El hallazgo de Rosenblatt resultó finalmente en la creación del Mark I Perceptron, en 1960, el primer ordenador que realizara operaciones a base de prueba y error (Jay J.C. et al., 1960), es decir, siendo capaz de hacer correcciones en base a sus propios outputs, como haría un ser humano, llevando a la práctica el modelo matemático del perceptrón de Rosenblatt. El objetivo del ordenador sería el de clasificar imágenes en dos categorías. Este fue el primer computador capaz de aprender, y, por tanto, aunque de una forma muy limitada, simular el funcionamiento de las neuronas del ser humano.

Sin embargo, esta hazaña quedaría en un principio en anécdota tecnológica, ya que, el libro publicado en 1969, "Perceptrons" por Marvin Minsky, compañero de universidad de Rosenblatt, en el que se estudiaba el perceptrón, mostró algunas limitaciones del algoritmo, siendo una de ellas, y tal vez la más grave, la incapacidad de este de realizar operaciones XOR, una operación lógica básica de clasificación en el

mundo de la electrónica digital (Minsky, 1969). Estas limitaciones acabarían por generar una visión pesimista para el futuro de la IA y una pérdida de interés en la investigación del campo de los algoritmos de aprendizaje automático. No sería hasta la década de los 80 que, gracias a los avances computacionales y en especial, de un nuevo algoritmo, que se empezaría a vislumbrar un futuro más optimista para el campo de las IAs.

2.2.2. La complejidad añadida en los perceptrones multicapa (1980)

En la década de 1980, la IA volvió a propiciar un alto interés en la comunidad científica, alentado por varios avances en el desarrollo de algoritmos y en el poder computacional de los ordenadores de la época, que seguirían cumpliendo la Ley de Moore.

Uno de los motivos por los que es interesante incidir en los avances de esta época, es debido a que fue, en 1986, gracias a los estudios de Geoffrey Hinton y su equipo, que se desarrollara el algoritmo de aprendizaje automático que daría forma a la complejidad de los algoritmos de Deep learning que existen hoy día.

En 1986, Geoffrey Hinton, junto a sus compañeros de la Universidad de California, David Rumelhart y Ronald Williams, publicaría un artículo que revolucionaría la IA, un documento en el que se detalla por primera vez el funcionamiento del algoritmo de propagación trasera, o backpropagation (Hinton et al., 1986). Este algoritmo puede considerarse una versión mucho más avanzada del modelo de perceptrón, tal como aparece definido en las primeras líneas de su trabajo.

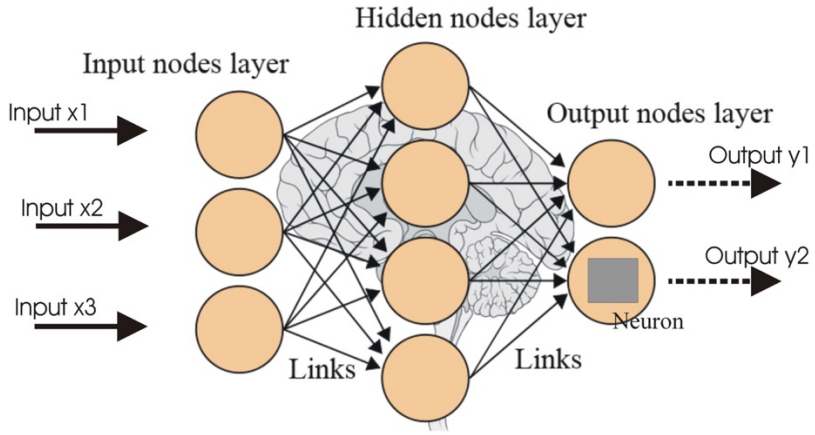

Figura 3. Esquema de un algoritmo de propagación trasera. Recuperado de https://github.com/XavierCarrera/Preguntas-Frecuentes-Data-Science-Machine-Learning

El algoritmo de propagación trasera, al igual que el perceptrón original, corrige los pesos asignados a los distintos inputs a partir del resultado en la salida. La gran diferencia con el modelo previo, como puede verse en la Ilustración 3, es que en esta aparece una capa oculta, una gran red de neuronas, con sus inputs y sus outputs, que dan resultados distintos que se propagan hasta el final, que ajustan sus pesos a partir de los resultados a la salida, y que funcionan en sintonía como un organismo vivo.

Este procedimiento recuerda, guardando las distancias, al proceso sináptico de una red de neuronas en el cerebro, y es aquí donde pudiera radicar una de las grandes preocupaciones de la ciencia ficción en su momento. Suponiendo que se pudiera desarrollar una red neuronal a una escala de grandes dimensiones, ¿podría llamarse al proceso algorítmico que realizara esta red, un pensamiento? En este punto de la historia, ya varios filósofos comenzarían a entrever que la tecnología que existiría detrás de las IAs podría suponer un problema al largo plazo, en especial, cuando al aprendizaje de los algoritmos se le comparaba al aprendizaje humano con términos como ‘red neuronal’. Uno de los primeros filósofos en abordar la problemática tecnólogica, y en especial, de la autonomía de los algoritmos de IA, sería John Searle.

2.2.3. John Searle: El debate entre los tipos de IA según la funcionalidad del algoritmo (1986)

En esta historia del avance de los algoritmos de las inteligencias artificiales a lo largo de la historia hasta nuestros días, llama la atención el origen de éstas en la

neurociencia. Tanto McCulloch, como Rosenblatt son profesionales del área de la neurociencia, y sus aportaciones científicas han sido las que han moldeado las bases de las inteligencias artificiales.

Por lo tanto, al hablar de la historia de los algoritmos de IA, tratar tan solo de lo que supone ésta a nivel tecnológico sería limitar el enfoque del trabajo. Es interesante por tanto conocer las aportaciones de distintos científicos de distintas áreas de conocimiento a lo largo de la historia, que han moldeado el debate de los algoritmos hasta nuestros días.

Teniendo en cuenta que, si el perceptrón es, en un principio, el modelo matemático del funcionamiento de una neurona humana, un perceptrón multicapa, en sus versiones más avanzadas, supondría la imitación de una red de neuronas (de ahí el nombre de red neuronal). Es por ello por lo que, podría extrapolarse que podemos obtener máquinas con el raciocinio humano necesario como para tomar decisiones humanas, definir qué es lo correcto, o influir en el propio desarrollo de la humanidad. Incluso podría aventurarse uno a decir que, con la suficiente potencia computacional, y teniendo en cuenta que un sistema informático puede, en principio, ser escalable hasta el infinito, un algoritmo de aprendizaje profundo actual podría llegar a tener una capacidad mayor de toma de decisión que un ser humano. Y, si ese fuera el caso, ¿seríamos realmente prescindibles en situaciones de toma de decisión de alta carga moral? Antes de entrar en materia, es importante introducir el concepto de la Chinese Room, ideado por John Searle, y que introduce el dilema de la existencia de la conciencia artificial.

2.2.3.1. IA Fuerte

John Searle es un filósofo estadounidense, el cual es conocido por realizar una de las primeras críticas a la inteligencia artificial. En la habitación china (Searle, 1980), Searle expone el siguiente escenario:

Hay una persona encerrada en una habitación. Por una pared, recibe mensajes escritos en chino. A pesar de no conocer el idioma, tiene instrucciones de traducirlos al inglés (lengua materna de Searle). Para ello, en la habitación hay un conjunto de reglas en su idioma que explican como convertir los distintos caracteres chinos al idioma inglés, siguiendo una serie de pasos, como una especie de algoritmo. Finalmente, dichos mensajes son enviados fuera de la habitación traducidos correctamente al inglés. Desde

fuera, pareciera que quien estuviera dentro de la habitación debiera entender el idioma chino para traducirlo correctamente. Sin embargo, como ya sabemos, la persona que está dentro solo ha seguido los pasos que le han sido dados, sin realizar cualquier otro juicio acerca de su situación.

Searle extrapola este ejemplo a los algoritmos de inteligencia artificial, y al porqué cree que carecen de conciencia. Aunque los resultados obtenidos en la actualidad sean impresionantes y parecieran imitar al ser humano, lo cierto es que estamos tratando de algoritmos sin capacidad de autodeterminación, y, por tanto, de razón, al carecer de autonomía.

Searle llama a estas IAs con supuesta capacidad de conciencia y autodeterminación con el término de IA fuerte, en contraposición a la IA débil, de la que se tratará más adelante.

Por otro lado, el argumento de Searle ha sido criticado bastante en la comunidad científica, en especial, destacan los argumentos de Daniel Dennett, aclamado filósofo especializado en la conciencia humana, que, además, critica el argumento de Searle de forma sarcástica. Dennett critica al argumento de Searle por ser demasiado simplista. Considera que la conclusión a la que llega Searle es un engaño, fruto de un sistema simple, que poco tiene que ver con la realidad de la IA, y guiado artificialmente a un final del que Searle quiere convencernos, el cual considera fácilmente rebatible (Dennett, 1995). Como se sabe, en el experimento de la Habitación China hay un hombre encerrado, que poco entiende de lo que está haciendo, pero lo hace. Como símil, es el equivalente a decir que ocurre lo mismo en los algoritmos de IA. En la IA, hay muchos subsistemas (como se ha visto en las redes multicapa), incapaces de producir entendimiento por sí mismos, pero, a la salida del programa, si esta es lo suficientemente elaborada, el conjunto puede ser capaz de imitar el comportamiento humano. A partir de este punto, Dennett compara el caso que nos ocupa con la teoría materialista. Si la conciencia del ser humano es fruto del cerebro, y este está compuesto de distintos subsistemas que la producen, y que sabemos que por separado son incapaces de producir entendimiento, no puede negarse entonces que ocurra lo mismo en un sistema artificial, o al menos, sería contradictorio. Termina Dennett su crítica indicando que, para que la teoría de Searle tenga sentido, la complejidad del sistema es muy importante, y no puede dejarse fuera de la ecuación. Teniendo en cuenta

ahora la importancia de la complejidad del sistema queda por ver qué es capaz de deparar el futuro, con unas redes neuronales que, junto a las especificaciones de los nuevos sistemas de computación que mejoran año tras año, no hacen más que hacerse más grandes y complejas en lo que a subsistemas se refiere.

2.2.3.2. IA Débil

Por otro lado, se define la categoría de IA débil (Searle, 1990) como la inteligencia artificial dedicada a solucionar problemas complejos, sin la posibilidad de tener conciencia o raciocinio sobre sus propias acciones.

Para Searle, la IA débil no supone un problema tan urgente como lo es la IA fuerte. Sin embargo, en la actualidad, los algoritmos que entran dentro de la categoría de IA Débil, como se verá en el capítulo tres, han sido ya, foco de debates sobre la moralidad y legitimidad de la IA.

2.2.4. Los nuevos algoritmos y la llegada del Deep Learning (1990s-actualidad)

A lo largo de los años siguientes, a partir de la invención del algoritmo de propagación trasera, la IA seguiría avanzando en su desarrollo hasta nuestros días. Sucedería antes un segundo invierno de la IA entre los años 1987 y 1993 debido a la falta de interés y financiación (Abeliuk, A., & Gutiérrez, C., 2021). Sin embargo, un hito de la técnica haría que la IA volviese a estar en el punto de mira, tanto por su potencial y sus capacidades como por los temores (a veces infundados) de lo que podría suponer para la humanidad que una computadora superara al ser humano en intelecto. Este punto en la historia se daría gracias a la computadora DeepBlue.

2.2.4.1. DeepBlue

DeepBlue es una computadora que fue capaz, en 1997, mediante el uso de algoritmos de inteligencia artificial y, de forma autónoma, de ganar al entonces campeón mundial de ajedrez Garry Kasparov (Campbell, M. et al., 2002).

La primera versión del ordenador, el DeepBlue I, perdería en 1996 contra el campeón de ajedrez. Sin embargo, la versión mejorada de este, en mayo de 1997,

conseguiría ganarle. La diferencia entre el nuevo modelo y el anterior estuvo simplemente en el hardware. DeepBlue II contaría con un nuevo chip, que supondría un incremento, de 6400 a 8000 funciones, además de otras mejoras de velocidad y de optimización para lograr su tarea. De nuevo, este hecho verifica la hipótesis inicial del capítulo, en la que el avance de la IA ocurre debido al progreso de la tecnología computacional, reflejada a su vez en la Ley de Moore.

DeepBlue es claramente un ejemplo que formaría parte de lo que se definió en el campo de la ética como **IA débil**, una máquina de la que no se espera raciocinio propio, pero que sí sea capaz de demostrar conocimiento para el propósito con el que ha sido construido. Sin embargo, esta máquina ya presentaría preocupación en su tiempo, que, sumada a la aún acertada ley de Moore, daba pie a formular los siguientes interrogantes: ¿Hasta dónde sería capaz de llegar esta tecnología? ¿Podría llegar a suponer un problema al largo plazo?

2.2.4.2. Aprendizaje profundo (Deep Learning)

No sería hasta la década de 2010 que se introduciría la idea del aprendizaje profundo, o Deep Learning (Abeliuk, A., & Gutiérrez, C., 2021). Gracias al avance de la computación en nuestros días, nacerían nuevos proyectos de algoritmos con objetivos más ambiciosos. Entre ellos destaca AlphaGo, en 2016, un superordenador capaz de ganar al entonces campeón mundial de Go, Lee Sedol. Esta anécdota resulta similar a lo sucedido en 1997 con DeepBlue, la diferencia entre ambas estaría en que el Go es un juego en el que es posible realizar hasta 300 movimientos en cada turno, mientras que son 30 en el caso del ajedrez. Esta mayor complejidad implica una demanda computacional mucho mayor para realizar cálculos (Silver, D., et al., 2016), algo inalcanzable en la era del Deep Blue. Es entonces en esta época, cuando nace el Deep Learning, modelos neuronales, tan grandes y complejos que resultaría imposible para un ser humano comprender o controlar su funcionamiento al nivel fundamental.

Los algoritmos de **Deep Learning** son modelos compuestos por múltiples capas de procesamiento, que les permiten aprender representaciones de datos con múltiples capas de abstracción. La aplicación de métodos como la propagación trasera adaptados a redes con grandes volúmenes de datos, ha mejorado de forma dramática el estado del arte

en aplicaciones de reconocimiento facial, de voz, reconocimiento de objetos, y su implementación se ha extendido a campos como la génetica y la medicina (LeCun, et al., 2015).

Por otro lado, a partir del Deep Learning, nacería la rama de algoritmos de IA Generativa, los cuales, debido a su naturaleza creativa, distinta a la de los algoritmos que habrían aparecido previamente, y a su reciente impacto en la actualidad, merecen ser reconocidos en el siguiente punto.

2.2.5. Algoritmos de IA Generativa

La IA Generativa aparece definida en "Generative Deep Learning" como una rama del aprendizaje automático que implica entrenar a un modelo para producir nuevos datos que sean parecidos a los de un dataset (Foster, D., 2022). Esta información generada, será información nueva, creada al momento por el algoritmo a partir de los datos con los que haya sido entrenado previamente, es decir, hablamos de una inteligencia creativa, no limitada por la realización de tareas automatizadas.

Visto desde el punto de vista de los algoritmos, las IAs generativas utilizan redes neuronales de Deep Learning, profundas y complejas, de miles de parámetros y capas ocultas, pero en su ejecución, se diferencian en cierta parte de una red neuronal común. El secreto del buen funcionamiento de estas IAs se basa en el uso de Redes Generativas Adversariales (Goodfellow, I., et al., 2014). En este modelo, se emplean dos redes neuronales, una red generativa, creadora de contenido, y, por otro lado, una red adversaria, encargada de discriminar el contenido que la red generativa pueda crear y que resulte falso, inverosímil, o dañino para el usuario. Llama la atención que, en este caso, sea un algoritmo el que regule el funcionamiento de otro algoritmo y no el ser humano, relegando su propia autonomía y responsabilidad a la tecnología, en pro de reducir tiempos y carga de trabajo. Ante este caso, cabe preguntarse, ¿podemos estar seguros de que la red discriminatoria sea capaz de hacer bien su trabajo? ¿Qué es aquello que debiera discriminar la IA para asegurar un correcto funcionamiento? ¿Cómo sabe qué debe discriminar?

Sin embargo, suponiendo que la red discriminatoria hiciera bien su trabajo, introducir una red neuronal que discrimine lo que haga su red hermana en paralelo, sin

darle al usuario ningún recurso visual que le muestre lo que está sucediendo, es de nuevo, un ejemplo de la falta de explicabilidad en los algoritmos de IA actuales.

Como ejemplo de un algoritmo de IA Generativa bien documentado, se hablará a continuación del algoritmo GPT, el cual es usado en numerosas aplicaciones de IA Generativa. Se mostrará su potencial, y se indicarán cuáles son algunos de los problemas que podemos encontrarnos en él, y por qué se cree necesario hablar de ética en la aplicación de los algoritmos.

2.2.5.1. El algoritmo GPT

GPT (Generative Pre-trained transformer) es el algoritmo de la inteligencia artificial generativa que ha revolucionado al mundo el último año. El modelo que presenta está basado en los algoritmos de Deep learning vistos anteriormente. Con miles de millones de parámetros siendo corregidos al instante, la explicabilidad del funcionamiento y de la toma de decisiones de este algoritmo se vuelve algo imposible, sin embargo, el buen rendimiento y el potencial de esta tecnología parece haberle restado importancia al problema. El algoritmo en este caso no sería únicamente aplicable a, por ejemplo, la clasificación de imágenes, como sucediera con el ordenador del Perceptrón o con otros algoritmos con un propósito claro. En este caso, la IA se convierte en un modelo de propósito general, en el sentido en el que es capaz de generar cualquier tipo de información.

La función de GPT, y por la que más se conoce, es su integración en ChatGPT. ChatGPT es un chatbot construido sobre la base de GPT, que utiliza un modelo de lenguaje natural para entender las peticiones a partir del lenguaje humano, y responder de forma acorde a cualquier consulta del usuario (Lock S., 2023). Recuerda esto al ejemplo de la Habitación China, pero a una escala mucho mayor, de naturaleza generalista, en el que el hombre encerrado en la habitación pasa de ser un traductor que sigue órdenes, a convertirse en el genio de la lámpara mágica.

Puede verse la capacidad multipropósito de esta tecnología en noticias recientes como que este algoritmo fuera capaz de aprobar en el presente año el Uniform Bar Exam (Arredondo, P. et al., 2023), el examen de acceso a la abogacía en EEUU.

El algoritmo ha demostrado un gran nivel de versatilidad, siendo capaz de generar imágenes y vídeo, como lo demuestra su integración con Dall-E, un modelo de difusión. Los sistemas de difusión son modelos de IA que tienen como propósito la creación de imágenes a partir de las peticiones de un usuario. Sin embargo, a pesar de lo interesante que resulta la propuesta, surgen nuevas dudas: ¿existe un control en qué imágenes es capaz de generar el algoritmo? ¿Puede éste censurarse a sí mismo? ¿En base a qué genera las imágenes el algoritmo? Además de estas cuestiones, GPT parece no haber solucionado aún el problema de los sesgos (Luccioni, A. S, et al., 2023), problema que se le atribuye a los algoritmos cuando los datos con los que se entrenan no se encuentran normalizados, y, por lo tanto, producen resultados sesgados.

Una vez comprendida cuál es la naturaleza de los algoritmos, qué son en realidad, y cuál es el estado de la técnica hasta los algoritmos de Deep learning e IA Generativa, puede comenzar a analizarse el impacto social, y la problemática que pueden arraigar unos algoritmos que, por el desconocimiento, sean susceptibles de ser aplicados de forma indiscriminada y sin supervisión para resolver de forma fácil problemas, que, sin un estudio ético previo, puedan acabar perjudicando más que favoreciendo a las personas. A esto se le añade la capa de complejidad que han añadido los algoritmos de Deep Learning sobre el entendimiento de cómo funcionan éstos en su interior, y de cómo llegan a tomar sus decisiones. Al ser sistemas tan alejados de la operación humana, cabe preguntarse ¿hasta qué punto podemos confiar en las decisiones que tomen estos sistemas? (Ras, G., 2022). La historia de los algoritmos de IA demuestra que el avance de la técnica ha dejado de lado la explicabilidad de los algoritmos en pro de favorecer el rendimiento.

Se verá en el siguiente capítulo, cuál es el impacto que producen los algoritmos de Deep Learning en la actualidad, tanto beneficios como puntos negativos de éstos. Además, se dedicará un punto en especial a los algoritmos generativos, pues han originado nuevos problemas recientemente. Tras esto, se verán las propuestas que ofrecen distintas instituciones en lo referente a mitigar los daños que pudieran producir los algoritmos.

3. IMPACTO DE LOS ALGORITMOS DE INTELIGENCIA ARTIFICIAL EN LA ACTUALIDAD

Ya conociendo cuál el origen de los algoritmos que propulsan las inteligencias artificiales de hoy día puede comenzar a plantearse cuál es el impacto que estos tienen en la vida humana. De esa forma, se verá qué beneficios pueden estos aportar a la humanidad, y a su vez, qué problemáticas plantean en ésta y en su desarrollo de cara a futuro.

Para ello, se inicia el análisis con un enfoque en el estado actual de los algoritmos de IA, comenzando con las bondades de los algoritmos de DeepLearning, lo que implica su uso en los distintos campos de la ciencia y los nuevos dilemas que éstos acarrean. Posteriormente, debido a cuán diferente son los dilemas de IA Generativa, y a la relevancia de éstos, se les dedica a las problemáticas de esta tecnología un apartado. Como se verá, su falta de transparencia y su capacidad de generar contenido nocivo, ha sido origen de debates y controversias alrededor de su concepción y su impacto.

Por último, se detallan las propuestas que existen hasta ahora por parte de las instituciones para afrontar estos dilemas, y se analiza la necesidad de trabajar en una ética de la tecnología que pueda ser aplicada en el desarrollo de futuros algoritmos.

3.1. BENEFICIOS Y DILEMAS ÉTICOS DE LA APLICACIÓN DE ALGORITMOS DE DEEP LEARNING

Los algoritmos de Deep learning, en la actualidad, comienzan a ser capaces de realizar nuevas tareas que en el pasado solo el ser humano habría tenido la capacidad de hacer. Este progreso sugiere la posibilidad de que estos algoritmos sean capaces de aprender distintas disciplinas y puedan ser aplicados en diversos campos de la ciencia, como anticipara Schank en 1987. Sin embargo, a pesar de que gracias a ellos es posible conseguir una mayor automatización del trabajo, e incluso propulsar los avances en investigación y desarrollo de la ciencia, estos algoritmos podrían propagar sus problemas a otros campos en los que, por su naturaleza, en la que deben abordar temas de mayor delicadeza, el impacto producido pudiera ser mucho más grave. Si esto es así, ¿cuál es el interés de aplicar estos algoritmos?

Para responder a esta cuestión, se buscará mostrar en este punto el beneficio que pueden producir estos algoritmos aplicados a distintos ámbitos de desarrollo humano, como la educación o la medicina. No obstante, es necesario tratar los problemas intrínsecos que poseen los algoritmos, y cómo estos pueden ser peligrosamente propagados a las disciplinas previamente mencionadas.

3.1.1. Algoritmos aplicados a otras ciencias

Existen varias razones por las cuales interesa el desarrollo y la aplicación de algoritmos de IA, pues son herramientas que permiten la automatización de tareas tediosas, una mayor eficiencia y velocidad en realizar trabajos. A continuación, se enumeran algunos de los beneficios que esta tecnología es capaz de aportar, al ser integrados como herramientas.

3.1.1.1. Asistentes virtuales

Los asistentes virtuales son aplicaciones capaces de entender instrucciones de voz y de realizar acciones para el usuario (Moshayedi, A.J et al., 2022). Entre estos asistentes, los más conocidos son Alexa de Amazon, Siri de Apple o Cortana de Microsoft. Su funcionalidad puede ir desde resolver tareas simples como poner recordatorios en el calendario, a hacer funciones que controlen elementos del hogar como las luces, la temperatura u electrodomésticos.

Como se verá a continuación los algoritmos de Deep Learning también pueden ser aplicados a distintas industrias para hacerlas prosperar y ayudarlas a crear mejores productos para sus usuarios.

3.1.1.2. Industria del automóvil

En la industria automovilística existen muchos casos de uso en los que se utilizan técnicas de predicción algorítmica. A continuación, se muestra un listado con algunos de ellos (Luckow A. et al., 2016):

- **Conducción autónoma:** Es necesario aplicar técnicas de DeepLearning para diversos aspectos de la conducción autónoma. Por ejemplo, necesitará procesar grandes cantidades de datos que provengan de los sensores y cámaras, aprender de situaciones de conducción y del comportamiento del conductor.

- **Interfaces conversacionales:** Los vehículos pueden tener asistentes conversacionales que permiten una interacción manos libres con el coche.
- **Analítica en redes sociales:** Las aplicaciones de visión por computadora se utilizan para analizar imágenes en redes sociales de usuarios con sus vehículos con el fin de recabar información para hacer análisis de datos.

3.1.1.3. Eficiencia en la atención médica

La medicina es uno de los ámbitos en los que más se están aplicando algoritmos de IA. Comenzando con la atención médica, y teniendo en cuenta que se ha cubierto previamente el campo de los asistentes virtuales es necesario añadir que existen asistentes hechos con algoritmos, especializados en el campo de la salud. Uno de ellos es Woebot, una IA conversacional que ayuda a sus pacientes a tratar la ansiedad y la depresión (Pataranutaporn, P., et al., 2021). No obstante, existen muchas más utilidades de los algoritmos de Deep learning en el campo de la medicina se listan a continuación (Moshayedi, A.J et al., 2022):

- **Generación de imágenes médicas:** Pueden utilizarse los algoritmos en imágenes de resonancia magnética para detectar signos de Alzheimer.
- **Procesado de datos médicos:** El algoritmo DeepCare, de medicina predictiva. Este algoritmo lee registros médicos de los pacientes, guarda su historial de enfermedades y realiza predicciones médicas futuras (Pham, T. et al., 2016).
- **Genética:** Destaca en este ámbito el algoritmo DeepBind, utilizado para predecir regiones de interacción entre proteínas y ácidos nucleicos.

Estos algoritmos, para su entrenamiento, necesitan ser alimentados con datos reales para ser capaces de llevar a cabo predicciones, por lo tanto, cabe preguntarse si estos datos se han obtenido de forma ética, es decir, no se ha violado la privacidad de los pacientes al ser utilizados.

3.1.1.4. Utilidades de los algoritmos generativos

Los algoritmos generativos, por su parte, son capaces de realizar nuevas tareas creativas que resultarían imposibles para los algoritmos de Deep Learning. Burgos, Suárez y Benzadón (2023) recopilan una serie de utilidades que ChatGPT puede ofrecer en el campo de la investigación:

- **Brainstorming de ideas:** Puede organizar y desarrollar las ideas del investigador.
- **Resumir conocimiento:** Puede utilizarse para analizar grandes volúmenes de información de distintas fuentes de información acerca de un tema.
- **Búsqueda bibliográfica:** Puede ayudar a los investigadores en la búsqueda bibliográfica de información sugiriendo artículos relevantes para la investigación
- **Análisis de datos:** Es capaz de reconocer datos que hayan sido pegados al chat y de realizar un código, en base a nuestras peticiones, que pueda utilizarse en softwares específicos de análisis.

Otra utilidad se encuentra en la creatividad de los algoritmos para crear contenidos mediáticos. Franganillo (2023) lista algunas de estas funcionalidades:

- **Redacción automática de noticias:** La IA es capaz de producir noticias sobre economía, deportes y meteorología, entre otros temas. Franganillo expone que periódicos como el Daily Mail o el Wall Street Journal ya hace uso de algoritmos generativos para escribir artículos cortos, como, por ejemplo, los vaivenes de la bolsa, o resultados deportivos.
- **Síntesis de imágenes a partir de texto:** Los algoritmos generativos permiten generar imágenes a partir de texto escrito. Esto puede tener valor periodístico pues permite crear imágenes conceptuales que acompañen a noticias. Desde junio de 2022 se han visto varios ejemplos de uso, uno de ellos, en la portada de la revista The Economist, para ilustrar una portada dedicada a la IA.

3.1.2. Principales dilemas éticos

Una vez vistos los beneficios, cabría preguntarse, ¿por qué sería necesario aplicar un marco ético en beneficio de la humanidad, si el enfoque de estos algoritmos es en primera vista antropocéntrico? Lo cierto es que, existen algoritmos que, como veremos a continuación, a pesar de su enfoque humano, en el camino terminan por violar algunos derechos fundamentales como la privacidad o la integridad de las personas, o pueden llegar a fomentar la injusticia.

Se comienza entonces a hablar de los problemas que más se encuentran en la literatura, y se completa el listado con algunos dilemas que han surgido a lo largo de los años.

3.1.2.1. Falta de transparencia en los algoritmos

En los Estados Unidos, los procesos penales están siendo agilizados mediante el uso de algoritmos. En el caso Loomis, en el cual se juzgó a Eric Loomis por participar en un tiroteo y conducir un coche robado, se utilizó un algoritmo de predicción, el COMPAS, un algoritmo desarrollado por la empresa Equivant, que indicaba la probabilidad de un individuo de que volviese a delinquir. En ese caso, el algoritmo clasificó a Loomis como un individuo de alto riesgo, y por ello se le condenó a la pena máxima. El caso fue muy criticado ya que, a pesar de que el algoritmo indicaba que su predicción era de alta precisión, y los datos utilizados eran correctos, se desconocían los razonamientos que el algoritmo utilizaba en su interior para tomar decisiones (Washington, A. L., 2018). Además, al ser un algoritmo desarrollado por una empresa privada, su funcionamiento interno era desconocido por el público.

En este caso se denota, por una parte, la falta de transparencia del algoritmo, pues su razonamiento carecía de explicación, y por otra, la falta de responsabilidad de aquellos que usaron el algoritmo, pues relegaron su autonomía en la toma de decisiones en el algoritmo.

3.1.2.2. Privacidad y seguridad en los datos utilizados

Los sistemas de IA necesitan de datos para ser entrenados para realizar sus funciones. Muchas veces, el origen de los datos utilizados para entrenar a estos sistemas suele no ser claro. Aún se recuerda el escándalo de Cambridge Analytica, en el que Facebook dio acceso a dicha empresa a los datos de 87 millones de usuarios con el fin de realizar predicciones y recomendaciones (Berghel, 2018). En ocasiones, al utilizar tales cantidades de información, se desconoce si el dato utilizado pudiera infringir la privacidad de terceros, aunque otras, como en el caso anterior y el que se verá a continuación, demuestran un uso deliberado de este.

En la medicina este problema es aún mayor pues esos datos son de carácter sensible y, aunque sirvan para entrenar a modelos que puedan ayudar a mejorar la asistencia médica, si el paciente no da su consentimiento, trabajar con ellos implica una violación de su privacidad e integridad. Un ejemplo de esto puede verse en el proyecto Nightingale, una iniciativa de Google con la que buscaba entrar en el mercado de la sanidad estadounidense, a expensas de tratar, sin su consentimiento, información sensible

de 50 millones de usuarios de Ascension, una empresa proveedora de asistencia médica (Schneble, C. O. et al., 2020). Los datos que utilizaron los empleados de Google, además, contenían información no anonimizada de cada uno de los pacientes, y en ningún momento se les habría notificado del uso que se le estaría dando a su información. provienen de pacientes que deben mostrar su consentimiento expreso (Keskinbora, K. H., 2019).

3.1.2.3. Sesgo

Los datos sesgados envenenan a los algoritmos de IA, causando que puedan dar resultados racistas sexistas o xenófobos como se muestra en los siguientes ejemplos. En el caso de los sistemas de reconocimiento facial, por ejemplo, los algoritmos representan una diferencia de rendimiento con respecto a las razas, siendo con las mujeres negras con quienes representan un peor rendimiento, fruto de su falta de representación en las bases de datos con las que se entrenan al algoritmo (Ferrante, E., 2021).

El sesgo algorítmico puede volverse un problema aún mayor si se tiene en cuenta que algunos algoritmos pudieran usarse para el diagnóstico médico. Zou y Schiebinger (2018) ponen como ejemplo un algoritmo de Deep Learning creado en 2017, cuya función sería la detección de cáncer de piel. Al haberse usado menos de un 5% de imágenes de personas negras para entrenar al algoritmo, el rendimiento de éste variaría en función del paciente que lo usara.

Más allá del dato, Mónica Villas y Javier Camacho (2023), en su manual, reconocen los siguientes sesgos presentes en el diseño de algoritmos:

- **Sesgo de aprendizaje:** Surge cuando las opciones de modelado amplifican la disparidad de rendimiento entre distintos conjuntos de datos. Las opciones en cuestión se refieren a las métricas de optimización que se usan en los algoritmos para que aprendan de los datos, las cuales podrían influir en los resultados que ofrece el modelo.
- **Sesgo de evaluación:** En este caso, los datos de referencia no representan a la población que utiliza el algoritmo. Este sesgo se origina debido a que la verificación de la calidad de los algoritmos a menudo se realiza

utilizando datos que provienen de algunos bancos de datos online de referencia. Verificar en función de dichos conjuntos de datos esperando que esto produzca un mayor rendimiento en el modelo, plantea el riesgo de que el algoritmo generalice en función de estos nuevos datos. A esto se le añade el problema de que dichos datos de referencia pueden no mostrar una buena representación de la sociedad y, por lo tanto, aparezcan ya sesgados.

- **Sesgo de agregación:** Se produce cuando se utilizan datos de diferentes grupos y se asume que la vinculación de las distintas entradas de datos con las etiquetas asignadas a esos datos es consistente para todos los grupos, a pesar de que podría no serlo. El sesgo puede derivar de problemas de diseño en el algoritmo, como etiquetar automáticamente de manera errónea debido a fallos en el diseño o, en el caso en que se analicen mensajes de usuarios en una plataforma como Twitter, clasificar de manera diferente a personas que utilizan terminología distinta para describir lo mismo, según señalan Mónica Villas y Javier Camacho.
- **Sesgo de implementación:** Sucede cuando existe una desconexión entre el problema que el algoritmo busca resolver y la manera en que se implementa. Se produce entonces el fenómeno "framing trap", en el que los algoritmos funcionan bien en un entorno controlado o autónomo, pero fallan al ser implementados dentro de estructuras sociales.

3.1.2.4. Problema de la atribución de la responsabilidad

Es importante determinar dónde se halla la responsabilidad en el caso en el que los algoritmos de IA fallen. Coeckelbergh (2020) pone como ejemplo los vehículos autónomos; en el caso en el que fallen y produzcan un accidente ¿quién es responsable? Por otro lado, también se pregunta, ¿cómo pueden construirse los sistemas de IA de forma en que se garanticen que estos se utilicen de forma responsable?

Coeckelbergh reconoce que la agencia de la responsabilidad es más sencilla de adjudicar cuando el causante directo de un accidente es un ser humano, dado que la ética, desde la perspectiva aristotélica y a lo largo de la historia, estudiaría casos en los que la acción es realizada por un agente humano. Coeckelbergh asume que los algoritmos no cumplen los criterios para poseer agencia moral completa, como poseer libertad y

conciencia, por lo que no pueden ser considerados responsables. Compara el caso con las condiciones aristotélicas para ser el responsable de un acto moral: (1) realizar una acción y (2) ser consciente de realizar dicha acción. Bajo esta perspectiva declara que "no tiene sentido pedir que la IA actúe voluntariamente o sin ignorancia ya que un agente de IA carece de las precondiciones para esto".

Otra opción sería atribuirle la responsabilidad total de las decisiones del algoritmo al ser humano. No obstante, si el humano carece de control sobre el algoritmo, ¿seguiría siendo responsable?

3.1.2.5. Desigualdad económica en los algoritmos

Los algoritmos pueden fomentar la apertura de la brecha social entre distintos grupos socioeconómicos. Cotter y Reisdorf (2020) hablarían de la disparidad que produce en las sociedades el conocimiento algorítmico. El conocimiento algorítmico implica comprender cómo funcionan y evolucionan los algoritmos, un elemento crucial para comprender el mundo digital de hoy día. Cotter y Reisdorf debaten entonces sobre cómo el contexto socioeconómico de las personas influye en su conocimiento algorítmico, en especial, la educación, la cual supone un factor relevante según el estudio estadístico que exponen en su artículo. Por tanto, la desigualdad en el conocimiento algorítmico dejaría a algunas personas con mayor conocimiento con mayores posibilidades para usar estos sistemas y beneficiarse de ellos. Por el contrario, aquellos con menor conocimiento sobre los algoritmos, podrían llegar a aceptar información de éstos como verdades incuestionables, siendo influenciados por los sesgos en su diseño. Finalizan su investigación diciendo que sus estudios sugieren que los grupos privilegiados estarían mejor posicionados para beneficiarse del conocimiento algorítmico que aquellos que no lo son.

A continuación, debido a su relevancia actual y al haber generado problemas éticos que no se encontraban en los algoritmos de Deep Learning convencionales, se dedica un espacio al dilema de la información creada por los algoritmos generativos. Se pone el enfoque en el algoritmo GPT, del cual se habló en profundidad en el capítulo anterior, y se tratan los dilemas que traen su implementación en ChatGPT para la generación de información y en Dall-E, para la generación de imágenes.

3.1.3. El problema de la información en los contenidos producidos por GPT

ChatGPT es una herramienta basada en un algoritmo de IA generativa, entrenado a partir de información existente, cuya capacidad creativa depende en gran medida de la integridad de los datos de entrenamiento introducidos. Esto quiere decir que, si los datos utilizados para entrenar al algoritmo son incorrectos o sesgados, esto se propagará a los resultados que el algoritmo sea capaz de reproducir. Además, como se mencionó anteriormente, GPT está integrado por una red neuronal que discrimina el contenido que debe o no ser mostrado al usuario. Sin embargo, no puede saberse si esta segunda red está bien entrenada, ni qué razonamiento sigue para tomar sus decisiones, debido a su falta de explicabilidad. Es por ello por lo que deben analizarse con ojo crítico los resultados que estas herramientas puedan producir, pues pueden no encontrarse libres de sesgos, infracciones a la propiedad intelectual o pueden llegar a proveer información inexacta. A continuación, se muestran algunos de los problemas que se atribuyen al algoritmo debido a la falta de transparencia y a la posible falta de precisión en su diseño.

3.1.3.1. Desinformación

La desinformación producida por los algoritmos de IA Generativa, en concreto GPT, plantea cuestiones éticas que deben tratarse. Por un lado, estos algoritmos pueden utilizarse de forma maliciosa con la intención de generar desinformación en la sociedad con el fin de manipularla. Un ejemplo de esto puede verse en la creación de deepfakes, imágenes generadas por sistemas como Dall-E, cuyo fin es el de generar contenido visual engañoso. En el pasado, se han generado, con algoritmos, imágenes falsas sobre líderes políticos o sobre hechos actuales, que pueden utilizarse con el fin de manipular la integridad de procesos democráticos, como las elecciones (Diakopoulos, N., & Johnson, D., 2021). El peligro de la desinformación se encuentra en que una vez recibida, es difícil de contrarrestarse. A nivel cognitivo, la información falsa, a pesar de haberse recibido tan solo una vez, persiste en el individuo (Lewandowsky et al., 2012). Esta desinformación puede terminar por manipular la capacidad de elección de la sociedad, siendo altamente nociva para el desarrollo humano. No obstante, estos algoritmos, debido a su baja explicabilidad, también podrían producir desinformación en el usuario

Por otro lado, también existe la posibilidad de usar estas herramientas con la intención de suplantar la identidad en redes sociales, mediante imagen, vídeo o audio, y

con ello manipular a los usuarios (González Arencibia, M., & Martínez Cardero, D., 2020).

Franganillo (2023) alerta también de los riesgos de los deepfakes en los medios de comunicación. En mayo de 2022, el canal de televisión France 3 estrenó un formato que utilizaba la tecnología deepfake para recrear a celebridades fallecidas y hacerles hablar con el presentador. Ante esto, se plantea si es ético utilizar la imagen de personas fallecidas diciendo cosas que quizá jamás pronunciarían.

3.1.3.2. Robo de información

Una de las preocupaciones acerca del contenido generado por IA, y concretamente del contenido artístico, como imágenes o vídeo es: ¿está el contenido creado, completamente libre de derechos de autor? ¿Se puede considerar robo el producir contenido con estas tecnologías?

En el caso de OpenAI, la compañía reconoce haber tomado datos de obras con copyright, con el fin de entrenar a sus programas de IA. Sin embargo, algunos de sus stakeholders consideran que, el uso que dan a sus datos entraría dentro del fair-use, pues no infringen directamente el copyright de sus obras (Zirpoli, 2023). Por lo tanto, la responsabilidad en la creación y distribución de contenido originado con estas herramientas pasaría a los usuarios de estas aplicaciones. Se atisba aquí un dilema de no saber sobre quién recaería la responsabilidad de la creación de estos contenidos ¿sobre el usuario, el creador del algoritmo, o el propio algoritmo?

3.1.3.3. Perpetuación de sesgos

Por otro lado, los algoritmos generativos poseen al igual que los de Deep Learning problema ante el sesgo algorítmico. Si los datos con los que se entrena al algoritmo tienen sesgos raciales en los que pueda haber una mayoría de personas de una determinada raza, el output que saldrá de éste estará claramente determinado por esos datos iniciales. Este problema se hace aún más patente en algoritmos generativos de imagen, donde las imágenes creadas dependen de las peticiones del usuario. En ese sentido, el algoritmo podría generar imágenes con personas blancas por defecto, y añadir personas de otras

razas cuando este considerara que debe añadir a una persona 'exótica', reforzando estereotipos raciales y contribuyendo a la discriminación (Zhou K. Q., 2023).

Como puede verse, estos problemas, a pesar de que dependen en parte del buen uso del usuario, están ligados a cómo se encuentra diseñado el algoritmo. Incluso al hablar de los deepfakes, que son responsabilidad de quien los crea, pues es el algoritmo el que da al usuario la capacidad de realizar estas prácticas engañosas. Estos han sido algunos de los problemas que producen estos algoritmos desde el punto de vista de la generación de información.

A continuación, veremos qué propuestas ofrecen distintas organizaciones mundiales acerca de un marco ético para la creación y uso de algoritmos.

3.2. PROPUESTAS DE LAS INSTITUCIONES

En esta sección, se exploran las directrices que proponen diversas instituciones expertas en los algoritmos de IA, con respecto a alcanzar una ética en el desarrollo de algoritmos. Varias organizaciones alrededor del mundo han aportado para el establecimiento de una regulación que proteja tanto al usuario final que utilice estos algoritmos como a terceros, cuyos datos o integridad puedan verse comprometidas, del uso, creación, y entrenamiento de los algoritmos de IA.

Se adentra en las guías de recomendaciones y buenas prácticas que proponen diversas organizaciones relevantes en la aplicación de la ética a los algoritmos de IA. En concreto este punto investigará acerca de la propuesta de AlgorithmWatch, y de Trustworthy AI.

3.2.1. AlgorithmWatch

La iniciativa AlgorithmWatch nace como un repositorio de recomendaciones acerca de la IA y se enfoca en mostrar cómo los algoritmos toman las decisiones que acaban teniendo impacto en la sociedad. Para esta organización, ningún algoritmo es neutro, y es importante conocer cómo estos han sido entrenados, de dónde han obtenido los datos, y cómo éste interpreta los resultados (Villas M., Camacho J., 2023). Concuerda esto con la imagen de la tecnología que muestra Oriol Quintana (2023): La tecnología no es neutra, aunque carezca de utilidad en su nacimiento. El hecho de afirmar que ésta es

neutral predispone a aceptar cualquier tipo de novedad tecnológica con la excusa de que aún carece de propósito (Quintana, 2023).

AlgorithmWatch pone el foco del problema donde este trabajo busca incidir, el algoritmo como raíz del dilema moral. En la plataforma se encuentran recopilados diversos artículos en los que se hacen críticas y recomendaciones de uso de los algoritmos de IA, por lo que resulta un repositorio que debieran.

Sin embargo, AlgorithmWatch se queda limitado en ser un repositorio de recomendaciones, careciendo de la autoridad necesaria en el desarrollo de algoritmos. Por ese motivo, en el siguiente punto, se verá cuál es la propuesta de la Comisión de la UE para crear una guía ética para el desarrollo de algoritmos de IA.

3.2.2. Trustworthy AI

Se escoge la propuesta de la Trustworthy AI por ser una de las más reconocidas a nivel global y por provenir de una institución respetada como lo es la Unión Europea. Se trata ésta de uno de los referentes más destacados en el ámbito ético de la Inteligencia Artificial, respaldada por una prestigiosa institución. Además, se elige esta propuesta como un buen ejemplo de propuesta ética ya que, como afirman sus responsables, los tratados de derechos humanos en la UE ofrecen una prometedora base para identificar principios éticos (Fukuda-Parr, S., & Gibbons, E., 2021).

La Comisión Europea nombró a un grupo de expertos en inteligencia artificial, que asesorarían a la UE en materia ética. El grupo de expertos de alto nivel para la inteligencia artificial (o AI HLEG por sus siglas en inglés) definiría en 2019 unas directrices éticas para una IA confiable, la Trustworthy AI (Smuha, N. A., 2019). Esta guía estaría compuesta de cuatro principios éticos, y siete requisitos que se espera que los algoritmos de IA cumplan.

Los principios éticos en los que se basa serían el respeto por la autonomía humana, prevención del daño, la justicia y explicabilidad de los algoritmos. Además, se añaden a los principios estas siete directrices: Supervisión humana, robustez y seguridad técnica, privacidad, transparencia, diversidad, bienestar social y medioambiental, y responsabilidad.

Sin embargo, a pesar de lo interesante de la propuesta, ésta puede quedar algo limitada para nuestro enfoque por los siguientes motivos:

- **La extensión del proyecto:** El proyecto tardó en llevarse a cabo 6 meses, un tiempo muy limitado, aunque el resultado terminó siendo bastante satisfactorio. Sin embargo, el problema no reside ahí, y es que, pese a que este tratado está redactado por expertos, teniendo en cuenta el avance de la tecnología, una declaración de principios realizada en ese breve espacio de tiempo puede quedarse obsoleta en el tiempo. Es necesaria la existencia de una ética específica, descentralizada, de los algoritmos, dinámica, que sea capaz de adaptarse a los avances de la tecnología.
- **La carencia de un planteamiento ético aplicable:** La propuesta del HLEG, carecería en principio de una guía de actuación para los desarrolladores y usuarios finales, aunque, finalmente, en abril de 2021, después de tres años de trabajo gracias a las aportaciones del AI HLEG, se publicó la Regulación Europea sobre IA, enfocada a proporcionar el marco jurídico para una IA confiable.
- **El tamaño del equipo:** El AI HLEG está compuesto por 52 expertos. Si bien, son expertos, la naturaleza de su propuesta podría encontrarse sesgada por los intereses de sus integrantes.

Queriendo hacer énfasis en este último punto, debe decirse que en el equipo habría una pequeña fracción de expertos en ética, de los que unos pocos de ellos serían filósofos. Aunque es importante que la industria tecnológica esté involucrada en el desarrollo de procesos de regulación de algoritmos, que el grupo estuviera tan desbalanceado entra en conflicto con la ambición de proveer una buena guía ética. Sin embargo, esta composición parecería no ser coincidencia, sino estar en la línea de las preferencias políticas de la Comisión por poner en mejor posición a las empresas europeas en el mercado global (Heilinger, J. C., 2022).

En contraparte a lo anterior, nuestro interés estará principalmente en obtener una ética desinteresada, no involucrada en procesos políticos, que sea una ética centrada en el ser humano y en salvaguardar los derechos humanos. Esta debe estar alejada del concepto de ethics-washing, el cual se refiere a que las grandes empresas entren en el debate de los

algoritmos solo con el fin de prevenir futuras regulaciones (Fukuda-Parr, S., & Gibbons, E., 2021).

Por otro lado, será interesante revisar una de las más recientes regulaciones de la Inteligencia Artificial, el EU AI Act, que nacería dos años después de la creación de esta guía de principios.

3.3. ASPECTOS REGULATORIOS DE LA IA EN LA UE

Finalmente, es necesario dedicar un punto al estado actual de la regulación de la Inteligencia Artificial, con el fin de mostrar el estado de la ética de los algoritmos en el ámbito jurídico. Por los mismos motivos expuestos en la elección del marco de la Trustworthy AI, además de por esperarse que entre en vigor en los próximos años, se expone a continuación la reciente EU AI Act, la primera ley de regulación de la IA publicada en la Unión Europea. La Comisión Europea publicaría la propuesta de la EU AI Act el 21 de abril de 2021, una ley que buscaría crear un marco normativo para el desarrollo, comercialización, y uso de las IAs.

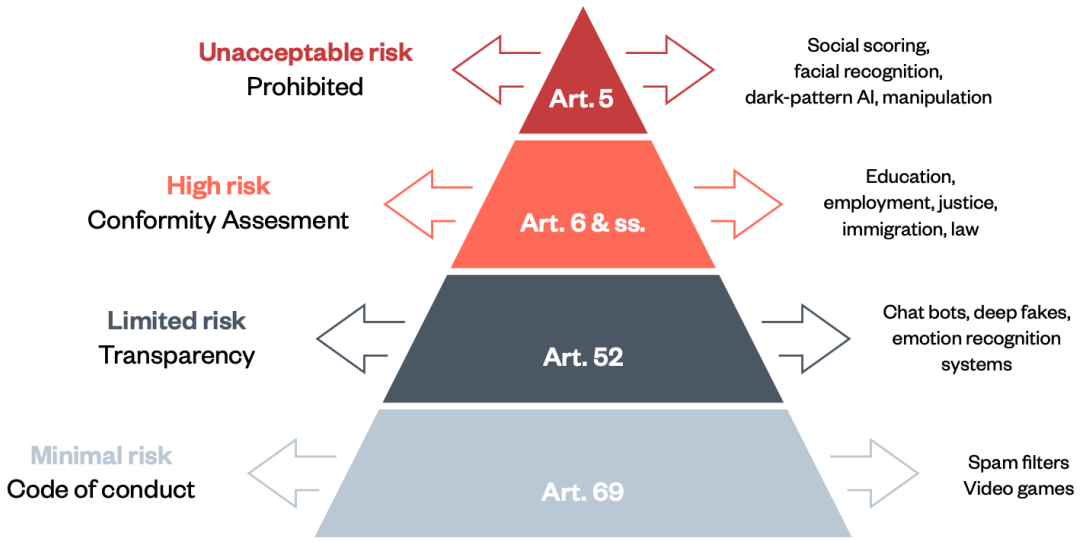

Figura 4. Pirámide de riesgos establecida en el AI Act. Recuperado de https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/

Los algoritmos estarían regulados en esta ley según una escala de riesgos de cuatro niveles, la cual se basaría en la intención con la hubieran sido creados (Edwards, L., 2021):

- **Riesgo inaceptable:** Estos algoritmos estarían prohibidos por violar derechos fundamentales. Serían aquellos relacionados con la manipulación, explotando vulnerabilidades de un grupo de personas, la puntuación social, o el reconocimiento facial en tiempo real en lugares públicos.
- **Alto riesgo:** Algoritmos extensamente regulados, pero que no suponen tal riesgo como para ser prohibidos. Se consideran de alto riesgo aquellos que incidan sobre la educación, la medicina o la justicia
- **Riesgo limitado:** Estas aplicaciones solo debieran mostrar transparencia en sus procesos. Entre ellos se encuentran chatbots.
- **Bajo riesgo:** Aplicaciones de bajo riesgo como filtros de spam o videojuegos que utilicen IA. Sobre ellas, la Comisión indica que deben ceñirse a sus códigos de conducta.

En función del nivel de riesgo de los algoritmos, las regulaciones impuestas a éstos serán más o menos restrictivas.

Hacker, Engel y Mauer (2023) criticarían a la ley del AI Act por su reducido enfoque, en el que avances recientes como los modelos generativos o excesivamente complejos, no tendrían cabida o no estarían correctamente definidos:

- En el Artículo 3(1b), se define a los modelos generativos de forma genérica, sin tener en cuenta la diversidad de estos.
- Una definición más limitada tampoco evitaría otros problemas. Debido a la versatilidad de estos sistemas, los desarrolladores no serían capaces de cumplir lo que se ve en el Artículo 4c, en el que se responsabiliza al desarrollador de eliminar los casos de uso de alto riesgo. Lo cierto es que, aunque éste eliminase todos los posibles casos potencialmente peligrosos, aún el usuario podría utilizar el sistema para casos de uso sensibles, como, por ejemplo, resumir información confidencial como registros médicos o información bancaria.
- La ley tendría consecuencias adversas en el mercado competitivo de la IA, pues se regularizaría por igual a aquellos sistemas de código abierto, ya que muchos de estos desarrollarían por filantropía o investigación. Para estos desarrolladores sería muy costoso cumplir con todos los requisitos de la ley, que ha sido mayormente pensada para empresas como Meta, Google o Microsoft. Sin

embargo, la ley indica que no se trataría por igual a aquellos casos de uso, habría que ver que ocurre con el paso del tiempo con la ley aplicada.

3.4. NECESIDAD DE UN PLANTEAMIENTO ÉTICO APLICABLE

Vistos algunos de los problemas actuales, queda claro que la ética debe estar presente en cada etapa del desarrollo tecnológico. Como se ha visto anteriormente, la aplicación de la ética en este campo no es algo novedoso, ya que ya existen unos principios y regulaciones que aquellos interesados en el uso y desarrollo de algoritmos debieran seguir. Sin embargo, parece que estos principios no son suficientes pues, o no atacan correctamente el problema de raíz, o no ofrecen información acerca de cómo llevar a cabo la consecución de sus principios.

Heilinger (2022) recopila un listado de cuestiones de distintos estudios de ética sobre la IA que se considera esencial que una ética tecnológica pudiera resolver:

- ¿Cómo podemos **entender y explicar** los sistemas de IA?
- ¿Sobre quién recae la **responsabilidad** de sus algoritmos?
- ¿Cómo prevenir a los algoritmos de ser envenenados con los **sesgos**?
- ¿Cómo puede mantenerse la **privacidad** de la gente si los datos personales son tan fáciles de recopilar y analizar?
- ¿Cómo puede protegerse la **autonomía humana** en la toma de decisiones?
- Y, por último, ¿cómo puede prevenirse que una IA maligna imponga sus objetivos sobre los objetivos humanos?

Se hace patente entonces la necesidad de hallar un planteamiento ético que pueda ser llevado a la práctica en el diseño y uso de algoritmos, que tenga en cuenta estas cuestiones. Para este planteamiento, será necesario recordar que la tecnología existe para cubrir las necesidades humanas, y facilitar en lo posible la vida de las personas (Quintana, 2023), por lo que el enfoque de esta búsqueda debe ser antropocéntrica. En el momento en que se pierde este enfoque y por consiguiente se acaba creando primero la tecnología para posteriormente buscarle una utilidad, nos ponemos al servicio de nuestra propia creación, siendo esclavos de la tecnología.

En el siguiente punto, se tratarán algunas propuestas éticas con el fin de encontrar un marco ético bajo el que se pueda construir una propuesta de ética aplicada que pueda utilizarse a la hora de desarrollar nuevos algoritmos de aprendizaje profundo.

4. ÉTICA APLICADA PARA MITIGAR LOS PROBLEMAS DE LOS ALGORITMOS

Con el fin de conseguir realizar una propuesta que atenúe los dilemas algorítmicos, y teniendo en cuenta el razonamiento del punto anterior, se buscará un enfoque antropocéntrico, que ponga al ser humano en el centro, y no a la tecnología. Es necesario controlar a los algoritmos para asegurar ese enfoque en el ser humano, todo ello con el fin de asegurar la justicia y evitar un mayor sufrimiento humano (Bryson, J. J., & Theodorou, A., 2019).

Vistos los puntos anteriores, se buscará un marco ético que cumpla los siguientes requisitos:

- **Reconocer la importancia de un enfoque multidisciplinar:** Se busca que la ética reconozca la necesidad de que la aplicación de sus principios debe estar presente desde el diseño del algoritmo. Propone una ética cuyo eje central está en los algoritmos, reconociendo la importancia de la complementación ético-técnica.
- **Reconocer el problema en los algoritmos:** Se busca una ética que tenga como objetivo atacar todos los problemas que aparecen debido a la implementación de algoritmos de Deep Learning.
- **Sea antropocéntrica:** El objetivo final que tenga tal marco ético debe siempre ser poner al ser humano por delante de la tecnología, y que luche por garantizar los derechos humanos, libre de otro tipo de intereses económicos o políticos.

En la búsqueda de ese marco ético, se opta por proponer el uso de la algor-ética, una reciente propuesta de principios éticos, la cual, como se verá a continuación, cumple los tres puntos expuestos. Sin embargo, la algor-ética, a pesar de su propuesta, es un campo que en la literatura aún no está bien definido, pues aún pocos autores conocen de su existencia, y es un concepto relativamente reciente. Por ello, el objetivo final de crear una propuesta de ética aplicada será además una aportación a esta teoría ética que se espera que siga desarrollándose a lo largo de los años venideros.

Debe tenerse en cuenta que esta propuesta de ética aplicada no tiene carácter impositivo, sino más bien, se considera una guía de recomendaciones que buscaría acercar la ética del plano teórico al plano práctico en el diseño y uso de algoritmos. En este sentido,

además, el uso de la algor-ética para este fin debiera servir para influenciar las políticas públicas, sin que en ningún caso se perciba como un marco regulatorio jurídico.

Por otro lado, se buscará completar la propuesta de la algor-ética con una propuesta tecno-sostenible, incluyendo la dimensión de la sostenibilidad como uno de los puntos que deben cubrir los algoritmos de IA en el futuro, pues su elevado coste de computación comienza a ser problemático, debido a las altas emisiones de CO2 de los servidores que los sostienen.

4.1. LA ALGOR-ÉTICA

El papa Francisco mostraría a finales de 2020 su preocupación acerca del avance de los algoritmos: "Recemos para que el progreso de la robótica y la Inteligencia Artificial esté siempre al servicio del ser humano". Entonces reconocería, al igual que nosotros ahora, la necesidad del enfoque antropocéntrico del avance tecnológico. El Vaticano apoyaría a su vez ese mismo año la propuesta de la algor-ética en la *Rome Call for AI Ethics,* conferencia en la que, junto a directivos de IBM, Microsoft y el gobierno italiano, se trataría el problema ético del avance incontrolado de los algoritmos.

La algor-ética es una propuesta que sugiere atacar la problemática ética de la inteligencia artificial desde el paradigma de los algoritmos. El principal impulsor de esta propuesta es Paolo Benanti, teólogo italiano, quien desde la fundación RenAIssance, fundada en 2021, investiga sobre la ética en la tecnología. Parte de la base de que los problemas que pueden producir las IAs, están originados desde lo que las construye de base. Por ello, la algor-ética aboga por una creación y uso responsable de los algoritmos, con el fin no de limitar la IA sino al contrario, de ponerla al servicio del ser humano (Jobin, A. et al., 2019).

Debe hacerse énfasis en que la algor-ética busca ejercer influencia sobre las políticas públicas, sin ejercer un carácter impositivo desde una esfera superior. Si bien es necesario abordar los problemas de la ética desde un carácter político y jurídico, esta propuesta no busca imponer una regulación sobre los algoritmos. Sin embargo, puede y debiera servir de influencia y apoyo para futuras iteraciones de leyes sobre los algoritmos como la anteriormente vista AI Act.

¿Por qué la idea de la algor-ética es distinta a lo visto anteriormente? Con la algor-ética se abre un nuevo campo de la ética, se estudia la tecnología y su relación con el ser humano, y se desarrolla a la par con el avance de los algoritmos. La algor-ética es un campo de estudio dinámico que abandona el ser limitada a una lista de recomendaciones como ocurriera con la Trustworthy AI.

Paolo Benanti haría una aportación acerca de la algorética que concuerda con el punto anterior. En ella diría que la algor-ética no debe limitarse a una lista de precauciones que los desarrolladores debieran tomar para evitar que los algoritmos sean peligrosos, "un arma bien diseñada no es menos peligrosa por un virtuoso cuidado en su diseño", sino que debe ser capaz de guiar a la humanidad, y que, para ello, necesitamos una ética que nos proteja y guíe hacia dónde queremos ir como especie. Refiriéndose a la algor-ética, "nuestra contribución va en la dirección de ayudar a la humanidad a decidir por sí misma" (Benanti, P., 2023).

Refuerza aquí Paolo Benanti la idea de la necesidad de una guía acerca de cómo utilizar los algoritmos de IA de forma ética, que no se limite a evitar que los algoritmos sean peligrosos, sino que sirvan como catalizadores del desarrollo humano, y que tengan una perspectiva antropocéntrica. Antes de definir los principios éticos por los que lucha esta propuesta, se presenta el humanismo digital como otra teoría ética que complementará a la algor-ética, pues ésta considera de nuevo que la tecnología debiera ser antropocéntrica. Gracias al humanismo digital, se podrá precisar aún más, como deberían crearse futuros algoritmos de IA.

4.1.1. Humanismo digital

Es importante entender la postura del humanismo digital para entender la algor-ética, pues esta, defiende la postura de la algor-ética en lo referente a guiar al ser humano mediante la tecnología para alcanzar el mayor bien. El humanismo digital, al igual que la algor-ética, tiene un planteamiento antropocéntrico, donde el bienestar del ser humano es la principal prioridad.

Nida-Rümelin, filósofo impulsor de esta idea, afirma que el humanismo digital aboga por el uso de las nuevas tecnologías con el fin de mejorar las condiciones de vida humanas y fomentar la sostenibilidad del medio ambiente, teniendo en cuenta los

intereses de generaciones futuras. A la vez, el humanismo digital se opone a la auto-depreciación de la competencia humana en la toma de decisiones, y a que el ser humano relegue su autonomía a los algoritmos autónomos (Nida-Rümelin, J., 2022).

4.1.2. Principios de la algor-ética

El 28 de febrero de 2020, el "Llamamiento de Roma para la ética de la IA" (Rome Call for AI Ethics) definió, basándose en la algorética, los principios éticos bajo los que debieran construirse los futuros algoritmos (Pegoraro, R., 2023). Estos principios coinciden con lo que se lleva viendo en la literatura desde hace años, y, además, coinciden con las problemáticas que más se han repetido en la literatura. Se expondrán a continuación dichos principios y se comparará, en función a lo visto anteriormente, en qué estado se encuentra la técnica con respecto a estos principios:

4.1.2.1. Transparencia

Los algoritmos deben ser explicables y comprensibles para los usuarios. Como se explicó previamente, los algoritmos han llegado a un grado de complejidad en el que su funcionamiento y concretamente, conocer qué lleva a un algoritmo a tomar una decisión, se ha vuelto imposible. En el capítulo 2 se trató de hacer un acercamiento, explicando el funcionamiento de los algoritmos de IA a lo largo de la historia, con el objetivo de mostrar que los algoritmos de IA han mutado en complejidad a lo largo del tiempo, a la vez que han sido sumergidos cada vez más bajo capas de abstracción, hasta el punto de que resulta imposible para sus propios creadores poder explicar el comportamiento de una red neuronal una vez está en funcionamiento.

Se vuelve necesario que, si buscamos confiar en el funcionamiento de un algoritmo, este sea de por sí explicable, y que haga entender al usuario el por qué ha tomado una decisión y no otra, pues se busca que la tecnología no arrebate la autonomía en la toma de decisiones al ser humano.

4.1.2.2. Inclusión

Los algoritmos deben tener en cuenta las necesidades de todos los seres humanos y ser capaces de beneficiarles por igual. Todos los individuos deben tener las mismas oportunidades de expresarse y desarrollarse con estas herramientas. Por ello, además, los algoritmos que se desarrollen bajo este marco deben luchar por la no exclusión de ningún

usuario, así como de evitar el fenómeno visto anteriormente con el conocimiento algorítmico, con el fin de evitar abrir más aún la brecha socioeconómica.

4.1.2.3. Responsabilidad

Aquellos que diseñan e implementan algoritmos de IA deben proceder con responsabilidad y transparencia. Esto es, por la parte de la transparencia, deben dar cuenta de cómo está creado su algoritmo, antes de comercializarlo, como ocurriera en el caso de COMPAS. En este punto, y tras analizar los problemas que origina la IA generativa, se añade el matiz de que la responsabilidad no debería estar limitada a los desarrolladores de los algoritmos, sino a los que hagan uso de ellos, siempre que los desarrolladores sean transparentes con el origen de la información de sus sistemas.

4.1.2.4. Imparcialidad

No crear o usar algoritmos que fomenten el sesgo en los humanos, con el fin de salvaguardar la dignidad humana. Este punto puede confundirse con el punto de inclusión, ya que van muy alineados. Sin embargo, un algoritmo puede cumplir ser inclusivo, en el sentido en el que se crea para cumplir derechos humanos fundamentales, pero debido a problemas de sesgo en su diseño, pudiera acabar por favorecer a un grupo determinado, convirtiéndose en un algoritmo no imparcial.

4.1.2.5. Fiabilidad

Los algoritmos de IA deben asegurar un resultado confiable, es decir, el algoritmo debe estar diseñado para cumplir su propósito de forma correcta, pues de ello dependerá que se confíe en ellos para ser aplicados en campos vistos previamente como la medicina. Además, deben ser capaces de dar al usuario razones para ser confiable, por lo que este principio se complementaría con el principio de transparencia.

4.1.2.6. Seguridad y privacidad

Los algoritmos de IA deben asegurar la seguridad y la privacidad de sus usuarios. Como se ha visto previamente, uno de los graves problemas existentes es el referente a los datos utilizados para diseñar estos modelos. Si bien, pueden obtenerse mejores resultados cuanto mayor sea la muestra de datos utilizada, estos datos en ningún momento deberían comprometer la privacidad del individuo ni contener información que pudiera

ser considerada sensible, a menos que el usuario diera el expreso consentimiento para su tratamiento.

Vistos los principios que propone el tratado, surge un problema, y es que, a pesar de atacar a todos los problemas que se han mencionado previamente, como ocurriera con la Trustworthy AI del AI HLEG, se trata tan solo una declaración de principios, y carece de unas líneas de actuación que los diseñadores y usuarios de algoritmos pudieran seguir de cara a futuro. Sin embargo, esto nos da la posibilidad de poder construir nuestra aportación; una guía de aplicación de la ética que complemente a estos principios, y que complemente a la iniciativa de la algorética. En búsqueda de completar una propuesta antropocéntrica en los algoritmos, se añadirá en el siguiente punto la idea de la creación de algoritmos sostenibles.

4.2. SOSTENIBILIDAD DE LOS ALGORITMOS

Por último, si se busca un diseño de algoritmos antropocéntrico, es interesante añadir la dimensión de la sostenibilidad a nuestra propuesta. La sostenibilidad, como dijera Nida-Rümelin, es uno de los pilares que conforman el humanismo digital. Esta es considerada uno de los principios éticos que más preocupan a los investigadores en la literatura (Jobin, A. et al., 2019), pero parece que no se le da la importancia que se debiera con respecto al resto de principios. Y es que el avance de los algoritmos, siendo cada vez más computacionalmente exigentes, suponen un mayor consumo energético que puede no ser sostenible a largo plazo. Estos algoritmos que buscan resultados sin plantearse su impacto medioambiental son denominados Red AI.

Por otro lado, los algoritmos verdes, o Green AI, son aquellos algoritmos que tratan de poner la eficiencia como primer criterio, junto a la precisión, fomentando una reducción en los recursos utilizados. Éstos, buscan reducir la huella de carbono, y aumentar su inclusividad (Schwartz, R., et al., 2020).

En el siguiente punto, se explicarán formas de promover los algoritmos verdes y de conseguir llevar a la práctica lo visto en este capítulo.

4.3. PROPUESTA APLICABLE

Conocidos los marcos éticos bajo los que se pretende enmarcar el campo de la investigación en los algoritmos de IA de cara a futuro, solo resta definir cuáles deberían ser las directrices que debieran seguir estos algoritmos para poder integrar la algorética en futuros proyectos en algoritmos de IA. Para ello se tendrán en cuenta los principios que se establecieron en la Rome Call for AI Ethics, extraídos directamente de la algor-ética.

Mediante la propuesta de la algor-ética, se buscará conseguir crear una guía de ayuda que sirva a empresas, desarrolladores, y usuarios, a utilizar los algoritmos de inteligencia artificial de una forma ética. Así, se fomentará un desarrollo sostenible de la tecnología, además de una comercialización y un uso ético de estos algoritmos. Para este enfoque, es fundamental que los principios éticos estén presentes durante todo el ciclo de vida de los algoritmos de IA, si se busca asegurar que estos sean diseñados, implementados y comercializados de forma ética (Zhou, J. et al., 2020).

Por ello, uniendo técnicas de diseño de algoritmos con teorías éticas, se propondrá a continuación un marco ético-práctico basado en los principios de la algor-ética, que pueda servir como apoyo para desarrollar futuros algoritmos de IA. Con ello, se cumplirá el tercer objetivo de este trabajo, proponer una solución ética aplicable a los problemas de los algoritmos.

4.3.1. Cumplimiento de los principios de la algor-ética

¿Cómo podemos garantizar el cumplimiento de los principios de la algor-ética? A continuación, se mostrará una recopilación de tecnologías, buenas prácticas y teorías éticas, que nos ayudarán a confeccionar una guía de ética aplicada, la cual servirá para complementar a la propuesta de la algorética. Para cada uno de los principios, se presentan las siguientes soluciones.

4.3.1.1. Explicabilidad

La dimensión de la explicabilidad en los modelos es necesaria para justificar por qué un algoritmo toma una decisión y no otra. Ya que es problemático. La consecución de este principio es complicada, pues requeriría que el algoritmo implementase esa funcionalidad en su etapa de desarrollo. Con el fin de conseguir aplicar la explicabilidad a los algoritmos, nacería el programa XAI.

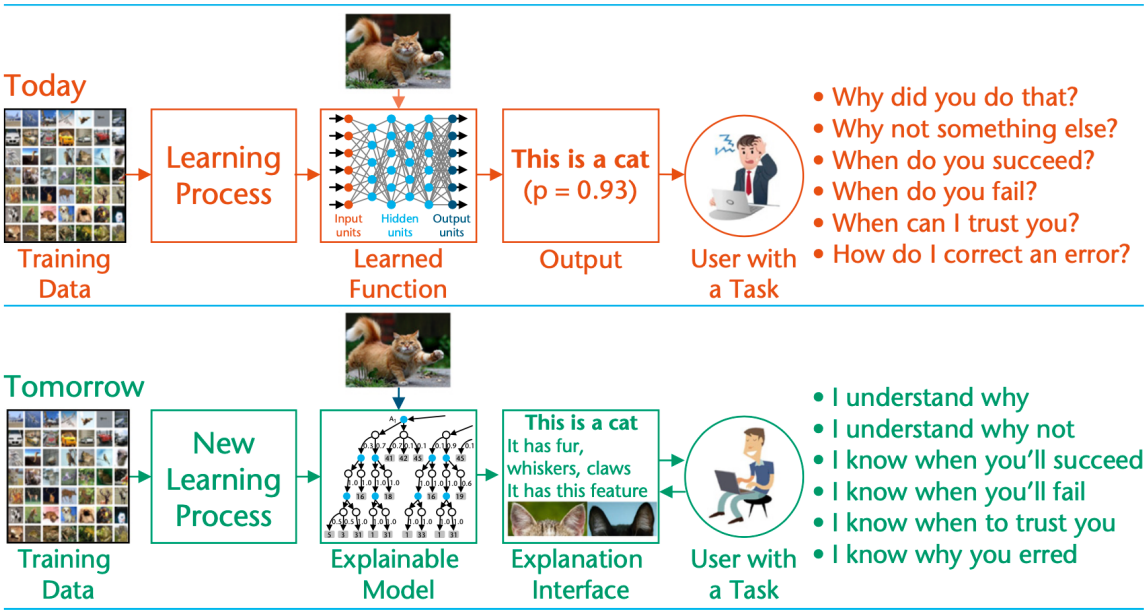

Figura 5. Modelo XAI. Recuperado de "DARPA's explainable artificial intelligence (XAI) program". Gunning, D., & Aha, D. (2019). AI magazine, 40(2), 44-58.

XAI (eXplainable AI) es un concepto creado por DARPA en 2017, con el propósito de hacer la IA más explicable. El programa XAI tiene como objetivo crear nuevos algoritmos de IA que, combinados con técnicas efectivas de explicación, ayuden al usuario final a entender, confiar y gestionar a los sistemas de IA que se creen con estos algoritmos (Gunning, D., & Aha, D., 2019). La figura muestra cómo funcionaría el modelo, el cual mostraría al usuario explicaciones que le ayudarían a entender por qué un modelo se comporta de una determinada forma, y así facilitarle a tomar sus propias decisiones con respecto a esta información.

Por otro lado, dentro de la dimensión de la transparencia, también se considera importante luchar por la buena comprensión de la explicación de los sistemas por parte del usuario. Es importante incidir en esta dimensión, no solo por evitar posibles peligros sino por los beneficios que pueden aportar los sistemas transparentes: Un usuario capaz de comprender los resultados que devuelva un sistema estará mejor informado, y, por tanto, mejor capacitado para tomar sus propias decisiones.

Finalmente, en la transparencia de los algoritmos se incluye la necesidad de la transparencia de los datos, pues debe conocerse en base a qué datos los algoritmos toman

sus decisiones. Yanisky-Ravid, y Hallisey, proponen un Modelo de Transparencia para garantizar la transparencia de los datos, basado en cuatro componentes (2019):

- Los stakeholders en la industria de la IA deberían llevar auditorías y examinar los datos a los que sus algoritmos de IA están expuestos, en función del tipo de dato recopilado y del riesgo de su mal uso.
- Los stakeholders deberían retener los datos que hubieran usado para entrenar a los algoritmos en caso de que en un futuro fuera necesario investigar cómo se desarrolló el sistema de IA.
- Las auditorías deberían estar estandarizadas y llevadas a cabo por terceros que fueran objetivos y que no fueran desarrolladores del algoritmo a investigar. Éstos terceros deberían también certificar los resultados de las auditorías.
- Finalmente, los operarios de sistemas de IA deberían poder tener algunas protecciones de responsabilidad cuando cumplan con los requisitos del Modelo, pero, a pesar de ello, se produzca algún daño.

Este último punto se opone a lo visto en el AI Act, ley en la que el desarrollador se consideraría responsable del mal uso que se diera de su herramienta, a pesar de haber evitado cualquier tipo de mal uso en ésta.

4.3.1.2. Inclusión

Con el fin de evitar que los algoritmos agraven la brecha socioeconómica y, por el contrario, aboguen por cerrarla, se verá que existen diversas propuestas que luchan por hacer de la IA una herramienta que defienda la inclusividad. Flores-Vivar y García-Peñalvo (2023) muestran varias ideas en las que aplicar algoritmos a la educación serviría para garantizar una educación universal, accesible y de calidad. Algunas de estas propuestas son:

- **Garantizar que las aulas estén disponibles para todos** independientemente de si tienen alguna discapacidad auditiva o visual al poder, por ejemplo, implementar subtítulos en tiempo real en las presentaciones de clase con PowerPoint Translator.
- **Sistemas de tutoría inteligentes**. Pueden realizarse servicios de tutoría inteligentes para los alumnos en función de las dificultades que tengan. Un

ejemplo en el que se aplica esta tecnología está en el sistema de tutoría SHERLOCK para pilotos de fuerza aérea.

- **Colaboración entre profesores e IA.** La visión de la IA en la educación prevé que esta tecnología y los profesores puedan trabajar juntos para obtener el mejor resultado para los estudiantes.

Sin embargo, la implementación de estas propuestas en la educación posee muchos retos que deben afrontarse antes de aplicar estas medidas ya que, no formar a los docentes en materia de estos algoritmos, convertiría a los algoritmos en herramientas de exclusión. Estas herramientas deben servir para facilitar la labor de los docentes en la educación y no para apartarles de ella.

4.3.1.3. Imparcialidad

La dimensión de la imparcialidad debe luchar contra el problema del sesgo en los algoritmos. Por lo tanto, una de las primeras cuestiones que debiera abordarse, se hallaría en determinar el origen de dicho sesgo. Popularmente, se cree que la solución para conseguir la imparcialidad en los sistemas de predicción se encuentra únicamente en los datos, por lo que la solución pasaría por suministrar al algoritmo de datos no sesgados. Sin embargo, en contra de esta creencia, el diseño técnico del propio algoritmo contribuye también a amplificar estos sesgos (Hooker, S., 2021).

Por otro lado, es cierto que parte de la culpa la tendrán los datos utilizados, y que normalizarlos atenúa el problema del sesgo. Sin embargo, a veces se hace muy costoso, y otras, el propio algoritmo podría priorizar unos datos sobre otros debido a cómo éste estaría construido.

Sara Hooker, investigadora de Google Brain, equipo de investigación sobre la Inteligencia Artificial en Google, afirmaría que los algoritmos no son imparciales, y que siempre estarán influenciados por las decisiones en su diseño. Algunas decisiones, como implementar técnicas de privacidad, por ejemplo, añadir ruido en las imágenes, pueden afectar al error en la clasificación de fotografías de personas de piel más oscura del dataset normalizado, Diversity in Faces. Otras decisiones de diseño sutiles como modificar la curva de aprendizaje de los modelos y la duración de su entrenamiento, también parecen influir en este fenómeno.

Hooker termina indicando que es necesario una mayor investigación sobre el origen del sesgo desde el punto de vista algorítmico ya que, atribuir la responsabilidad de este problema a únicamente el dato puede considerarse difusión de la responsabilidad. La solución se hallaría entonces en mitigar el daño que tanto el dato como el algoritmo conjuntamente pudieran realizar.

4.3.1.4. Responsabilidad

Con el fin de asegurar la implementación de algoritmos de IA responsable, debe matizarse que, asegurar la responsabilidad sobre estos no solo implicaría realizar la tarea de supervisar su ciclo de vida, desde su concepción hasta su aplicación, sino también, cuestionar la autonomía de estas herramientas. Como argumenta Adela Cortina: Le hemos dado la etiqueta de autónomo a herramientas creadas por el ser humano por el hecho de haberles dado la capacidad de tomar decisiones. La autonomía es algo propio del ser humano, no consiste solo en ser capaz de elegir, sino también en tener la capacidad de auto legislarse y autodeterminarse (Cortina, 2022). Además, si buscamos construir algoritmos antropocéntricos, es contraproducente relegar nuestra autonomía en pro de diluir nuestra responsabilidad.

Si bien, se defiende la importancia de las decisiones humanas sobre las algorítmicas en nuestros modelos, debe ahora evitarse la reducción de nuestra responsabilidad en las acciones que involucren la relación con los algoritmos. Dicho de otra manera, debe evitarse la 'responsabilidad distribuida' en los algoritmos. Los actos y decisiones de los sistemas de IA suelen ser resultado de múltiples interacciones entre muchos actores, tanto diseñadores, desarrolladores, usuarios, e incluso, software y hardware, lo que origina un fenómeno conocido como agencia distribuida (Taddeo, M., & Floridi, L., 2018). Es con esta agencia distribuida que nace la responsabilidad distribuida, es decir, la disolución de la responsabilidad del individuo. En base al problema de la responsabilidad tecnológica, han surgido nuevas teorías éticas que buscan abordar este dilema.

Interesa la teoría de Luciano Floridi (2016) con respecto a la responsabilidad distribuida en la tecnología. Floridi explica cómo las acciones realizadas por distintos actores, sin una intencionalidad moral explícita, pueden influir en un sistema distribuido, como pudiera ser un algoritmo de IA. Estas acciones, al ser agregadas pueden finalmente

dar resultados éticamente positivos o negativos. Luciano reconoce que la ética clásica no puede afrontar este problema, pues no tiene en cuenta grandes sistemas interconectados como los que existen en la actualidad. Por lo tanto, considera necesaria una ética que considere la responsabilidad distribuida para abordar los nuevos dilemas tecnológicos.

Por otro lado, es también interesante aplicar el principio de responsabilidad de Hans Jonas a la algor-ética (Terrones Rodríguez, A. L., 2018), pues puede servir como punto de partida al tratar la responsabilidad en los algoritmos. Hans Jonas ya adelantaría la necesidad de crear una ética sobre la tecnología, pues la posibilidad de hacer el mal, en la actualidad, no recaería solo en las relaciones interpersonales, como adelantara Floridi en el párrafo anterior. Hans Jonas elaboraría entonces el principio de responsabilidad, el cual implica que el ser humano debe vivir en la actualidad de forma que pueda garantizar su supervivencia en el futuro. Esta propuesta se relaciona además con la sostenibilidad, pues ese sería el fin de una sostenibilidad antropocéntrica, como la que buscamos tratar con la algor-ética.

Terrones Rodríguez (2018) introduciría el concepto del principio de la responsabilidad a los dilemas de la IA. Un acercamiento interesante es el que propone con respecto a los algoritmos de IA y la automatización de trabajos que antes realizaran los seres humanos. Concluye que no debemos tener miedo a la tecnología, pero sí debemos pensar, desde la ética de la responsabilidad, antes de aplicar la IA en campos en los que profesionales pudieran verse afectados, pues no queremos atentar contra la esencia de determinadas profesiones si se quiere garantizar un futuro de estabilidad.

4.3.1.5. Fiabilidad

La fiabilidad de un algoritmo, va a depender en gran medida de la consecución del resto de principios.

Para que un sistema sea fiable, una de las primeras dimensiones que debe cumplir es que éste sea transparente. La fiabilidad se complementaría con la transparencia, pues el usuario será capaz de confiar más en sistemas que sean capaces de explicar qué hacen y por qué lo hacen (Larsson, S., & Heintz, F., 2020).

Por último, un sistema fiable deberá garantizar la seguridad y privacidad del usuario. En el siguiente punto, se verán pautas de cómo puede lograrse esto.

4.3.1.6. Seguridad y privacidad

Uno de los problemas que poseen los algoritmos de Deep Learning, y que se ha visto previamente en casos como el del proyecto Nightingale, es que se haga uso de datos de usuarios con información sensible para entrenar a los modelos. Una solución que existe para evitar comprometer la privacidad de las personas a quien perteneces esos datos, pasa por la anonimización del dato. Esto es, la eliminación de cualquier elemento identificativo que pudiera encontrarse en la información que se suministre al algoritmo.

Los desarrolladores de algoritmos poseen en la actualidad herramientas que pueden utilizarse con el fin de anonimizar los datos que utilicen en sus modelos. Estas herramientas, desarrolladas como librerías de Python, se implementarían en la etapa de diseño del algoritmo. Mónica Villas y Javier Camacho (2023) hacen un listado de algunas de estas herramientas:

- **PrivacyPanda:** librería que serviría para eliminar identificadores sensibles de usuarios.
- **AMNESIA:** Herramienta que elimina identificadores directos, como el nombre y DNI, y transforma secundarios, como la fecha de nacimiento y el código postal.
- **ARX:** Ofrece diversas transformaciones de datos, así como análisis de riesgos y evaluación de la utilidad
- **SdcMicro:** Paquete para la generación de datos anónimos que además incorpora métodos de estimación de riesgos.
- **The Cornell Anonymization Toolkit:** Una de las primeras herramientas de anonimización de datos. Transforma datos no sensibles en rangos de valores y ofrece la posibilidad de eliminación de los sensibles.

Sin embargo, la anonimización de datos, a pesar de ser una muy buena opción, no garantiza por completo la privacidad de los usuarios, pues estos datos son susceptibles de ser deanoniminizados a partir de la información no sensible (Villas M., Camacho J., 2023). Se hacen entonces necesarias medidas complementarias para garantizar la privacidad, siendo una de estas el marco 'Privacy by Design'. Este marco es una guía de desarrollo de productos o servicios que considera que la privacidad debe ser parte fundamental del ciclo de vida de éstos. Posee siete principios, listados por Everson (2016), quien además lo hace desde la perspectiva de aplicaciones que hacen uso del big data:

- **Privacidad proactiva no reactiva.** La privacidad debe ser preventiva y no un remedio.
- La privacidad debe estar **por defecto** en el sistema
- La privacidad debe estar **involucrada desde la etapa de diseño**
- Una privacidad completamente funcional en el ciclo de vida del producto
- **Seguridad de punto a punto**. La privacidad debe estar en todo el ciclo de vida del producto
- **Visibilidad y transparencia.** Debe conocerse cuál es el ciclo de vida del dato: dónde se origina, hacia dónde va, para qué propósito va a utilizarse, y cómo va a destruirse. En el momento en el que no puede responderse alguna de estas cuestiones, existe un riesgo a la privacidad.
- Respeto por la privacidad del usuario, mantener un **enfoque centrado en el usuario**

4.3.2. Creación de sistemas sostenibles

Por último, y aunque no sea uno de los principios expuestos en la Rome Call for AI Ethics, es importante considerar la dimensión de la sostenibilidad en esta propuesta, pues como se ha visto, es una de las principales preocupaciones que aparece en la literatura de la ética de los algoritmos.

Desde el punto de vista técnico, Roy Schwartz junto a su equipo (2020), proponen distintas técnicas para medir la eficiencia energética de los algoritmos:

- **Emisiones de carbono:** Llama la atención la medición de las emisiones de carbono, pues es una métrica que interesa reducir. Sin embargo, medir estas emisiones es muy difícil, pues depende de la infraestructura eléctrica en la que se ejecute el modelo, y de muchos parámetros como la zona en que se realice o el momento del día.
- **Uso de electricidad:** Las GPUs ofrecen hoy en día la posibilidad de ver cuánta electricidad consumen. Sin embargo, esta electricidad variaría en función de la máquina que ejecutara el algoritmo.
- **Tiempo:** El tiempo empleado en ejecutar la respuesta de un algoritmo es proporcional al trabajo computacional que tendría que realizar la máquina.

- **Número de parámetros:** Una forma común de medir eficiencia es mirando el número de parámetros que se utilizan para entrenar al modelo.
- **FPO (Floating Point Operations):** Desde el artículo Green AI, proponen contar el número de operaciones de punto flotante que realizan los modelos, como forma concreta de medición de la eficiencia. Este método está comúnmente aceptado como forma de medición de la huella energética del software, pero no ha sido adoptado en IA. El artículo da los siguientes motivos para utilizar esta métrica:
    - Computa la cantidad de trabajo realizado por la máquina y está vinculada a la cantidad de memoria consumida.
    - El método no depende del hardware en el que se ejecute el modelo.
    - Este método, además, tiene en cuenta el tiempo que el modelo lleva ejecutándose.

Por otro lado, Mehlin, V., Schacht, S., y Lanquillon, C. (2023), muestran una serie de herramientas que servirían para realizar una estimación de la huella de carbono de los algoritmos si se implementan en los modelos. Algunas de estas son:

- **Eco2AI:** Librería de Python que monitoriza la energía consumida por la CPU y GPU del ordenador. Realiza una estimación de las emisiones de carbono teniendo en cuenta la región desde la que se realiza el análisis.
- **CodeCarbon:** Paquete que se puede integrar en Python. Estima el CO2 producido por el ordenador o por el servicio en la nube que tenga desplegado el algoritmo. Además, da recomendaciones para optimizar el código para reducir las emisiones, y sugiere servidores en la nube para desplegar algoritmos en regiones donde se utilicen energías renovables.
- **GreenAlgorithms:** Herramienta para medir la huella de carbono de las computaciones algorítmicas. Requiere poca información y tiene en cuenta varias configuraciones de hardware en las que podría desplegarse el algoritmo.

5. CONCLUSIONES

La creación de un marco que regule la ética de los algoritmos, como se ha visto, es una tarea complicada, pues el desarrollo tecnológico avanza rápido y a pasos grandes. Como ejemplo, la ley AI Act, la cual aún no ha llegado a implementarse, ha quedado obsoleta a día de hoy por no tener en cuenta la diversidad en la aplicación de los algoritmos generativos. Es por tanto necesaria una ética de los algoritmos que pudiera servir como apoyo como la algor-ética, que sea cambiante y capaz de adaptarse a la tecnología según se desarrolla.

A través de este trabajo, se ha proporcionado en el capítulo 2 una base técnica sólida que permita comprender la naturaleza del algoritmo a aquellos que busquen influenciar y aportar en su diseño desde una postura ética y filosófica, además de hacer entender el problema que implica el avance tecnológico descontrolado de la tecnología en la transparencia. Posteriormente, en el capítulo 3, se han expuesto los beneficios y dilemas que han traído los algoritmos en la actualidad, junto a las propuestas éticas que provienen de diversas instituciones, tanto en lo referente a políticas públicas como a lo relativo a marcos regulatorios, y se debate por qué, tal vez, esas propuestas no sean suficientes para solventar dichos problemas.

Por ello, con el fin de aportar una propuesta que pueda mitigar los problemas algorítmicos, el capítulo final se estructura en base a propuestas éticas recientes que abogan por construir una ética de los algoritmos, orientada a influir sobre las políticas públicas, como la algor-ética. En este contexto, se ha diseñado una guía de buenas prácticas en el diseño y uso de algoritmos bajo los principios de la Rome Call for AI Ethics. Sin embargo, debido a las limitaciones impuestas por tratarse este proyecto de un Trabajo de Fin de Grado, la propuesta pudiera quedar un poco limitada en proyectos a gran escala, y aún más teniendo en cuenta la diversidad de los algoritmos existentes, con sus propias funcionalidades y problemas.

Una dimensión que no ha podido tratarse en esta guía debido al enfoque del proyecto, pero que queda abierta a futuros trabajos de investigación, se encuentra en el marco regulatorio, concretamente en lo relativo a la IA Generativa. Una guía de buenas prácticas en el desarrollo de algoritmos no es suficiente para limitar el uso malicioso de los algoritmos por parte de los usuarios que, como se ha podido comprobar, pueden

realizar con ellos técnicas de deepfakes o generar información engañosa. A pesar de los esfuerzos para abordar la problemática de la agencia distribuida en la responsabilidad, se argumenta la necesidad de una regulación que sancione a aquellos usuarios que busquen hacer daño mediante el uso de esta tecnología.

A pesar de las limitaciones impuestas por el alcance de este proyecto, su contribución sigue siendo significativa para la algor-ética, pues sirve como un paso hacia la construcción de esta ética de cara a futuro, además de servir para la construcción de nuevos algoritmos que sean responsables y confiables. Se hacen necesarias entonces pequeñas aportaciones que puedan sumar con el fin de obtener una ética robusta que abogue por un diseño antropocéntrico de la tecnología en los próximos años.

BIBLIOGRAFÍA

Abeliuk, A., & Gutiérrez, C. (2021). Historia y evolución de la inteligencia artificial. *Revista Bits de Ciencia*, (21), 14-21.

Bali, J., Garg, R., & Bali, R. T. (2019). Artificial intelligence (AI) in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required?. *Indian journal of ophthalmology*, *67*(1), 3–6.

Benanti, P. (2023). The urgency of an algorethics. *Discover Artificial Intelligence*, *3*(1), 11.

Berghel, H. (2018). Malice domestic: The Cambridge analytica dystopia. Computer, 51(05), 84-89.

Bryson, J. J., & Theodorou, A. (2019). How society can maintain human-centric artificial intelligence. *Human-centered digitalization and services*, 305-323.

BURGOS, L. M., SUÁREZ, L. L., & BENZADÓN, M. (2023). INTELIGENCIA ARTIFICIAL CHATGPT Y SU UTILIDAD EN LA INVESTIGACIÓN: EL FUTURO YA ESTÁ AQUÍ. *MEDICINA (Buenos Aires)*, *83*, 0000.

Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002). Deep blue. *Artificial intelligence*, *134*(1-2), 57-83.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, *26*(4), 2051-2068

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). *Introduction to algorithms*. MIT press.

Cortina Orts, A. (2019). Ética de la inteligencia artificial. In Anales de la Real Academia de Ciencias Morales y Políticas (pp. 379-394). Ministerio de Justicia.

Cortina, A. (2022). Los desafíos éticos del transhumanismo. Pensamiento. Revista de Investigación e Información Filosófica, 78(298 S. Esp), 471-483.

Cotter, K., & Reisdorf, B. C. (2020). Algorithmic knowledge gaps: A new horizon of (digital) inequality. *International Journal of Communication*, *14*, 21.

Dennett, D. C. (1995). Conciencia explicada. Madrid: Paidós.

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, *23*(7), 2072-2098.

Edwards, L. (2021). The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*, *1*.

El Papa propone que los avances en robótica e inteligencia artificial sean "humanos" para que supongan un progreso real (6 Nov. 2020). *EuropaPress.* https://www.europapress.es/sociedad/noticia-papa-propone-avances-robotica-inteligencia-artificial-sean-humanos-supongan-progreso-real-20201106140328.html

Ergen, M. (2019). What is artificial intelligence? Technical considerations and future perception. *Anatolian J. Cardiol*, *22*(2), 5-7.

Everson, E. (2016). Privacy by design: Taking ctrl of big data. *Clev. St. L. Rev.*, *65*, 27.

Fernández Fernández, J. L. (2021). Hacia el Humanismo Digital desde un denominador común para la Cíber Ética y la Ética de la Inteligencia Artificial. X

Ferrante, E. (2021). Inteligencia artificial y sesgos algorítmicos ¿Por qué deberían importarnos? *Nueva sociedad*, (294), 27-36.

Flores-Vivar, J. M., & García-Peñalvo, F. J. (2023). Reflexiones sobre la ética, potencialidades y retos de la Inteligencia Artificial en el marco de la Educación de Calidad (ODS4).

Floridi, L. (2013). Distributed morality in an information society. *Science and engineering ethics*, *19*, 727-743.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160112.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535-545.

Foster, D. (2022). *Generative deep learning*. " O'Reilly Media, Inc.".

Franganillo, J. (2023). La inteligencia artificial generativa y su impacto en la creación de contenidos mediáticos. *methaodos. revista de ciencias sociales*, *11*(2), 15.

Fukuda-Parr, S., & Gibbons, E. (2021). Emerging consensus on 'ethical AI': Human rights critique of stakeholder guidelines. *Global Policy*, *12*, 32-44.

González Arencibia, M., & Martínez Cardero, D. (2020). Dilemas éticos en el escenario de la inteligencia artificial. *Economía y Sociedad*, *25*(57), 93-109.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, *40*(2), 44-58.

Hacker, P., Engel, A., & Mauer, M. (2023, June). Regulating ChatGPT and other large generative AI models. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1112-1123)

Heilinger, J. C. (2022). The ethics of AI ethics. A constructive critique. *Philosophy & Technology*, *35*(3), 61.

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, *2*(4).

J. C. Hay, B. E. Lynch, and D. R. Smith, "Mark I perceptron operators' manual," Cornell Aeronautical Lab., Cornell Univ. Library, Buffalo, NY, USA, Rep. VG-1196-G-5, 1960.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399.

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, *22*(14), 2081-2096.

Keskinbora, K. H. (2019). Medical ethics considerations on artificial intelligence. *Journal of clinical neuroscience*, *64*, 277-282.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, *13*(3), 106-131.

Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*.

Luckow, A., Cook, M., Ashcraft, N., Weill, E., Djerekarov, E., & Vorster, B. (2016, December). Deep learning in the automotive industry: Applications and tools. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 3759-3768). IEEE.

Mantini, A. (2022). Technological sustainability and artificial intelligence algor-ethics. *Sustainability*, *14*(6), 3215.

Marvin, M., & Seymour, A. P. (1969). Perceptrons. *Cambridge, MA: MIT Press*, *6*, 318-362.

McCulloch, W. S. y W. Pitts (1943), "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Mathematical Biophysics*, 5:115-133.

Moshayedi, A.J., Roy, A.S., Kolahdooz, A., & Shuxin, Y. (2022). Deep Learning Application Pros And Cons Over Algorithm. *EAI Endorsed Transactions on AI and Robotics*.

Nida-Rümelin, J., & Weidenfeld, N. (2022). Digital Humanism: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age (p. 127). Springer Nature.

Nida-Rümelin, J. (2022). Digital Humanism and the Limits of Artificial Intelligence. Perspectives on Digital Humanism, 71-75.

Galaviz, J. (2016). La mente en la máquina. Academia Mexicana de Ciencias. Vol. 67, 1.

Lock, S. (2023, 6 febrero). What is AI Chatbot phenomenon ChatGPT and could it replace humans? *the Guardian*. https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans

Olmeda, M. V., & Ibánez, J. C. (2022). *Manual de ética aplicada en Inteligencia Artificial*. Anaya Multimedia.

Pataranutaporn, P., Danry, V., Leong, J., Punpongsanon, P., Novy, D., Maes, P., & Sra, M. (2021). AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, *3*(12), 1013-1022.

Pegoraro, R., & Curzel, E. (2023). Rome call for AI Ethics: the birth of a movement. *Medicina y ética*, *34*(2), 315-349.

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2016). Deepcare: A deep dynamic memory model for predictive medicine. In *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II 20* (pp. 30-41). Springer International Publishing.

Quintana, Oriol (2023), Sobre la Tecnología.

Rome Call For Ethics. 2020. Recuperado de: https://www.romecall.org/wp-content/uploads/2022/03/RomeCall_Paper_web.pdf

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533-536.

Sag, M. (2023). Copyright safety for generative AI. *Forthcoming in the Houston Law Review*.

Schank, R. C. (1987). What is AI, anyway? *AI magazine*, *8*(4), 59-59.

Schneble, C. O., Elger, B. S., & Shaw, D. M. (2020). Google's Project Nightingale highlights the necessity of data science ethics review. *EMBO molecular medicine*, *12*(3), e12053.

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green ai. *Communications of the ACM*, *63*(12), 54-63.

Searle, J. R. (1980): "Minds, Brains and Programs", Behavioral and Brain Sciences 3 (3), 417-457.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *nature*, *529*(7587), 484-489.

Smuha, N. A. (2019). The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, *20*(4), 97-106.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751-752.

Terrones Rodríguez, A. L. (2018). Inteligencia artificial y ética de la responsabilidad. *Cuestiones de Filosofía; Volumen 4, número 22 (Enero-Junio 2018)*.

Villas Olmeda, M. y Camacho Ibánez, J. (2023). Manual de ética aplicada en Inteligencia Artificial. Anaya

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ*, *17*, 131.

Yanisky-Ravid, S., & Hallisey, S. K. (2019). Equality and privacy by design: A new model of artificial intelligence data transparency via auditing, certification, and safe harbor regimes. *Fordham Urb. LJ*, *46*, 428.

Ras, G., Xie, N., Van Gerven, M., & Doran, D. (2022). Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, *73*, 329-396.

Zirpoli, C. T. (2023). Generative Artificial Intelligence and Copyright Law.

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., & Savage, S. (2020, December). A survey on ethical principles of AI and implementations. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 3010-3017). IEEE.

Zhou, K. Q., & Nabus, H. (2023). The Ethical Implications of DALL-E: Opportunities and Challenges. *Mesopotamian Journal of Computer Science*, *2023*, 17-23.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair.