Facultad de Ciencias Económicas Empresariales (ICADE)

LA ÉTICA APLICADA AL USO DE LA INTELIGENCIA ARTIFICIAL GENERATIVA

Autor: Santiago Gil Peñaranda

Tutor: Javier Fuertes Pérez

MADRID| Junio, 2023

RESUMEN

A lo largo de los últimos años, el avance imparable de la Inteligencia Artificial (IA) ha permeado en todos los aspectos de nuestra existencia, catapultando a la humanidad hacia la cuarta revolución industrial. Este trabajo emprende una investigación detallada sobre cómo la sociedad puede y debe adaptarse a estas transformaciones tecnológicas radicales, en especial en lo que se refiere a la aparición de la Inteligencia Artificial Generativa (IAG) y los Modelos de Lenguaje de Gran Escala (LLM, por sus siglas en inglés). Más allá de su mera adopción, es imprescindible que estos desarrollos se realicen de manera éticamente consciente y responsable. Se investiga los diferentes tipos de Inteligencia Artificial, enfocando como principal tipo la Inteligencia Artificial Generativa. Se estudia el entrenamiento y funcionalidad de estos híper sistemas de computación con el fin de entender porque surgen tantos problemas éticos en el desarrollo y la implementación de ellos. Finalmente, se propone un análisis de ciertos principios éticos con la intención de entender cómo se pueden aplicar al diseño de la IAG. Se estudiarán estos principios a través del Humanismo Digital, y cómo se aplicará esta teoría para guiar el desarrollo y la implementación de estas tecnologías para evitar daños potenciales y asegurar que su utilización sea benefactora para los aspectos de la sociedad.

Palabras clave: Inteligencia Artificial (IA), Inteligencia Artificial Generativa (IAG), Humanismo Digital, LLM, Redes neuronales, ética, problemas éticos.

ABSTRACT

Over the past few years, the unstoppable advance of Artificial Intelligence (AI) has permeated every aspect of our existence, catapulting humanity into the fourth industrial revolution. This paper undertakes a detailed investigation of how society can and should adapt to these radical technological transformations, especially as it relates to the emergence of Generative Artificial Intelligence (GAI) and Large Scale Language Models (LLMs). Beyond their mere adoption, it is imperative that these developments are carried out in an ethically conscious and responsible manner. The different types of Artificial Intelligence are investigated, focusing on Generative Artificial Intelligence as the main type. The training and functionality of these hyper-computing systems is studied in order to understand why so many ethical problems arise in their development and implementation. Finally, an analysis of certain ethical principles is proposed with the intention of understanding how they can be applied to IAG design. These principles will be studied through Digital Humanism, and how this theory will be applied to guide the development and implementation of these technologies to avoid potential harm and ensure that their use is beneficial to aspects of society.

Keywords: Artificial Intelligence (AI), Generative Artificial Intelligence (GAI), Digital Humanism, LLM, Neural Networks, ethics, ethical issues.

INDICE

1. Introducción

El impacto de la inteligencia artificial (IA) en nuestras vidas es cada vez más profundo y, en consecuencia, el estudio de sus implicaciones éticas se ha vuelto una cuestión de importancia crucial. Particularmente, la IA generativa, que se refiere a aquellas técnicas de IA que generan nuevas instancias de datos en vez de simplemente analizar o clasificar datos existentes, ha desatado una amplia gama de desafíos éticos únicos.

Este trabajo trata de una reflexión y un estudio sobre la inteligencia artificial con capacidad generativa y los posibles riesgos éticos que conllevan. Buscamos explorar los principios éticos que deberían guiar la implementación de la IA generativa, analizar los desafíos éticos específicos que presenta esta rama de la IA y proponer recomendaciones para el desarrollo y uso éticos de la IA generativa. En una época en la que todos nuestros entornos están atravesando un tramo de digitalización, es importante mantener los principios éticos y tener en cuenta los mejores intereses del ser humano a la hora de tomar decisiones.

El trabajo está motivado por el rápido avance de la IA generativa y su creciente adopción en diversas industrias, desde la generación de contenido en redes sociales hasta la medicina personalizada, pasando por la creación de arte digital y la producción musical. En el contexto de los últimos años, estas aplicaciones se han hecho cada vez más frecuentes. Sobre todo, en los entornos profesionales y educativos, ya se ha visto cómo las nuevas y distintas aplicaciones de la Inteligencia Artificial están revolucionando el rendimiento y la metodología de estas entidades. Lo que antes requería esfuerzo y un proceso de reflexión ahora requiere menos, con el aumento de la automatización que trae consigo nuevas tecnologías que disminuyen la aportación humana y nos facilitan la vida. Aunque la IA generativa tiene el potencial de generar grandes beneficios, también plantea serios desafíos éticos que merecen un estudio detenido.

El crecimiento exponencial de la tecnología, amparado en la Ley de Moore, ha permitido la proliferación de dispositivos inteligentes y la acumulación masiva de datos, lo que ha alimentado el desarrollo y la eficacia de los sistemas de IA. Sin embargo, estos procesos deben de ser regulados

y formulados con un mismo criterio, que intenta mantener al ser humano como el centro del progreso tecnológico mundial. En otras palabras, es crucial garantizar que las consideraciones éticas no queden relegadas a un segundo plano.

En este contexto, este trabajo buscará dimensionar estos problemas y proponer vías de solución desde la ética aplicada. El objetivo es contribuir al desarrollo responsable de la IA generativa y ayudar a sentar las bases para una sociedad donde la tecnología y la ética puedan avanzar de la mano.

2. Objetivos y Metodología

El objetivo de este Trabajo Fin de Grado es analizar los métodos de Inteligencia Artificial Generativa y aplicar ciertos principios éticos a su desarrollo e implementación; y de este modo poder mantener al ser humano al centro del desarrollo tecnológico. Se hará un estudio sobre los criterios aplicados al diseño de la IAG, y una reflexión sobre sus problemas éticos

- Obj 1: Entender cómo la inteligencia artificial está afectando a nuestra sociedad
- Obj 2: Comprender la computación de la inteligencia artificial generativa y relacionar los problemas éticos que pueden surgir en el diseño de la IAG
- Obj 3: Construir una serie de sugerencia y reflexiones para encaminar el futuro de la IAG

Para conseguir estos objetivos, se estudiará la inteligencia artificial y sus aplicaciones recogiendo la información más actual en artículos y libros científicos. Se implementará esta metodología con el ánimo de entender cómo los usos de la IAG, sobre todo en los GLLMM (Generative Large Language Multi-Modal Model), afectan a la sociedad. De esta manera, se podrá analizar los problemas éticos en estos procesos que transformarán el mundo por completo. Finalmente, habrá una entrevista hecha a Pilar López, vicepresidenta de Strategic Partnerships – LSEG – Capital Markets en Microsoft. Pilar es exalumna de E2, y ha tenido una carrera exitosa dentro de Microsoft, una empresa que está muy involucrada en las inversiones de la IA, sobre todo la IAG. La entrevista a Pilar servirá para ver como las grandes empresas están tratando el tema de la IAG, para conocer casos reales de implementación de estas tecnologías y para entender cómo las estrategias corporativas se están adaptando para enfrentar este nuevo panorama. Asimismo, la perspectiva de Pilar ofrecerá valiosos insights sobre las implicaciones éticas y sociales de la IAG, un tema de creciente importancia tanto para las empresas como para la sociedad en su conjunto. Su experiencia en Microsoft brindará un contexto relevante y actualizado para esta discusión, considerando el liderazgo de la empresa en el desarrollo y aplicación de tecnologías avanzadas de IA, particularmente la IA generativa. He decidido entrevistar a Pilar por proximidad y

disponibilidad, pero también por su experiencia en una de las empresas líderes en las inversiones de IA generativa.

Las preguntas hechas a Pilar incluyen:

1. ¿Puedes contarme sobre algún caso en el que Microsoft ha tenido que abordar cuestiones éticas al desarrollar o implementar IAG?
2. ¿Cómo maneja Microsoft los posibles riesgos de sesgo y discriminación en las redes generativas, y qué políticas de igualdad se aplican en estos casos?
3. ¿Cómo están equilibrando Microsoft y otras multinacionales la innovación en la Inteligencia Artificial Generativa con las preocupaciones éticas?
4. ¿Cómo ves el futuro de la IAG en términos de ética? ¿Hay algún problema emergente que te preocupa particularmente?
5. ¿Existe alguna estrategia o plan de contingencia para manejar consecuencias imprevistas o no deseadas de la IA generativa?

3. Entendiendo la Inteligencia Artificial

"Never trust anything that can think for itself if you can't see where it keeps its brain" - **Arthur Weasley, Harry Potter and the Chamber of Secrets.**

3.1. Que es la Inteligencia Artificial

La Inteligencia Artificial (IA) es una disciplina de la ciencia de la computación que se centra en la creación de sistemas capaces de realizar tareas que normalmente requerirían la inteligencia humana, como el reconocimiento de voz, la toma de decisiones, la traducción de idiomas y la percepción visual. Ha experimentado una evolución notable desde sus inicios hasta hoy, aplicado en varios contextos y disciplinas.

La IA ha encontrado su lugar en una amplia gama de sectores, desde la medicina y la educación hasta el comercio y el entretenimiento, revolucionando la forma en que vivimos y trabajamos. A lo largo de su evolución, la IA ha demostrado una capacidad única para resolver problemas complejos y mejorar la eficiencia de los procesos, pero su aplicación no está exenta de dilemas éticos que deben ser considerados en la búsqueda de su utilización responsable y beneficiosa para la humanidad. Es imprescindible que el desarrollo y la implementación de la IA se haga con ánimo de ayudar al ser humano y no corruptirlo. Con el advenimiento de tecnologías como la Inteligencia Artificial Generativa, su relevancia y el impacto potencial en nuestra sociedad solo han aumentado, haciendo de la reflexión y el análisis sobre su uso una necesidad imperante.

3.2. Etapas de la Inteligencia Artificial

La Inteligencia Artificial (IA) ha pasado por varias etapas distintas a lo largo de su historia. Desde sus primeros días de formulación teórica y experimentación, hasta la era contemporánea, la IA se ha convertido en una parte integral de nuestras vidas diarias y actividades comerciales. Estas etapas de la IA no sólo son testigos de los avances tecnológicos, sino también de los cambios en la

percepción y aceptación de la IA en nuestra sociedad. En este apartado, exploraremos estas etapas de la IA, que incluyen su historia inicial, el resurgimiento de la IA tras un periodo de desilusión y la nueva primavera de la IA, el periodo de crecimiento y expansión acelerado que experimentamos. A lo largo de este recorrido, es importante recordar que la IA no es un ente estático, sino una disciplina en constante evolución que se adapta y crece con el paso del tiempo, reflejando las necesidades y expectativas de la sociedad (Russell & Norvig, 2021). El rápido avance de la tecnología significa un crecimiento exponencial en cuanto a las capacidades de la IA de cara al día de mañana. Al entender estas etapas, ganamos una apreciación más profunda de cómo la IA ha llegado a su estado actual y obtenemos una visión de las posibilidades que nos esperan en el futuro.

3.2.1 Historia de la Inteligencia Artificial

En los tiempos modernos, hemos presenciado cómo la "inteligencia artificial" ha ido evolucionando en los medios, pasando de ser un concepto de ciencia ficción a uno cada vez más común en las noticias científicas. Este avance incontenible de la IA está impulsando la cuarta revolución industrial, refiriéndose a las transformaciones digitales que se están produciendo en todas las industrias del siglo XXI.

Esta tecnología disruptiva ha tenido un impacto similar al de la electricidad, cambiando y permeando todos los procesos productivos. Sin embargo, a pesar de su uso frecuente, hay cierta confusión en torno a qué es exactamente la inteligencia artificial. En su sentido más estricto, la IA es un campo de la ciencia que estudia el comportamiento de los sistemas inteligentes.

Aunque sus raíces se encuentran en los trabajos de Norbert Wiener y las teorías de control y comunicación de la cibernética, la IA como campo de estudio fue formalmente fundada en una conferencia celebrada en Dartmouth College en 1956 organizada por John McCarthy, donde se acuñó el término "inteligencia artificial" (Russell & Norvig, 2016). McCarthy propuso un proyecto de investigación en el que invitó a diez académicos a compartir ideas sobre la "inteligencia artificial". Este proyecto, que solicitaba financiación a la Fundación Rockefeller, es uno de los primeros usos documentados del término.

El término "inteligencia artificial" se consolidó después de los años cincuenta para describir la capacidad de ciertas máquinas para realizar tareas que los humanos consideraban indicativas de

inteligencia. Sin embargo, no hay un consenso absoluto sobre la definición de un sistema inteligente. Durante mucho tiempo, la IA se centró en emular los procesos del cerebro y la conducta humana.

El famoso científico Alan Turing, mundialmente conocido como el "Padre de la Inteligencia Artificial", amplió estas ideas en su artículo "Computing machinery and intelligence", en el que propuso el famoso "test de Turing" o también conocido como el juego de la imitación (imitation game). Esta prueba consiste en evaluar si una máquina puede pensar por sí misma y pasar por un ser humano ante otro juez humano. Funciona de esta manera; un evaluador humano tiene conversaciones con dos interlocutores a través de una comunicación textual, siendo uno de ellos un humano y el otro la maquina en cuestión. Si después de una serie de preguntas el evaluador humano no puede distinguir cuál de los interlocutores es la maquina y cuál es el ser humano, se dice que la maquina ha pasado el "Test de Turing", lo cual se le atribuye capacidad "pensar" o actuar de una manera indistinguible a un ser humano. Pero la cuestión va más allá de si las maquinas pueden pensar, sino si la semántica de lo que entendemos por pensamiento se puede atribuir a lo que realizan estas máquinas. (Turing 1950)

De todas formas, antes del descubrimiento de Turing, estaban Walter Pitts y Warren McCulloch, que sentaron las bases de las redes neuronales artificiales con la creación del primer modelo de una neurona artificial. En su innovador artículo de 1943, "A Logical Calculus of the Ideas Immanent in Nervous Activity", propusieron un modelo de la neurona como una máquina lógica binaria, dando un paso crucial hacia la intersección de la neurociencia y la informática (McCulloch & Pitts, 1943). Los modelos que desarrollaron describen cómo las neuronas pueden producir resultados binarios, "encendido" o "apagado", en función de los estímulos de entrada que reciben. Este trabajo seminal sentó las bases para los futuros desarrollos en inteligencia artificial y redes neuronales, al establecer un puente entre la biología y la computación (Piccinini, 2004). Al presentar estas ideas, Pitts y McCulloch no solo influyeron en el entendimiento de cómo el cerebro procesa la información, sino que también contribuyeron a forjar el camino hacia la creación de sistemas de inteligencia artificial que intentan emular este procesamiento.

Sin embargo, estos modelos de redes neuronales no podían manejar la disyunción exclusiva (XOR), lo que Marvin Minsky y Seymour Papert, dos figuras de gran influencia en el desarrollo de la inteligencia artificial y la neurociencia computacional, señalaron en 1969. En su emblemático

libro “Perceptrones” (1969), Minsky y Papert exploraron las limitaciones de las redes neuronales de la época, concretamente los perceptrones de una sola capa, demostrando que estos modelos solo podían resolver problemas linealmente separables y que no podían aprender ciertas funciones lógicas simples, como la función XOR (Minsky & Papert, 1969). Esta crítica provocó una reorientación en la investigación de la inteligencia artificial hacia los enfoques simbólicos, desacelerando temporalmente el desarrollo de las redes neuronales.

No obstante, también es importante señalar que su trabajo allanó el camino para las futuras mejoras de las redes neuronales. En la edición extendida de “Perceptrones” (1988), Minsky y Papert expresaron su confianza en que los perceptrones multicapa podrían superar las limitaciones que habían identificado anteriormente. Aunque este trabajo recibió menos atención en su momento, en retrospectiva, ha tenido un papel crucial en la evolución de la inteligencia artificial y las redes neuronales profundas que hoy están en el corazón de muchos sistemas de IA (Boden, 2006).

El progreso en este ámbito atrajo inversiones sustanciales, sin embargo, la incapacidad para cumplir con las expectativas de avance condujo a la era conocida como el primer “invierno de la IA”. Este término hace alusión a los ciclos recurrentes de descubrimientos, entusiasmo, aportaciones financieras, obstáculos, retraso en los progresos y desilusión que resultan en la reducción de la financiación, seguidos de un lapso de apatía hacia la IA debido a los desafíos asociados con la tecnología de la época.

En la década de 1980, la IA volvió a ganar interés con los “sistemas expertos” y las aplicaciones comerciales de la IA. El aprendizaje automático y las redes neuronales empezaron a ganar atención como una forma de hacer que las máquinas aprendan de los datos. Estos sistemas buscan emular el conocimiento de un experto en una materia mediante la creación y codificación de reglas, a lo que ahora llamamos “IA símbolica” o “IA basada en reglas”. Sin embargo, el entrenamiento de estas redes era un desafío debido a las limitaciones computacionales y a la falta de los grandes conjuntos de datos que existen hoy en día.

A pesar de su utilidad, los sistemas expertos tienen la desventaja de ser costosos de desarrollar y notablemente rígidos. Estas debilidades técnicas fueron criticadas por John McCarthy, y la falta de correlación entre los nuevos descubrimientos y las aplicaciones comerciales viables llevó al segundo invierno de la IA.

3.2.2. Resurgimiento de la Inteligencia Artificial

El auge contemporáneo de la inteligencia artificial (IA) se atribuye en gran medida a los avances en el aprendizaje automático (machine learning, ML). Estos métodos están basados en teorías biológicas de la cognición y la comunicación neuronal, introducidas por primera vez en las teorías de Daniel Hebb sobre la comunicación y el ordenamiento de las neuronas. (Hebb, 1949). Posteriormente han evolucionado a lo largo del tiempo a través de trabajos pioneros como los de Arthur Samuel y Frank Rosenblatt, y han contribuido a la formación de las actuales redes neuronales artificiales (ANN). Estas redes se han beneficiado considerablemente de los algoritmos de retro propagación, que permiten a las redes aprender a partir de errores al retroalimentar estos a las capas previas de la red.

De acuerdo a lo que señaló Arthur Samuel, una figura precursora en el ámbito, el aprendizaje automático se describiría como el "Ámbito de investigación que otorga a las máquinas la habilidad para aprender sin una programación específica". No obstante, este es una interpretación reduccionista de las ideas centrales detalladas en su escrito, que sugería que el aprendizaje a partir de la experiencia debía eventualmente hacer innecesario un minucioso trabajo de programación, haciendo una clara alusión a las estrategias basadas en sistemas expertos y la "ingeniería del conocimiento" (Samuel, 1959). En esencia, el objetivo es resaltar la posibilidad de automatizar el aprendizaje, lo cual se logra mediante el análisis estadístico de los datos para construir el modelo, un patrón que sirve como plantilla para soluciones futuras.

En resumen, el aprendizaje automático representa una metodología que engloba un conjunto de técnicas que se apoyan en métodos que, mediante el uso intensivo de datos, habilitan a las máquinas para automatizar la creación y programación de modelos, a través de la identificación sistemática de patrones con significancia estadística en los datos disponibles (Chowdhury, Apon, and Dey, 4 de abril de 2017).

La esencia del ML es su capacidad para automatizar el aprendizaje a través de la exploración estadística de datos para construir modelos abstractos que proporcionen soluciones innovadoras. En lugar de programar explícitamente cada acción, los algoritmos de ML buscan patrones significativos en los datos que les permitan tomar decisiones automáticamente, clasificar nuevos

datos, predecir tendencias y en algunos casos, generar nuevos datos (Chowdhury, Apon, and Dey, 2017). Sin embargo, es crucial tener en cuenta que los datos utilizados en ML son fundamentales para la eficacia de los modelos generados. Por lo tanto, se necesita un volumen suficiente de datos de alta calidad para extraer correlaciones significativas, evitando sesgos que puedan llevar a modelos discriminatorios o inexactos.

Los enfoques de aprendizaje en ML se pueden clasificar en tres categorías: aprendizaje supervisado, no supervisado y por refuerzo.

1. Aprendizaje Supervisado: En este tipo de aprendizaje, a los algoritmos de ML se les proporcionan conjuntos de datos de entrada y salida correspondientes, conocidos como datos de entrenamiento. El objetivo de estos algoritmos es aprender a mapear entradas a salidas correctas. Una vez que se ha establecido este mapeo a través del entrenamiento, el algoritmo puede aplicarlo a nuevos datos de entrada para generar salidas correspondientes. Los ejemplos de aplicaciones incluyen clasificación de imágenes y predicción de tendencias de ventas.
2. Aprendizaje No Supervisado: Aquí, los algoritmos de ML reciben conjuntos de datos de entrada sin ninguna salida correspondiente. El objetivo del algoritmo es encontrar estructura o patrones en los datos. Por ejemplo, podría agrupar los datos en diferentes categorías basándose en similitudes. Un ejemplo común de aprendizaje no supervisado es el algoritmo de agrupamiento k-means que se utiliza para segmentar a los clientes en categorías distintas.
3. Aprendizaje por Refuerzo: Este es un tipo de aprendizaje que se basa en la idea de que las estrategias, también conocidas como políticas, deben ser recompensadas si llevan a un resultado deseado. En otras palabras, los algoritmos de aprendizaje por refuerzo aprenden de los errores y éxitos a lo largo del tiempo. Un ejemplo popular de esto es el programa AlphaGo de DeepMind que superó a los campeones mundiales en el juego de mesa Go.

Todos estos enfoques se basan en la idea fundamental de utilizar grandes volúmenes de datos para crear modelos precisos y útiles que puedan realizar tareas de forma autónoma, a menudo mejorando a lo largo del tiempo a medida que se les proporciona más datos y feedback. Sin embargo, también se enfrentan a desafíos, como el manejo de sesgos, el aseguramiento de la

calidad y representatividad de los datos y el equilibrio entre la identificación de tendencias significativas y la evitación de sobreajustes y subajustes.

3.2.3. Nueva Primavera de la IA

El principio de la década 2010 marcó un punto de inflexión en la historia de la IA y la IAG. El avance de los hardware especializados y el aumento en la disponibilidad de datos revolucionó las aplicaciones de redes neuronales y algoritmos de aprendizaje profundo. Estas empezaron a demonstrar resultados impresionantes en cuanto al reconocimiento de voz e imágenes. El primer ejemplo significativo de IAG vino con la introducción de las Redes Generativas Antagónicas (GANs) por Ian Goodfellow y sus compañeros en 2014 (Goodfellow et al., 2014). Este modelo utilizó dos redes neuronales, un generador y un discriminador, en un juego adversario para generar imágenes realistas. Desde entonces, hemos visto avances impresionantes en la IAG. Los modelos autorregresivos, como GPT-3 (GPT-4 recientemente) de OpenAI, pueden generar texto coherente y a menudo indistinguible de la escritura humana (Brown et al., 2020). Las GANs han sido utilizadas para generar arte original, deepfakes y mucho más.

El impacto significativo de los datos y el crecimiento masivo de Internet se pueden considerar como catalizadores cruciales en la reactivación de las técnicas de Machine Learning (ML), superando las restricciones inherentes que anteriormente limitaban su aplicación. En particular, el Internet de las cosas (IoT) y el auge de los dispositivos inteligentes de recopilación de datos han sido factores determinantes en la ampliación de la economía basada en datos, proporcionando la cantidad necesaria de información para alimentar estas técnicas. Sumado a la accesibilidad de la información, otros aspectos relacionados con el hardware, como la reducción de los costos de almacenamiento y el aumento de la capacidad de procesamiento acorde a la Ley de Moore, han facilitado este avance. Esto se refiere a una observación realizada por Gordon Moore, cofundador de Intel, en 1965. Moore notó que el número de transistores en un circuito integrado denso se duplicaba aproximadamente cada dos años. Este crecimiento exponencial ha sido la base del rápido aumento en el rendimiento de las computadoras y otros dispositivos digitales durante las últimas décadas.

En el presente, se discute la inteligencia artificial como la cuarta revolución industrial. Sin embargo, aún queda por ver si estos progresos se mantendrán de manera sostenida y efectiva como una ciencia aplicada, o si la disciplina encontrará un límite que conduzca a otro período de recesión o "invierno" para la IA. Esto parece poco probable a día de hoy, de hecho, muchos expertos se preocupan por el avance expedito de estas tecnologías. En abril de 2023, Business Insider público un artículo donde decía "Elon Musk, CEO de Twitter, junto a líderes empresariales de la talla de Steve Wozniak, cofundador de Apple, firmaron un manifiesto la semana pasada que pedía una moratoria de 6 meses en el desarrollo de la inteligencia artificial más avanzada.". Se teme que tanto desarrollo tecnológico en tan poco tiempo puede ser destructivo para la sociedad, un tema ético que se explorará a lo largo de este trabajo.

3.3 Los Tipos de Inteligencia Artificial

La Inteligencia Artificial es un campo enormemente diverso, compuesto por una variedad de tipos y subdisciplinas. Cada uno de estos tipos de IA tiene su conjunto de características, metodologías y aplicaciones únicas, contribuyendo a la complejidad y amplitud de la IA como campo de estudio y práctica. Al explorar los diferentes tipos de IA; como la IA débil, la IA general, los sistemas de aprendizaje automático, las redes neuronales y el deep learning, y finalmente la IA con capacidad generativa; nos adentramos en las complejidades y la riqueza de este fascinante campo. En esta sección analizaremos estos diversos tipos, su naturaleza única y sus aportaciones al campo de la IA en general. Estos distintos tipos de IA no solo nos permiten entender mejor su funcionamiento actual, sino que también nos dan una visión de su potencial y los retos futuros que podrían plantear.

3.3.1. La Inteligencia Artificial Débil o Especializada

La Inteligencia Artificial Débil (ANI, por sus siglas en inglés) se especializa en realizar tareas específicas con un alto nivel de competencia. También se le conoce como IA estrecha, limitada o especializada, y su desempeño está condicionado por su programación y los datos de entrenamiento. Los sistemas ANI, a diferencia de los humanos, no pueden adaptarse a nuevas situaciones o generalizar el conocimiento entre contextos sin una intervención humana.

Estos sistemas pueden clasificarse en dos tipos: los de aprendizaje supervisado y no supervisado. Los de aprendizaje supervisado se entrenan en conjuntos de datos etiquetados para aprender la relación entre la entrada y la salida deseada, y los de aprendizaje no supervisado, que se entrenan en conjuntos de datos no etiquetados y pueden identificar patrones y relaciones en los datos sin orientación.

La ANI es común en nuestras vidas, aunque no siempre estemos atentos a su presencio. Asistentes virtuales como Siri de Apple o Alexa de Amazon, sistemas de traducción como Google Translate o Deepl, y sistemas de diagnóstico médico son ejemplos de ANI. Pese a su alta sofisticación, estos sistemas se consideran “débiles” o “limitados” porque no pueden igualar la inteligencia humana en su totalidad. Sin embargo, son muy eficaces en la realización de tareas específicas, a menudo superando a los humanos en velocidad y precisión en cuanto a los resultados que proporciona.

La ANI se divide en dos categorías: IA simbólica y aprendizaje automático. La IA simbólica, también conocida como Buena y Antigua IA (GOFAI), sigue reglas definidas explícitamente por los programadores. Por otro lado, el aprendizaje automático se basa en el entrenamiento mediante ejemplos y utiliza estos ejemplos para predecir resultados y clasificar tareas.

Las aplicaciones de la ANI son diversas, desde sistemas de reconocimiento facial y asistentes virtuales hasta vehículos autónomos y motores de recomendación. A pesar de sus limitaciones, la ANI es una herramienta poderosa que continúa impactando profundamente en muchos aspectos de la vida diaria y profesional.

3.3.2. La Inteligencia Artificial Fuerte o General

La Inteligencia Artificial Fuerte, también llamada Inteligencia Artificial General (IAG) o IA General, es un tipo de IA teórica con una inteligencia y autoconciencia equiparable a la de los humanos. Esta forma de IA puede resolver problemas, aprender y planificar el futuro. A diferencia de la Inteligencia Artificial débil, que está especializada en tareas concretas, la IA fuerte puede realizar múltiples tareas a un nivel muy competente, similar a como lo haría un humano.

La creación de una inteligencia artificial general, es decir, una IA que sea igual o más que los seres humanos posiblemente sea la mayor ambición científica de la historia de la humanidad. (Clune,

2019). Sin embargo, hasta la fecha, la IA Fuerte sigue siendo un concepto teórico, no una realidad tangible. Aunque hay mucho interés en su desarrollo, aún no se ha alcanzado un consenso sobre cómo se definiría y mediría su éxito.

En 1950 se creó el Test de Turing, desarrollado por el famoso científico Alan Turing. El propósito de este test era para evaluar si el comportamiento de un maquina se puede distinguir al de un ser humano. En el test, un humano intenta determinar si las respuestas a una serie de preguntas provienen de un humano o una máquina. Si no puede distinguir entre los dos, la máquina se considera que ha pasado el test.

Sin embargo, hay críticos de este método, como John Searle, quien propuso el Argumento de la Habitación China (CRA). La siguiente cita resume bien su argumento, "Los cálculos se definen de forma puramente formal o sintáctica, mientras que las mentes tienen contenidos mentales o semánticos reales, y no podemos pasar de lo sintáctico a lo semántico simplemente teniendo las operaciones sintácticas y nada más...Un sistema, yo, por ejemplo, no adquirirá una comprensión del chino simplemente siguiendo los pasos de un programa informático que simule el comportamiento de un hablante de chino" (Searle, 1980). Aunque una máquina pueda proporcionar respuestas correctas en un idioma que no comprende mediante el uso de un manual de instrucciones, eso no significa que la máquina entienda de verdad el idioma. En otras palabras, la capacidad para simular la comprensión no equivale a la verdadera comprensión.

Además de la IA fuerte, se ha propuesto otra forma de IA: la Superinteligencia Artificial (ASI). Esta forma de IA iría más allá de la IA fuerte, superando la inteligencia y la capacidad humanas. Sin embargo, al igual que la IA fuerte, la ASI sigue siendo especulativa.

A pesar de las limitaciones actuales, el campo de la IA está innovando rápidamente, y la IA tiene un papel cada vez más importante en campos como la ciberseguridad, el entretenimiento y la creación de contenidos, y el reconocimiento y predicción de comportamientos. Sin embargo, estos avances también plantean cuestiones éticas, como los sesgos implícitos y la necesidad de una IA responsable, cuyas cuestiones se explorarán más a fondo a lo largo de este trabajo.

3.3.3. Los Sistemas de Aprendizaje Automático

Los sistemas de aprendizaje automático, ML por sus siglas en inglés (Machine Learning) son sistemas informáticos que utilizan modelos matemáticos de datos para aprender sin recibir instrucciones directas. Estos sistemas usan algoritmos para identificar patrones en los datos, que luego se convierten en modelos de datos capaces de realizar predicciones. A medida que estos sistemas se alimentan de más experiencia y datos, mejoran en precisión, de manera similar a cómo los humanos aprenden y mejoran con la práctica.

El aprendizaje automático (ML) se con otros campos, se considera un subconjunto de la IA, con algunos de los sistemas más avanzados imitando el razonamiento humano utilizando redes neuronales modeladas a partir del cerebro humano. También está relacionado con el análisis predictivo, aunque a diferencia de este último, que suele trabajar con un conjunto de datos estático, el ML puede actualizarse en tiempo real a medida que recibe más datos. El aprendizaje profundo, por otro lado, es una forma especializada de ML que utiliza redes neuronales para proporcionar respuestas y determinar su precisión (Microsoft Azure).

Las aplicaciones y beneficios del ML son numerosos y están en constante crecimiento. Algunas de las ventajas más destacadas incluyen la capacidad de descubrir información a través de la identificación de patrones en datos estructurados y no estructurados, mejorar la integridad de los datos a través de la minería de datos, mejorar la experiencia del usuario, reducir el riesgo de fraude, predecir el comportamiento del cliente y reducir costos a través de la automatización de procesos (Microsoft).

El ML utiliza diversas técnicas, como el aprendizaje supervisado, donde se utiliza un conjunto de datos etiquetado para “entrenar” al sistema; el aprendizaje no supervisado, que busca los patrones y las relaciones que existen en conjuntos de datos no estructurados y no etiquetados; y el refuerzo del aprendizaje, que utiliza un agente para determinar resultados basados en un bucle de retroalimentación.

La implementación de ML implica un proceso de cuatro pasos: recopilación y preparación de datos, entrenamiento del modelo, validación del modelo y finalmente la interpretación de los

resultados. Los ingenieros de ML juegan un papel crucial en este proceso, trabajando para convertir datos sin procesar en modelos de ciencia de datos que se pueden aplicar y escalar según sea necesario.

El ML tiene una gran cantidad de aplicaciones en varios sectores. En el sector financiero, por ejemplo, se utiliza para la gestión de riesgos y la prevención de fraudes. En atención sanitaria, puede ayudar en el diagnóstico, la monitorización de pacientes y la predicción de ataques. En el sector de transporte, puede optimizar rutas de entrega y facilitar el desarrollo de vehículos autónomos. En atención al cliente, puede mejorar la interacción con los clientes al responder a preguntas y proporcionar asistencia virtual. En el comercio minorista, puede optimizar precios y ofertas, y en la agricultura, puede mejorar la gestión de la mano de obra y la supervisión del estado del suelo (Microsoft).

3.3.4. Las Redes Neuronales y Aprendizaje Profundo (Deep Learning)

El aprendizaje profundo, normalmente llamado deep learning, es una técnica de inteligencia artificial que utiliza redes neuronales artificiales para emular la toma de decisiones humana. Esta tecnología puede aprender y mejorar su rendimiento a través de la experiencia, procesando grandes cantidades de datos para realizar predicciones precisas.

En ciertas circunstancias, es posible usar una estrategia llamada aprendizaje por transferencia, que permite aplicar conocimientos adquiridos en un problema a otro similar. Esto permite reducir la cantidad de tiempo, datos y recursos necesarios para entrenar un nuevo modelo, lo cual resulta especialmente útil cuando se carece de recursos o datos de entrenamiento abundantes.
El aprendizaje profundo es eficaz para identificar patrones en datos no estructurados, como imágenes, sonido, vídeo y texto. Por lo tanto, se está utilizando para transformar diversos sectores como la atención sanitaria, la energía, las finanzas y el transporte, reconfigurando procesos empresariales tradicionales.

El aprendizaje profundo tiene una variedad de aplicaciones, incluyendo:

1. **Reconocimiento de entidades con nombre**: Esta aplicación permite identificar y clasificar piezas específicas de información dentro de un texto, lo que puede ser útil para diversas tareas, como la creación de listas de direcciones o la validación de identidades.
2. **Detección de objetos**: Esta técnica implica identificar y localizar objetos específicos en imágenes. Se está utilizando en diversas áreas, incluyendo videojuegos, comercio minorista, turismo y vehículos autónomos.
3. **Generación de subtítulos para imágenes**: Esta tecnología puede etiquetar objetos en fotografías y convertir estas etiquetas en oraciones descriptivas, proporcionando una descripción coherente de las imágenes.
4. **Traducción automática**: El aprendizaje profundo puede traducir texto, audio y señales visuales de un idioma a otro, y transcribir la palabra hablada o la imagen como texto.
5. **Análisis de texto**: El aprendizaje profundo puede analizar grandes cantidades de datos de texto, identificar patrones y generar información organizada y concisa. Se usa para diversas aplicaciones, como la detección de negociaciones en el mercado de valores, el cumplimiento normativo gubernamental y la detección de fraudes en los seguros.

El aprendizaje profundo se basa en redes neuronales artificiales, que pueden ser de varios tipos, incluyendo:

1. **Redes neuronales de tipo feedforward**: El tipo más simple de red neuronal artificial, en las que la información se desplaza solo en una dirección, desde la capa de entrada a la de salida.
2. **Redes neuronales recurrentes (RNN)**: Estas redes almacenan y reenvían la salida de una capa a la capa de entrada, y son útiles para tareas como la predicción de series temporales, el aprendizaje de escritura a mano y el reconocimiento de idiomas.
3. **Redes neuronales convolucionales (CNN)**: Son especialmente eficaces en tareas como el reconocimiento de vídeo, el reconocimiento de imágenes y los sistemas de recomendación.
4. **Redes generativas antagónicas (GAN)**: Se utilizan para crear contenido original y realista, como imágenes y texto, y se compone de dos redes que compiten entre sí, el generador y el discriminador.

El aprendizaje profundo y las redes neuronales artificiales están ayudando a impulsar avances significativos en inteligencia artificial, abriendo nuevas posibilidades y aplicaciones en una variedad de campos y sectores.

3.3.5. La Inteligencia Artificial Generativa (IAG)

La Inteligencia Artificial Generativa es la rama de la IA que se utiliza para crear contenido completamente original, y es el último subcampo de la IA que se explorará a lo largo de este trabajo. La IAG funciona con sistemas de computación capaces de generar contenido nuevo como imágenes, texto, y música entre otros. Su habilidad de emular el razonamiento y pensamiento humano es verdaderamente sorprendente, y algo que ilusiona, pero que a la vez preocupa a los expertos de este campo. Se plantea la pregunta ¿Cómo vamos a controlar el desarrollo de estas tecnologías para beneficiar al ser humano?

La IAG está compuesta por dos redes; una generativa y otra discriminativa. La red generativa se encarga de producir datos sintéticos (como imágenes, sonido o texto) a partir del ruido aleatorio, mientras que la red discriminativa se encarga de determinar si los datos que se le presentan son reales o han sido generados por la red generativa. Cuando se entrenan, estas dos redes están en constante competencia; la red generativa trabaja en mejorar su capacidad para crear datos que parezcan ser reales, mientras que la red discriminativa trabaja en mejorar su capacidad para distinguir entre los datos que parezcan reales y los que lo son de verdad. A medida que la red generativa mejora, la red discriminativa debe adaptarse y mejorar para mantenerse al día, y viceversa.

Esta tecnología se puede aplicar a una gran variedad de campos, como;

1. **Creación de imágenes, arte y música**: Las GANs se han utilizado para crear obras de arte nuevas y únicas, como música, pinturas, y muchas más. Un ejemplo notable es el cuadro “Portrait of Edmond de Belamy”, que fue creado por una GAN y vendido en Christie's en Nueva York por casi medio millón de dólares, destrozando su estimación de diez mil dólares.
2. **Creación de texto coherente**: Las GANs también se han utilizado para generar texto coherente. GPT-4, el modelo de lenguaje natural desarrollado por la empresa OpenAI, es

es el ejemplo más de moda de esta aplicación. El sistema es capaz de redactar ensayos, responder preguntas, traducir idiomas, y más, todo con una coherencia y fluidez que asimilan mucho la del ser humano.

3. **Modelado 3D y diseño de productos**: Las GANs se utilizan en el modelado 3D y en el diseño de productos, donde pueden ayudar a generar una variedad de diseños potenciales basándose en los parámetros y criterios de diseño ingresados por el usuario, típicamente un humano.
4. **Sintetización de voz**: En el campo de la síntesis de voz, las GANs pueden generar voces sintéticas que suenan muy similares a las voces humanas. Esto tiene aplicaciones potenciales en la asistencia virtual, el doblaje de películas, los audiolibros y más. Esto también está teniendo repercusiones con el incremento de "fakes" que estan saliendo a la luz, algo que puede llevar un problema serio sobre el uso ético de estas tecnologías.

3.4. Porque la Inteligencia Artificial Generativa

Habiendo analizado los diferentes tipos de Inteligencia Artificial, he decidido enfocar el trabajo a la Inteligencia Artificial Generativa (IAG). Dada a la naturaleza de este trabajo, la IA en su totalidad es un campo demasiado extenso como para hacer una investigación sobre él. Debido a esto, he seleccionado el subcampo que más relevancia tiene a día de hoy. La IAG tiene alta capacidad de ser disruptiva en una gran cantidad de industrias. A parte, las noticias sobre esta nueva tecnología son abundantes, aludiendo a su rápido crecimiento exponencial que cambiará nuestra sociedad como la conocemos.

La IAG se centra en la creación de algo nuevo, en contraste con los tipos más tradicionales de IA que se centran en la identificación y clasificación de la información. Esta es una de las razones por las que es tan emocionante; en lugar de simplemente usar la IA para entender nuestro mundo, estamos comenzando a usarla para añadir a él. Es una herramienta que no sólo interpreta y procesa, sino que también imagina y crea.

Enfocar el trabajo hacía la IAG no significa que los otros tipos de IA fueran menos importantes o interesantes; el progreso de estos es esencial para el progreso en todas las áreas de la IA, pero hoy

la IAG tiene la mayor importancia en sus aplicaciones. Además, es importante reconocer que la IAG tiene el potencial de ser profundamente disruptiva. Por ello, es vital que comprendamos sus implicaciones y las maneras en las que se está desarrollando. Las preguntas sobre la ética de la IA generativa, su impacto en el empleo y la economía, y cómo debería ser regulada son de vital importancia. Por esto, este subcampo de la IA es un tima digno de una investigación detallada.

4. Técnicas Generativas: Inteligencia Artificial Generativa

“When you invent a new technology, you uncover a whole new class of responsibilities.” - Center for Human Technology

4.1. Que es la Inteligencia Artificial Generativa

La Inteligencia Artificial generativa (IAG) se refiere a la rama de la Inteligencia Artificial que se enfoca en crear sistemas de computación para generar contenido original y autónomo, como imágenes, texto, y música. A diferencia de otros tipos de Inteligencia Artificial cuya labor es solucionar problemas específicos, la IA generativa intenta imitar la capacidad de razonamiento normalmente atribuida a un ser humano, y así generar contenido que no se haya creado con esta creatividad humana. Estos híper-sistemas de computación se entrenan a través del aprendizaje automático utilizando un gran corpus de datos, que luego utilizan para generar contenido nuevo. Es el subcampo de la IA más emocionante actualmente, contando con descubrimientos pioneras casi semanalmente desde el inicio de 2022.

El crecimiento de la IAG está siendo impresionante en los últimos años, alcanzando cotas jamás previstas por los expertos de la inteligencia artificial. En 2021 se llevó a cabo un estudio en el que miles de pronosticadores profesionales, expertos en la inteligencia artificial, fueron cuestionados sobre la capacidad resolutiva de la IA (Steinhardt). El estudio se enfoca en la evaluación de dos conjuntos de datos para inteligencia artificial, que son MATH y Multitask. MATH es un conjunto de problemas matemáticos competitivos de nivel secundario y Multitask es un conjunto de datos más amplio que abarca una variedad de tareas. Enfocándonos en los problemas matemáticos, una de las preguntas planteadas fue “¿Cuándo será capaz la IA de resolver problemas matemáticos de competición (competition-level) con una precisión superior al 80 %?” (Steinhardt). La predicción ganadora estimaba que la IA podrá alcanzar una precisión del 52% para 2025, es decir en cuatro años, en cuanto a su capacidad de generar la respuesta correcta a este tipo de problemas. Sin embargo, la IA tardó menos de un año en cumplir este reto, aludiendo a no sólo la capacidad de

aprendizaje que tienen estos sistemas, sino a la incertidumbre profesional que los rodea. A pesar de la emoción que están generando estos avances a los expertos de la IA, la IAG sigue siendo un campo activo de investigación con muchas preguntas abiertas y desafíos por resolver.

La IAG tiene un gran potencial en una variedad de campos. En el diseño gráfico y la producción de películas, por ejemplo, podría utilizarse para generar automáticamente gráficos, efectos visuales o animaciones (Elgammal, Liu, Elhoseiny, & Mazzone, 2017). En la industria de la música, la IAG puede ser utilizada para componer nuevas melodías, armonías o incluso canciones completas (Briot, Hadjeres, & Pachet, 2020). Además, en el periodismo y la escritura, la IAG puede ser utilizada para generar automáticamente artículos o historias, como las noticias generadas por bots en sitios web de noticias (Liu, 2019).

La IAG se basa en modelos como las Redes Generativas Antagónicas (GANs); tambien llamadas Redes Adversariales Generativas, las Máquinas de Boltzmann, los Autoencoders Variacionales (VAEs) y los Modelos Generativos Latentes de Lógica de Múltiples Modalidades (GLLMM). Cada modelo utiliza estrategias y técnicas para generar contenido, pero todos comparten el objetivo de producir contenido coherente, de alta calidad e indistinguible de los contenidos creados por humanos.

Pero la IAG también plantea desafíos éticos y legales significativos. Por ejemplo, las cuestiones de derechos de autor pueden surgir cuando estos sistemas generan contenido que se asemeja demasiado a obras protegidas por derechos de autor. Además, la posibilidad de que se generen noticias o información falsa plantea problemas de desinformación (Chesney & Citron, 2019).

La IAG es un área emocionante de la IA con un enorme potencial, pero que requiere una consideración cuidadosa de las implicaciones éticas y legales. Es importante, y la cuestión de este trabajo, saber cómo posicionar al ser humano en el centro de todos estos avances tecnológicos. Es decir, construir y desarrollar estas tecnologías con intención de ayudar al ser humano de cara al futuro. Esto conlleva la discusión sobre la automoción y el riesgo de los trabajadores, sobre la originalidad normalmente atribuida a un ser humano, etc. Estas cuestiones se estudiarán a lo largo de este trabajo.

4.2. Cómo funciona la IAG

Los modelos de inteligencia artificial generativa (IAG) aprenden los patrones y la estructura de sus datos de entrenamiento de entrada, y luego generan nuevos datos que tienen características similares.

La IA Generativa ha mostrado aplicaciones potenciales en una amplia gama de industrias, incluyendo el arte, la escritura, el desarrollo de software, la atención sanitaria, las finanzas, los juegos, el marketing y la moda. Ha habido un aumento en la inversión en IA generativa durante la década de 2020, con grandes empresas como Microsoft, Google y Baidu, así como numerosas empresas más pequeñas, desarrollando modelos de IA generativa.

Históricamente, el campo del aprendizaje automático ha utilizado modelos estadísticos, incluyendo modelos generativos, para modelar y predecir datos. A finales de los 2000, el aprendizaje profundo impulsó el progreso y la investigación en el procesamiento de imágenes y vídeos, el análisis de texto, el reconocimiento de voz y otras tareas. En 2017, la red Transformer permitió avances en los modelos generativos, dando lugar al primer Transformador Pre-entrenado Generativo (GPT) introducido por la empresa OpenAi en 2018

La IA generativa puede ser unimodal o multimodal; los sistemas unimodales toman solo un tipo de entrada (por ejemplo, texto) mientras que los sistemas multimodales pueden tomar más de un tipo de entrada (por ejemplo, texto e imágenes). Ejemplos de sistemas de IA generativa entrenados en palabras o tokens de palabras incluyen GPT-3, LaMDA, LLaMA, BLOOM, GPT-4, entre otros. Estos son capaces de procesamiento de lenguaje natural, traducción automática y generación de lenguaje natural y pueden usarse como modelos base para otras tareas 7.

4.3. Como se entrenan

Entrenar un modelo de Inteligencia Artificial Generativa (IAG) implica la aplicación de diversas técnicas de aprendizaje automático, dependiendo del tipo específico de modelo generativo que se esté utilizando. Este proceso no es uniforme y puede variar considerablemente dependiendo de la naturaleza del modelo generativo específico.

Los modelos de IAG, como las redes generativas antagónicas (GANs), los autoencoders variacionales, y los GLLMM (Generative Large Language Multi-Modal Model), entre otros, tienen cada uno sus propios enfoques y desafíos de entrenamiento. Por ejemplo, las GANs involucran una interacción competitiva entre dos redes neuronales, una generadora y una discriminadora, que trabajan en conjunto para producir resultados realistas. Por otro lado, los GLLMM aprenden patrones y estructuras lingüísticas, permitiéndoles generar oraciones, párrafos e incluso documentos enteros que se leen de manera natural y coherente.

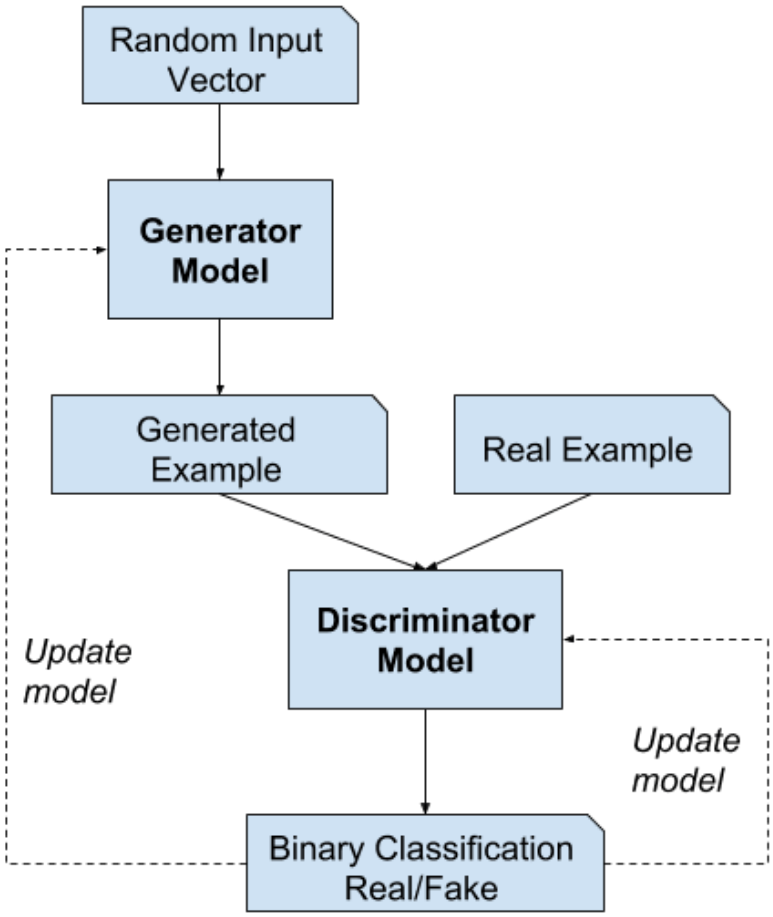
Además, cada uno de estos modelos puede requerir la aplicación de técnicas especializadas de optimización, regularización y ajuste de hiperparámetros para garantizar un aprendizaje eficaz y resultados de alta calidad. A medida que avanzamos en la comprensión de estos distintos modelos de IAG, es esencial tener en cuenta que cada uno de ellos representa una faceta única de la IAG y que las técnicas de aprendizaje automático aplicables pueden variar considerablemente de un tipo a otro (Goodfellow, 2014).

4.3.1 Las Redes Adversariales Generativas (GANs)

Una de las técnicas más populares de IAG son las Redes Adversariales Generativas, GANs por sus iniciales en inglés (Generative Adversarial Nets). En una GAN, hay dos componentes principales: un generador y un discriminador, que se entrenan juntos en un proceso iterativo. El generador crea nuevas instancias a partir de un espacio latente y el discriminador evalúa si estas instancias son reales (tomadas del conjunto de entrenamiento) o falsas (creadas por el generador). Este proceso de entrenamiento, que se asimila a un juego de dos jugadores, se lleva a cabo hasta que el generador se vuelve lo suficientemente bueno como para engañar al discriminador, creando instancias indistinguibles de las reales (Goodfellow et al., 2014).

Las Redes Adversariales Generativas (GANs) son una clase de algoritmo de IA diseñado para resolver el problema del modelado generativo. El objetivo de un modelo generativo es estudiar una colección de ejemplos de entrenamiento y aprender la distribución de probabilidad que los generó. Las GANs son capaces de generar más ejemplos a partir de la distribución de probabilidad estimada. Estas han sido aplicadas con éxito a una amplia variedad de tareas, pero continúan presentando desafíos y oportunidades de investigación únicos porque se basan en la teoría de

juegos, mientras que la mayoría de los otros enfoques de modelado generativo se basan en la optimización.



En esencia, las Redes Adversarias Generativas (GANs) se entrenan para aprender la distribución de probabilidad inherente a un conjunto de datos, añadiendo variaciones aleatorias (ruido) para prevenir que los contenidos producidos sean una réplica exacta de los datos de entrenamiento. En este proceso, el generador selecciona valores aleatorios de entrada para convertirlos en contenido nuevo. No obstante, para poder confundir al discriminador, la red generadora también intenta producir datos que aparenten originarse de la misma distribución de probabilidad que los datos de entrenamiento.

Por otro lado, el discriminador actúa como un clasificador, su labor es distinguir si el contenido es auténtico, es decir, proveniente del conjunto de datos, o si son datos sintéticos creados artificialmente.

Durante cada ciclo de esta competencia (de ahí el término “adversarial”), ambas redes neuronales evolucionan, una creando datos cada vez más similares a los "reales", mientras que la otra se perfecciona en el arte de identificar las falsificaciones.

En este enfrentamiento mutuo, ambas aprenden y se mejoran, culminando el entrenamiento cuando el discriminador ya no puede distinguir entre lo real y lo fabricado.

A diferencia de otras técnicas que consisten en alterar la información original (manipulación de imágenes, audio, video, etc.), lo destacable es que el contenido completo generado por estas redes es completamente sintético.

4.3.2. Los Autoencoders Variacionales

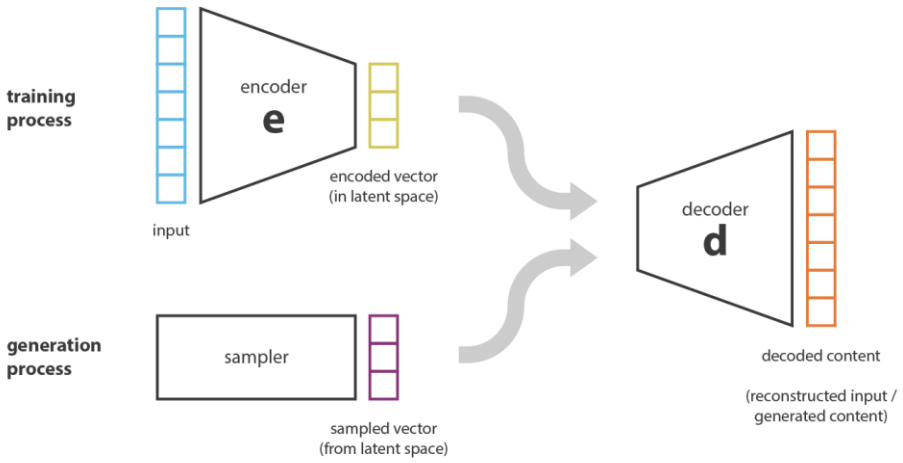
Los Autoencoders Variacionales, o VAEs por sus iniciales en inglés (Variational Auto Encoders) son un tipo especial de modelo de aprendizaje automático que tiene un papel importante en el campo de la Inteligencia Artificial Generativa (IAG). Son especialmente útiles para tareas como la generación de imágenes y la modelización de datos complejos.

Los VAEs son parte de una clase de redes neuronales conocida como autoencoders. Un autoencoder es una red neuronal que se entrena para copiar su entrada a su salida. Para lograr esto, tiene una capa oculta que describe un código utilizado para representar la entrada. Un autoencoder variacional difiere de un autoencoder estándar en la forma en que se diseñan y entrenan. En lugar de aprender una representación fija de la entrada, un VAE aprende los parámetros de una distribución de probabilidad que representa la entrada (Kingma & Welling, 2013).

Este modelo consta de dos componentes principales: un codificador que toma los datos de entrada y los transforma en un conjunto de parámetros que definen una distribución de probabilidad, y un decodificador que toma muestras aleatorias de esta distribución y las usa para reconstruir la entrada. Durante el entrenamiento, el codificador y el decodificador trabajan juntos para minimizar la diferencia entre los datos de entrada y los datos reconstruidos, así como para asegurar que la

distribución de probabilidad esté cerca de una distribución priora, generalmente una distribución normal estándar (Kingma & Welling, 2013).

En términos de aplicaciones éticas de los VAEs, hay varias cuestiones a tener en cuenta. Como muchos modelos de IAG, los VAEs pueden ser utilizados para generar contenido falso o "deepfakes". También existen problemas relacionados con la privacidad de los datos y el sesgo en los modelos entrenados.



4.3.3. Las Redes Adversariales Creativas (CANs)

En el ámbito de la creación de software inventivo, existen múltiples enfoques propuestos. Uno de estos métodos involucra el uso de algoritmos genéticos, los cuales funcionan mediante la generación iterativa de candidatos que se evalúan a través de una función de ajuste. Esta función se modifica posteriormente en función del rendimiento de los candidatos superiores (Holland, 1975).

Estos sistemas creativos evolutivos surgen de la computación evolutiva y representan una categoría de algoritmos de búsqueda que se inspiran en la evolución darwiniana. Los algoritmos genéticos (GA) y la programación genética (GP) son los ejemplos más populares de estas técnicas.

Estos enfoques resuelven problemas complejos codificando una población de posibles soluciones generadas de manera aleatoria como “conjuntos de instrucciones genéticas”. Luego, evalúan la habilidad de cada uno para resolver el problema mediante una función de ajuste predefinida, mutando y/o cruzando lo mejor para producir una nueva generación. Este proceso se repite hasta que uno de los descendientes produce una solución aceptable (De Jong, 2006).

Tradicionalmente, se ha considerado que la intervención humana es necesaria para hacer una valoración artística de los artefactos por sus cualidades estéticas. Sin embargo, se ha demostrado que esta intervención humana puede ser reemplazada a través de una función de aptitud automática. En estos sistemas interactivos, la computadora explora el espacio creativo y el humano juega el papel de observador, cuya retroalimentación es esencial para impulsar el proceso (McCormack, Gifford, & Hutchings, 2019).

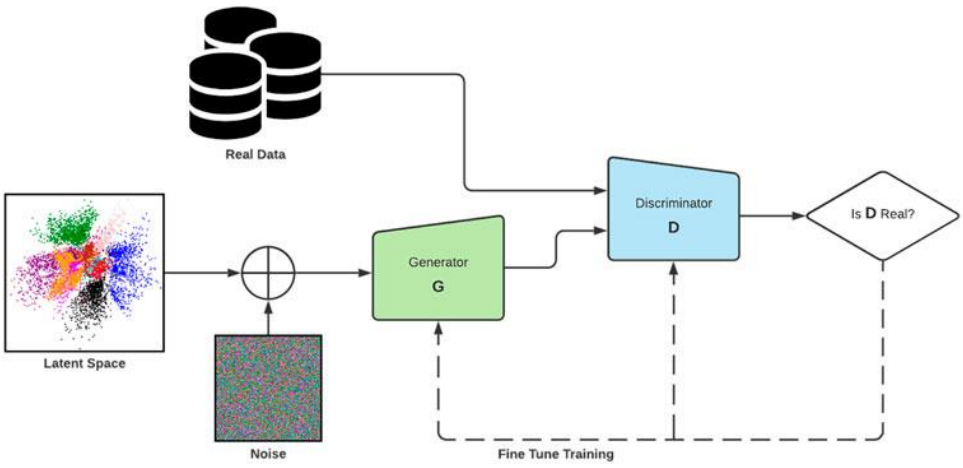
Elgammal y sus colaboradores (2017) enfatizan un aspecto esencial de los procesos de creatividad humana: la necesidad de una previa inmersión en el arte. Los artistas humanos han estado constantemente expuestos al trabajo de sus pares a lo largo de su vida. No obstante, la forma en que estas influencias se procesan y subliman a nivel cognitivo sigue siendo un misterio.

Pese a ser un terreno ampliamente no investigado, la jurisprudencia de derechos de autor en los Estados Unidos ha sentado un precedente sobre este fenómeno intrincado. En un caso particular, se emitió un fallo basado en una similitud “inconsciente” substancial en el caso de George Harrison y la canción “My Sweet Lord” (Bright Tunes Music Corp. v. Harrisongs Music, Ltd., 1976).

En su intento de elaborar un modelo que pueda ser emulado de manera computacional, Elgammal se apoya en las teorías de Martindale, que interpretan la creatividad desde un punto de vista psicológico. De acuerdo con estas teorías, los artistas buscan aumentar el potencial de excitación (arousal potential) de sus obras para evitar la repetición de lo cotidiano. Sin embargo, este aumento debe ser moderado para evitar respuestas adversas por parte de los espectadores. Un exceso de novedad provoca rechazo, mientras que muy poca novedad no suscita interés artístico (Martindale, 1990).

El concepto detrás de las CAN es el diseño de una estructura que percibe dos estímulos del discriminador, los cuales funcionan como fuerzas opuestas para lograr tres objetivos: la producción de trabajos innovadores, asegurando que dicha innovación no sea tan extrema como para causar

repulsión, y fomentando la amplitud del estilo en la obra. El objetivo final es crear trabajos que, a la vez, posean un valor artístico y sean distintos a los datos de entrenamiento utilizados.



4.4. GLLMM: Generative Large Multi-Modal Model (Gollem AIs) (LLM)

Gollem AIs - **“**In the jewish folklore, the idea of these innanimate objects that suddenly gain their own capacities, emerging capacities that were not in the innanimate clay with which you baked the product.” - Center for Human Technology

Los LLM (Large Language Models) son modelos de inteligencia artificial que han sido entrenados en grandes cantidades de texto. Su capacidad para producir contenido textual que parece notablemente humano se deriva de su aprendizaje a partir de patrones lingüísticos, conocimientos generales y ciertos razonamientos incluidos en los textos utilizados para su entrenamiento (DiarioBitcoin). El modelo más conocido mundialmente en este momento es GPT-4 de OpenAi, pero muchas otras empresas como Google, Facebook, Microsoft están sacando sus propios modelos con el ánimo de liderar esta nueva revolución industrial.

Los LLM producen resultados impresionantes y llegan a conclusiones de forma autónoma, y por eso su implementación en diversas industrias ha incrementado, incrementando así su aprendizaje y capacidad intelectual. Sin embargo, nuestra comprensión de cómo estos modelos de lenguaje funcionan internamente es aún limitada. Es un desafío descifrar por qué un modelo responde de

cierta manera a una situación específica, y es igualmente difícil determinar cómo estos modelos aumentan su conocimiento, si incorporan sesgos o si se involucran en el engaño.

Hay mucha incertidumbre que rodea las formas en las que se entrenan estos modelos, sobre todo a la hora de saber la procedencia los datos y si las respuestas que proporcionan los modelos son fiables o no. Al respecto, OpenAi ha lanzado una herramienta que permite analizar qué componentes del modelo son responsables de cada función y comportamiento.

Esta herramienta, como se detalla en un comunicado publicado en el sitio web de OpenAI, se basa en un proceso automatizado que utiliza el modelo de lenguaje GPT-4 para "generar y calificar explicaciones en lenguaje natural del comportamiento de las neuronas" de un LLM específico, en este caso, GPT-2, y posteriormente aplicarlo a las neuronas en otro modelo de lenguaje.

Los modelos LLM están compuestos por neuronas que reconocen patrones específicos en el texto que el usuario introduce para generar una respuesta correspondiente. Con esto en mente, la herramienta de OpenAI descompone el modelo GPT-2 en varios componentes y lleva a cabo un proceso de tres pasos.

Primero, se crea una explicación sobre un tema al mostrar secuencias de texto pertinentes en el modelo que se está evaluando, en este caso, GPT-2. En base a esto, la herramienta examina cada vez que se activa una "neurona". Luego, se suministra toda esta información al modelo GPT-4, que con estos datos genera una explicación y una predicción de cómo se comportaría la neurona dependiendo del caso. Finalmente, la herramienta de OpenAI utiliza nuevamente GPT-4 para comparar el comportamiento de la neurona simulada en el segundo paso y la neurona real de GPT-2.

Además, esta herramienta califica la comparación de ambas neuronas, detallando cómo la explicación de GPT-4 coincide con el comportamiento real del LLM evaluado, como explica la empresa.

Uno de los responsables del proyecto, William Sanders, mencionó en declaraciones a TechCrunch que con esta herramienta la compañía espera desarrollar "maneras de prever cuáles serán los problemas con un sistema de IA" (Sanders, 2023), lo que implica un mejor entendimiento de su funcionamiento interno para poder determinar la confiabilidad de los resultados del modelo y las respuestas que genera (OpenAi).

Aunque esta herramienta se encuentra en una fase temprana, los investigadores han logrado generar explicaciones para las 307.200 neuronas en GPT-2, y han recopilado estas en un conjunto de datos que se ha publicado junto con el código de la herramienta en GitHub.

Entre sus funciones, incluye mejorar el rendimiento de los modelos de lenguaje e identificar situaciones como el uso de sesgos en su comportamiento. Sin embargo, la herramienta aún está en fase de prueba, por lo que se esperan más funcionalidades en el futuro.

5. La Aplicación de la Dimensión Ética a la Inteligencia Artificial Generativa

"We must therefore be vigilant and work to ensure that the discriminatory use of these instruments does not take root at the expense of the most fragile and excluded." - Pope Francis after Roma Call for AI Ethics.

Teniendo ya una base fortalecida sobre los modelos y conceptos de tanto la Inteligencia Artificial como la Inteligencia Artificial Generativa, es momento de investigar el porqué del desarrollo e implementación de estas tecnologías. Surge la imperiosa necesidad de analizar e incorporar la dimensión ética. Este capítulo tiene el objetivo de explorar esta necesidad y analizar cómo los principios éticos pueden informar y mejorar no solo la práctica de la inteligencia artificial generativa, sino su desarrollo también. Al abordar los principios como, la transparencia, la igualdad, la seguridad, y la privacidad; buscamos un equilibrio entre la explotación de la tremenda capacidad de estos modelos y la salvaguarda de los valores fundamentales humanos en un mundo cada vez más digital. Investigaremos estas cuestiones partiendo del Humanismo Digital; que pone al ser humano en el centro de todo desarrollo tecnológico.

5.1. El Humanismo Digital: Definir e introducir concepto

El Humanismo Digital se refiere a la corriente filosófica que implica la tecnología digital y la conectividad como los vínculos para mejorar la vida humana y promover los valores y principios humanísticos. En esta filosofía, se mantiene al ser humano como el centro de todo desarrollo tecnológico, de esta manera asegurando que estas expandan las habilidades humanas, y no destrozarlas.

5.1.1. El Humanismo

Antes de embarcarnos al Humanismo Digital, es importante definir correctamente el término Humanismo. En su libro; *Digital Humanism: A Philosophy for 21st Century Digital Society*,

Christian Fuchs, sociólogo austriaco en la Universidad de Westminster, define el Humanismo como un enfoque filosófico que hace hincapié en las capacidades activas y transformadoras del ser humano en el mundo social. Añade que el Humanismo se presenta como un sistema de pensamiento que concentra las grandes preguntas del mundo, basándose en la ciencia y la razón como herramientas invaluables que los humanos podemos y debemos aplicar a todas las áreas de la vida (Ley, 2011). En este sentido, promueve la importancia de la ciencia y la razón en la comprensión del mundo, algo muy importante en la era actual, donde la información y el conocimiento se han convertido en los principales motores del progreso y desarrollo humano. En un mundo cada vez más interconectado y digitalizado, el Humanismo nos recuerda la importancia de la perspectiva humana, la autonomía individual y la cooperación colectiva para enfrentar los retos y desafíos que se nos presentan.

El Humanismo también se describe como una perspectiva del mundo que enfatiza que los valores que valoramos en la vida humana no son una ilusión. Como seres humanos, podemos encontrar desde nuestros propios recursos los valores morales compartidos que necesitamos para vivir juntos y los medios para crear vidas significativas y satisfactorias para nosotros mismos. El rechazo a la creencia religiosa no debe ser una causa de desesperación (Norman, 2004).

Por otro lado, el Humanismo es una filosofía o conjunto de creencias que sostiene que los seres humanos logran un sistema de moralidad a través de su propio razonamiento más que a través de una creencia en cualquier ser divino (Andrews, 2010).

Desde el siglo XIV hasta finales del siglo XIX, el Humanismo mínimamente significaba: un programa educativo basado en los autores clásicos y centrado en el estudio de la gramática, la retórica, la historia, la poesía y la filosofía moral; un compromiso con la perspectiva, los intereses y la centralidad de las personas humanas; una creencia en la razón y la autonomía como aspectos fundamentales de la existencia humana; una creencia de que la razón, el escepticismo y el método científico son los únicos instrumentos apropiados para descubrir la verdad y estructurar la comunidad humana; una creencia de que los fundamentos para la ética y la sociedad se encuentran en la autonomía y la igualdad moral (Luik, 1998).

5.1.2. El Humanismo Digital

El Humanismo Digital defiende las capacidades activas y transformadoras de los seres humanos en la era digital. A medida que la tecnología digital permea cada vez más aspectos de la vida cotidiana, es imperativo que reflexionemos sobre el papel que juegan los humanos y la tecnología en la sociedad.

El Humanismo Digital se basa en el principio ético de que no debemos transformar a los humanos en máquinas ni identificar a las máquinas como humanos. Al revés, debemos mantener la especificidad del ser humano y utilizar las tecnologías digitales para expandir nuestras habilidades, no para derrotarlas. El humanismo digital se desglosa en tres dimensiones: la epistemología, la ontología y la axiología.

La epistemología del Humanismo Digital sostiene que las tecnologías de la computación y las máquinas en general son diferentes de los humanos, ya que carecen de razón, conciencia, moralidad y pensamiento crítico. Aunque los robots, los big data, los métodos computacionales y digitales y la inteligencia artificial son herramientas útiles, no pueden ni deben reemplazar el papel del ser humano en la sociedad.

La ontología del Humanismo Digital afirma que las tecnologías y los ordenadores no son entidades humanas ni societales. Los seres humanos y sus actividades, relaciones y conexiones sociales constituyen la sociedad. En las sociedades contemporáneas, las tecnologías digitales dan forma y son formadas por los humanos y sus relaciones sociales, pero estas tecnologías no son actores autónomos. Esto les hace ser diferentes de los humanos, no pueden actuar de forma autónoma.

La axiología del Humanismo Digital sostiene que, dado que las máquinas digitales no son humanas y los humanos no son máquinas, es un imperativo moral no tratar a las máquinas como humanos y viceversa. Las tecnologías digitales deben ser moldeadas y utilizadas de manera que no perjudiquen a la sociedad y a los humanos, sino que apoyen el establecimiento de una sociedad buena y sobre todo humana.

Finalmente, en su libro, Fuchs introduce el concepto del Humanismo Digital Radical. Este término se refiere a una filosofía materialista que subraya las capacidades productivas, sociales, y transformativas de los seres humanos que les permite liberarse de la sociedad de clases digitales, el capitalismo digital, la explotación digital, la dominación digital y la ideología digital (Fuchs,

2021). En resumen, el Humanismo Digital Radical busca crear conocimientos sobre el mundo digital y las tecnologías digitales que apoyen el avance del socialismo democrático y digital.

Dicho esto, ahora exploraremos varios principios éticos y sacaremos conclusiones basadas en la entrevista con Pilar López.

5.2. Problemas éticos con la Inteligencia Artificial Generativa

En la era digital en la que vivimos, la Inteligencia Artificial Generativa (IAG) está tomando un papel cada vez más relevante, transformando varios sectores y cambiando las formas en que interactuamos con la tecnología. Sin embargo, a medida que la IAG avanza y su integración en nuestra sociedad se convierte más presente, también emergen desafíos éticos significativos que requieren una consideración cuidadosa. Este capítulo explorará los principales problemas éticos asociados con la IAG, tomando como referencia cuatro principios éticos fundamentales: la transparencia, la igualdad, la seguridad, y la privacidad. Estos principios nos permitirán analizar en profundidad la responsabilidad de los desarrolladores de IAG, la protección de los usuarios, la justicia en su implementación y la salvaguarda de la información personal y confidencial. Al abordar estos temas, nuestro objetivo es promover un entendimiento crítico y proponer posibles soluciones a estos dilemas éticos, con el fin de guiar el desarrollo de la IAG hacia un camino que respete los derechos y las libertades fundamentales de las personas, y que al mismo tiempo permita aprovechar los beneficios que esta tecnología puede aportar a nuestra sociedad.

A partir de este punto, muchas de las opiniones y datos se habrán sacado de la entrevista con Pilar López.

5.2.1. La Transparencia

En el campo de la IAG, se está incorporando la transparencia como principio fundamental en las políticas y estrategias empresariales. De hecho, en Europa, la nueva ley europea La Ley europea sobre Inteligencia Artificial (IA) aborda importantes cuestiones éticas y de derechos de autor en el uso de los datos por parte de las empresas de IA. Las empresas, como OpenAI, que desarrollan IA

generativa, deben ser transparentes acerca del uso de materiales protegidos por derechos de autor para el entrenamiento de sus algoritmos. Estas empresas también deben divulgar cualquier material protegido por derechos de autor utilizado para desarrollar sus sistemas. Este problema tiene una relevancia particular dado que ha habido conflictos entre autores y desarrolladores en relación con el uso de material protegido por copyright para el entrenamiento de IA (ObservatorioBlockchain).

Iniciado por la Comisión Europea hace casi dos años, el desarrollo de la Ley de Inteligencia Artificial (IA) surgió en respuesta al incremento notorio de la inversión y la popularidad en el campo de la IAG, en particular después del lanzamiento del asistente virtual ChatGPT por OpenAI. El propósito central de la propuesta legislativa es la salvaguarda de los derechos esenciales, la seguridad de los usuarios y el fomento de la confianza en el desarrollo y adopción de la IA. Esto es débido al rápido crecimiento, a la vez imprevisto, de la IAG que trae consigo numerosas cuestiones sobre su implementación y uso.

Los legisladores europeos han confirmado propuestas para imponer obligaciones más estrictas a los modelos básicos de IAG, incluyendo que los modelos generativos de IA se diseñen y desarrollen según la legislación de la UE y los derechos fundamentales. La ley en desarrollo también clasifica los sistemas de IA según cuatro niveles de riesgo, imponiendo obligaciones correspondientes a cada nivel.

Por último, la ley también pide medidas de rendición de cuentas y transparencia cuando las entidades públicas y privadas utilizan sistemas de IA en la UE. Deben revelar su uso de sistemas de IA de alto riesgo y publicar evaluaciones exhaustivas de su impacto en los derechos humanos. Este requisito es crucial para que las personas perjudicadas por los sistemas de IA puedan solicitar compensaciones.

La transición de Microsoft hacia una estrategia más centrada en la Inteligencia Artificial Generativa, y específicamente su inversión en OpenAI, está reformando de manera significativa el panorama de las búsquedas en línea, un espacio dominado hasta ahora por Google. Este cambio es revolucionario y tiene profundas implicaciones por varias razones.

En primer lugar, las búsquedas en línea realizadas a través de Google disminuyeron por la irrupción de modelos de IA como GPT, demostrando una brecha en el mercado que Microsoft desea llenar. Este cambio de comportamiento del usuario representa una oportunidad dorada para que Microsoft

desafíe la posición dominante de Google y se haga con una porción significativa de la cuota de mercado.

El enfoque de Microsoft es considerablemente diferente del que Google ha empleado históricamente. En lugar de basar los resultados de búsqueda en anuncios publicitarios, como hace Google, Microsoft se propone ofrecer una experiencia de búsqueda mucho más personalizada y centrada en el usuario. Por ejemplo, al buscar un restaurante, Microsoft tiene como objetivo brindar recomendaciones basadas en tus búsquedas previas y tus gustos personales, lo que resulta en una experiencia de usuario más intuitiva y relevante.

Este enfoque subraya la transparencia como un componente clave del modelo de negocio en el que invierte Microsoft. Al proporcionar un razonamiento para sus recomendaciones, Microsoft permite a los usuarios entender por qué se les presentan ciertos resultados, lo que aumenta la confianza en el sistema. Este nivel de transparencia y personalización contrasta con la opacidad del sistema de algoritmos de Google, actualizado para aumentar los ingresos de Google, no para mejorar la experiencia del usuario.

Aunque esta es solo la etapa inicial, el modelo que Microsoft propone tiene un potencial significativo para remodelar la categoría de software más grande del mundo: la de búsqueda. Al poner el enfoque en el usuario y en la transparencia, Microsoft está desafiando la manera en que se han realizado las búsquedas en línea hasta ahora.

Este cambio también subraya una tendencia más amplia hacia la personalización y la transparencia en la tecnología. A medida que los usuarios se vuelven más conscientes de la importancia de la privacidad de los datos y buscan un mayor control sobre su experiencia en línea, las empresas que pueden ofrecer estos elementos se situarán en una posición ventajosa. La decisión de Microsoft de invertir en OpenAI y de centrarse en mejorar la experiencia de búsqueda del usuario a través de la IA ilustra esta tendencia y sugiere un futuro prometedor para la búsqueda personalizada en línea.

5.2.2. La Igualdad

La IAG se está convirtiendo en una tecnología omnipresente en la sociedad y en el ámbito empresarial. Aunque puede beneficiar la vida diaria de las personas y la economía global, es crucial tener en cuenta los posibles prejuicios involuntarios que pueden derivarse de su aplicación.

La IAG, sin duda, está transformando el mundo a un ritmo acelerado. A la vez, también perpetua la discriminación y la desigualdad. Los modelos de IAG que reproducen y perpetúan prejuicios de género, raza, edad y otros grupos marginados pueden tener serias repercusiones en la sociedad, provocando exclusión, injusticia y falta de oportunidades.

Los prejuicios en la IAG pueden tener diversas raíces, como la falta de diversidad en los equipos de desarrollo de IAG, la escasez de datos equilibrados y representativos, y la ausencia de políticas y prácticas responsables en el desarrollo y uso de la IAG, o simplemente porque lo consideramos habitual ya que proviene de un grupo de personas reducido y homogéneo (Páez Pino, 2023).

Estos prejuicios podríamos catalogarlos como inconscientes, dado que las personas incorporan sus propios prejuicios en la tecnología. Esto puede resultar en modelos de IAG que perpetúan la discriminación y la desigualdad, y que impactan negativamente a los grupos marginados.

Por lo tanto, es vital incorporar una amplia gama de actores y perspectivas en el desarrollo y uso de la IAG, y adoptar políticas y prácticas responsables y transparentes para minimizar los prejuicios e impactos negativos en estos modelos.

Estos algoritmos tienen errores que no comprendemos, ya que no hay simetría en su uso. Estos errores tienen un pasado que puede ser arrastrado desde las injusticias de nuestra historia, que se ven reflejados en la IAG en términos de raza, género, y otros aspectos.

Es imprescindible tomar medidas proactivas para evitar que los modelos de IAG perpetúen y reflejen la discriminación y la desigualdad existentes en la sociedad. Es el momento de supervisar cada proceso en busca de prejuicios, ya que puede llevarnos a dos realidades: una en la que se mantiene el orden social, y otra que se usa para fines comerciales y generación de ingresos, alejándose del beneficio puro del interés público.

Además, es esencial monitorear y evaluar constantemente los modelos de IAG para corregir los prejuicios. La responsabilidad no recae en un solo grupo o individuos, sino en todos los actores

involucrados en el desarrollo y la implementación de la IA. Debemos trabajar juntos para garantizar la equidad y la justicia en la tecnología.

Existen varios tipos de prejuicios que pueden surgir en los modelos, incluyendo:

1. **Prejuicio en los datos**: ocurre cuando los datos utilizados para entrenar un modelo de IAG están sesgados. Por ejemplo, si los datos se recopilan únicamente de un grupo demográfico específico, el modelo puede tener un prejuicio hacia ese grupo.
2. **Prejuicio de confirmación**: nace cuando un modelo de IAG tiende a repetir o reforzar sus suposiciones o creencias previas sobre un grupo específico.
3. **Prejuicio de codificación**: se da cuando los desarrolladores de IAG no tienen en cuenta las desigualdades de género, raza, edad, orientación sexual, discapacidad, etc. en el diseño de un modelo.
4. **Prejuicio de interpretación**: sucede cuando los resultados generados por un modelo de IAG son interpretados de manera sesgada por humanos.
5. **Prejuicio de métricas**: acontece cuando las métricas utilizadas para evaluar el rendimiento de un modelo de IAG están sesgadas o no reflejan adecuadamente la diversidad de la población afectada por el modelo.

Es vital ser consciente de estos prejuicios y trabajar activamente para prevenirlos y minimizarlos en el desarrollo de sistemas de IAG. Esto incluye la revisión crítica de los algoritmos, datos y modelos de IAG, la implementación de pruebas rigurosas, y la consideración ética y diversa en el diseño, desarrollo, e implementación de sistemas de IAG.

Los avances de estas tecnologías aumentan rápidamente y son herramientas poderosas que podrían ser utilizadas de forma abusiva y generar consecuencias negativas para la sociedad. Por lo tanto, debemos implementar políticas y leyes y diversificar los datos de estos modelos. En la entrevista con Pilar López estresaba la importancia de esto, es necesario entrenar los modelos desde un criterio ético y con datos perfectos, ahora aún más con los modelos IAG, que funcionan por la probabilidad de que una palabra siga a otra. Nadie sabe lo que el futuro deparará, pero es una cuestión muy pendiente ya que puede tener unas repercusiones muy negativas a la sociedad.

Para finalizar, los prejuicios en la IAG son un problema significativo que requiere un esfuerzo conjunto de todas las partes involucradas para abordarlo y garantizar que la IAG se utilice de manera ética y responsable para el bienestar de todas las personas.

5.2.3. La Privacidad

El consentimiento al uso de datos personales es un tema de alta relevancia que involucra todo tipo de sistemas de IA. En este apartado nos enfocaremos en el caso de ChatGPT, el LLM (Large Language Model) más relevante a dia de hoy. Hace poco, una falla técnica en ChatGPT dejó al descubierto los títulos de las sesiones de chat de varios usuarios, en un incidente que podría considerarse una violación de datos. El CEO de OpenAI, Sam Altman, confirmó el problema y señaló que la solución implicaría que los usuarios no podrían acceder a su historial de chat durante un tiempo.

ChatGPT, más que una herramienta de búsqueda es una plataforma de interacción en la que los usuarios comparten ideas, consejos, emociones y más. En consecuencia, tiene una perspectiva única sobre nuestras preocupaciones, emociones, y cosas que nos gustan. Es un generador de lenguaje refinado para imitar nuestras formas de comunicación, lo que hace que sea más probable que compartamos información con él de manera similar a como lo haríamos con una persona conocida.

ChatGPT se basa en un volumen masivo de textos de Internet, incluyendo blogs, artículos, Wikipedia y foros públicos. La última versión, ChatGPT 4, incorpora 300 mil millones de palabras (OpenAi). Esto ha suscitado preocupaciones sobre la privacidad, ya que la información utilizada para afinar a ChatGPT puede haber sido generada por personas que no dieron su consentimiento para dicho uso.

OpenAI, la empresa matriz de ChatGPT, ha sido criticada por su falta de transparencia en la gestión de los datos personales. Aunque su política de privacidad detalla cómo se manejan los datos recopilados de los usuarios durante el registro y el uso del servicio, no proporciona información sobre cómo se manejan los datos personales de otras fuentes, como los que se utilizan para generar las respuestas de ChatGPT.

Paula Ortiz, una experta en derecho digital, ha señalado que la política de privacidad de OpenAI no se ajusta a las normas del Reglamento General de Protección de Datos (RGPD) de la UE. También se ha planteado el problema de que la privacidad de los datos es diferente para los usuarios de pago y los gratuitos, con una falta de claridad sobre cómo se manejan los datos de los usuarios de pago.

En resumen, aunque ChatGPT es una herramienta poderosa y cada vez más popular para la generación de texto, existen serias preocupaciones sobre su gestión de la privacidad de los datos. La falta de transparencia de OpenAI y la insuficiencia de las leyes existentes para regular las nuevas formas de IA son problemas que necesitan ser abordados. El desafío será asegurar que la protección de la privacidad mantenga el ritmo con la evolución de la tecnología.

Este problema de la privacidad de los datos se relaciona con el "capitalismo de vigilancia", un término que acuñó la socióloga Shosana Zuboff en su libro La Era del Capitalismo de la Vigilancia. El "capitalismo de la vigilancia" describe la recolección y utilización de los datos personales por parte de las grandes compañías tecnológicas con el ánimo de predecir el comportamiento de la sociedad (Zuboff, 2018). Este concepto plantea un desafío para la privacidad individual, ya que los usuarios no siempre son conscientes o aprueban el intercambio y venta de sus datos personales llevado a cabo por gigantes de la tecnología como Google, Amazon, Facebook, Apple y Microsoft, también conocidos como GAFAM.

Estas empresas no solo usan los datos para mejorar sus productos y servicios, sino que también los venden a empresas publicitarias en lo que Zuboff llama el "mercado de comportamientos futuros". Esta información, procesada mediante inteligencia artificial, permite a las empresas predecir comportamientos sociales y así afinar su publicidad. Las aplicaciones ofrecidas gratuitamente por estas empresas se actualizan constantemente para mantenerse atractivas y atraer a más usuarios, un ejemplo claro son las actualizaciones frecuentes de Instagram.

Sin embargo, el capitalismo de la vigilancia difiere de los sistemas capitalistas industrial, comercial o financiero. Es un sistema exclusivo de internet en el que los usuarios son la fuerza laboral, entregando sus vidas privadas a las empresas tecnológicas de forma gratuita. Sus datos personales se convierten en mercancía, que se negocia de manera oculta sin el consentimiento del individuo.

Zuboff sostiene que este sistema amenaza la democracia, ya que las empresas tecnológicas buscan maximizar sus beneficios a expensas de los derechos de los ciudadanos y resisten las leyes que limitan sus acciones. La Unión Europea y el estado de California en Estados Unidos han adoptado leyes para proteger los datos personales de los ciudadanos, pero el comercio de datos sigue prosperando. En la era digital, los perfiles de navegación y redes sociales se han convertido en el "petróleo del siglo XXI".

5.2.4. La Seguridad

El tema de la seguridad en la Inteligencia Artificial Generativa (IAG) está estrechamente vinculado con dos problemas sociales emergentes: la desinformación y el aumento exponencial de los "deepfakes". Este aspecto de seguridad tiene muchas ramificaciones que merecen una atención profunda.

En primer lugar, la desinformación es un problema cada vez más relevante en nuestra sociedad hiperconectada. La IAG, a través de sus avanzados sistemas de generación de contenido, puede contribuir a la propagación de las "fake news" ya sea intencionalmente por parte de actores malintencionados o por simple malentendido de los algoritmos. Esto se debe a que los sistemas de IAG pueden generar información rápida y masivamente, sin verificar necesariamente su veracidad. Esta capacidad puede ser explotada por aquellos que buscan difundir falsedades, desacreditar a individuos o entidades, o incluso influir en la opinión pública y las decisiones políticas. Pilar López añadió que este podía ser el mayor riesgo de estas tecnologías, dado que la manipulación de la información puede tener consecuencias de gran alcance, especialmente en una era donde la información se considera uno de los recursos más valiosos. Según Pilar, la era digital ha proporcionado un amplio acceso a la información, pero también ha abierto las puertas a la manipulación y tergiversación de esa información. Añade también que el desarrollo de la IAG está armando a los defensores de estas injusticias, por lo cual no va a ser una pelea fácil para los que buscan usar la IAG de manera maliciosa. Por último, Pilar afirma que es vital para nuestra sociedad desarrollar métodos más eficaces para detectar y combatir la desinformación generada por la IAG. Estos pueden incluir la promoción de la transparencia y la responsabilidad en el diseño y la implementación de la IAG, así como la educación de los usuarios sobre los riesgos de la desinformación y cómo identificarla. Solo a través de un enfoque multifacético y colaborativo

podemos esperar mitigar estos riesgos y aprovechar al máximo los beneficios que la IAG puede ofrecer.

En segundo lugar, los “deepfakes” son ejemplos de cómo la IAG puede ser utilizada para fines maliciosos. Los deepfakes son videos e imágenes falsos generados por IA, que son tan realistas que pueden engañar a las personas haciéndoles creer que son auténticos. Los individuos en estos deepfakes, a menudo figuras públicas, pueden manipularse para decir o hacer cosas que nunca ocurrieron. Con la creciente sofisticación de la tecnología, estos deepfakes están alcanzando niveles de realismo que dificultan su detección, lo que plantea serias preocupaciones en términos de seguridad y privacidad.

La proliferación de deepfakes puede tener consecuencias graves, desde la difamación y el engaño personal hasta la desestabilización política y la manipulación de elecciones. Por ejemplo, un deepfake convincente de un líder político haciendo comentarios incendiarios podría desatar una crisis política o incluso un conflicto militar.

Por lo tanto, es crucial que se implementen medidas de seguridad robustas en la IAG para mitigar estos riesgos. Esto podría incluir la construcción de algoritmos de IA que sean más transparentes y explicables, para que sea más difícil el uso indebido de la tecnología. Además, la detección de deepfakes y la verificación de noticias también son áreas donde la propia IA puede desempeñar un papel fundamental en la lucha contra estos problemas.

En resumen, la seguridad en la IAG es un tema crucial que necesita abordarse para prevenir la propagación de desinformación y la creación de deepfakes. Esta es una tarea que requerirá la colaboración de expertos en IA, legisladores, y la sociedad en general para garantizar un futuro digital seguro y confiable.

5.3. Como se están afrontando estos problemas

En la era de la digitalización y la inteligencia artificial, enfrentar los problemas éticos en el diseño e implementación de la IA generativa se convierte en una tarea de vital importancia. El creciente desarrollo y uso de estas tecnologías exponen una serie de dilemas éticos, desde la protección de la privacidad y la seguridad de los datos hasta la proliferación de desinformación y deepfakes.

Asimismo, los sesgos algorítmicos y la falta de transparencia representan desafíos adicionales. Por tanto, es esencial establecer estrategias y marcos éticos adecuados para garantizar que estas tecnologías se diseñen e implementen de manera responsable y equitativa, priorizando siempre el respeto por los derechos humanos y la dignidad de las personas. A lo largo de este apartado, exploraremos las múltiples maneras de abordar estos desafíos y garantizar la integridad ética de la IA generativa.

5.3.1. Aether Committee

El "Aether Committee" o comité Aether, fue fundado por Microsoft en 2017 para decidir sobre las aplicaciones de la IA. Aether, siglas de AI, Ethics and Effects in Engineering and Research (IA, ética y efectos en ingeniería e investigación), es el consejo asesor de Microsoft que delibera sobre cuestiones, problemas y retos que surgen al desarrollar y poner en marcha aplicaciones de IA (Microsoft, 2018). Toma decisiones sobre el desarrollo y la implementación de la IA dentro de Microsoft y también para las empresas clientes de esta. La meta es garantizar que las tecnologías de Microsoft sean desarrolladas e implementadas de manera responsable y que se alineen con los principios de la empresa en cuanto a la confiabilidad, seguridad, privacidad y transparencia.

El comité Aether está formado por un grupo interdisciplinario de miembros de Microsoft, incluyendo científicos de la computación, ingenieros, expertos en políticas, abogados y éticos. Este grupo diverso se reúne regularmente para discutir y debatir cuestiones emergentes en el campo de la IA y otras tecnologías avanzadas. Su trabajo ayuda a formar las políticas y prácticas de Microsoft en relación con estas tecnologías. Ahora más que nunca, su propósito ha aumentado en relevancia con el crecimiento exponencial de las aplicaciones de IA generativa.

Algunos de los temas que el comité Aether ha abordado incluyen la equidad y el sesgo en los sistemas de la IA, la seguridad y robustez de estos sistemas, la capacidad de explicar las decisiones de la IA, la colaboración humano-IA y los usos sensibles y potencialmente controvertidos de la IA, como el reconocimiento facial.

El comité Aether informa directamente al equipo de liderazgo senior de Microsoft, lo que subraya la importancia que la empresa otorga a las cuestiones éticas y sociales en el desarrollo y la implementación de tecnologías de IA, particularmente generativa en el día de hoy.

Es fundamental que comités como el de Aether existan en nuestra sociedad por varias razones. En primer lugar, estos comités proporcionan una plataforma crucial para el análisis profundo y el debate sobre los desafíos éticos, legales y sociales que emergen de la aplicación de tecnologías avanzadas como la IA. En segundo lugar, ayudan a fomentar la responsabilidad y la transparencia, asegurando que las decisiones y acciones tomadas en el desarrollo y uso de estas tecnologías estén en consonancia con los principios éticos y los derechos humanos. Además, su existencia puede fomentar una mayor confianza en la tecnología por parte del público, al demostrar un compromiso serio con la ética y la responsabilidad. Finalmente, estos comités juegan un papel crucial en la identificación y mitigación de posibles sesgos, desigualdades y otros efectos negativos que pueden surgir de la implementación de la IA y otras tecnologías similares.

5.3.2. Roma Call

El "El Llamado a la Ética de la IA" o "Roma Call" es un documento firmado por la Academia Pontificia para la Vida, Microsoft, IBM, la FAO y el Ministerio de Innovación, una sección del gobierno italiano en Roma el 28 de febrero de 2020 para impulsar un enfoque ético hacia la inteligencia artificial. La filosofía subyacente es fomentar un sentido de responsabilidad compartida entre organizaciones internacionales, gobiernos, instituciones y el sector privado con el fin de diseñar un futuro en el que la innovación digital y el progreso tecnológico otorguen a la humanidad su centralidad. Sugiriendo una nueva "algorethics" o algor-ética, entre la IA y la reflexión ética, los firmantes se comprometieron a promover el desarrollo de una inteligencia artificial que beneficie a cada individuo y a la humanidad en general; que respete la dignidad humana, permitiendo que todos puedan beneficiarse de los avances tecnológicos; y que no persiga únicamente un mayor beneficio económico o el reemplazo gradual de las personas en el entorno laboral.

El Llamado a la Ética de la IA de Roma se compone de 3 áreas de impacto y 6 principios.

Las tres áreas de impacto son ética; Todos los seres humanos nacen libres e iguales en dignidad y derechos, educación; Transformar el mundo a través de la innovación de la IA implica comprometerse a construir un futuro para y con las generaciones más jóvenes, y el derecho; El desarrollo de la IA al servicio de la humanidad y el planeta debe reflejarse en regulaciones y

principios que protejan a las personas, especialmente a las más vulnerables, y a los entornos naturales.

Los seis principios son, la transparencia; los sistemas de IA deben ser comprensibles para todos. Inclusión; estos sistemas no deben discriminar a nadie, ya que todos los seres humanos tienen igual dignidad. Responsabilidad; siempre debe haber alguien que se responsabilice de lo que hace una máquina. Imparcialidad; los sistemas de IA no deben seguir ni crear sesgos. Fiabilidad; la IA debe ser confiable. Seguridad y privacidad; estos sistemas deben ser seguros y respetar la privacidad de los usuarios.

El trabajo realizado por este grupo es reconocido por el Papa Francisco, quien elogia sus esfuerzos para proteger el bienestar de la familia humana, fomentar una ética compartida y la fraternidad entre todos, y estar alerta contra los usos maliciosos de la tecnología y la inteligencia artificial. El Papa reconoció que la inteligencia artificial es cada vez más prevalente en los aspectos de la vida cotidiana, alterando cómo comprendemos el mundo y a nosotros mismos, e instó a perseverar en este esfuerzo. Promovió la algor-ética, es decir, la reflexión ética sobre el uso de algoritmos, para que esté cada vez más presente no solo en el debate público, sino también en el desarrollo de soluciones técnicas. “De hecho, cada individuo debe poder disfrutar de un desarrollo humano y solidario, sin que nadie quede excluido.” (Pope Francis, 2023)

En conclusión, es positivo para la sociedad ver como este tipo de esfuerzos, como la Roma Call, se están llevando a cabo, ya que representan un paso significativo hacia la incorporación de principios éticos en el diseño y uso de la inteligencia artificial. El camino hacia la IA ética va a ser complejo y desafiante, pero iniciativas como la Roma Call demuestran que es posible, y es un camino que estamos dispuestos y decididos a recorrer.

6. Reflexiones y sugerencias

Después de un recorrido extenso y detallado por el panorama actual de la Inteligencia Artificial Generativa (IAG) y su enfoque ético, nos encontramos en una encrucijada histórica. Este análisis nos ha llevado a través de una variedad de temas, desde la comprensión fundamental de qué es la IA, su historia y evolución, pasando por el funcionamiento y entrenamiento de los sistemas de IAG, hasta el enfoque ético contemporáneo para la implementación de IAG.

La Inteligencia Artificial, con su increíble potencial y posibilidades, ha logrado situarse en el centro de nuestra sociedad. Sin embargo, es su vertiente generativa la que ha mostrado un potencial excepcional en diversas áreas e industrias, desde la creación artística hasta la generación de contenido escrito. Sin embargo, la implementación y el avance de esta tecnología no están exentos de dilemas éticos y morales.

El Humanismo Digital, que combina los principios fundamentales del humanismo con los desafíos y oportunidades que presenta la era digital, tiene una relación intrínseca con la ética en IA. Se enfoca en temas críticos como la transparencia, la igualdad, la privacidad y la seguridad en la implementación de la IA. Este enfoque humanista nos recuerda que cualquier avance tecnológico, por más innovador que sea, debe mantener al ser humano en su centro y respetar los derechos y la dignidad humanos. Para mí, este enfoque es lo más importante a día de hoy, donde hay avances en estas tecnologías diariamente, y son avances extremadamente disruptivos. Es fundamental que los gobiernos de cada país, junto a las grandes empresas, creen modelos para supervisar y acompañar el crecimiento de estas tecnologías para que vayan de la mano con la ética.

El problema tiene más complejidad. Primero, es casi imposible crear una serie de principios que tengan que seguir todos los países del mundo. Habrá países a los que le interesa más frenar el desarrollo de la IAG que a otros. Por ejemplo, aunque la Unión Europea este abierta a hacer esto, las grandes empresas junto al gobierno americano nunca cederán a esto, debido a que actualmente las empresas más innovadoras en esta industria son mismamente americanas. A parte, a los Estados Unidos les interesa este crecimiento para adelantarse a China, quien hasta el boom de OpenAi era el líder mundial en IAG. Dicho esto, frenar no es una opción. La innovación no se puede parar. Hay que buscar que los efectos negativos sean regulados, haya que gestionarlos, anticiparlos. En

vez de frenarla, debemos regular y gestionar sus efectos adversos, de anticiparnos a ellos. Es fundamental que dediquemos tiempo y esfuerzo a investigar y entender las potenciales consecuencias negativas de la IAG para poder mitigarlas y controlarlas, garantizando siempre que los beneficios superen a los posibles perjuicios.

No solo hay que evaluar los impactos negativos, sino que también los desarrolladores de la IAG deben tener sus propios principios, evaluar qué cosas quieren hacer y que cosas no quieren hacer, todo desde una perspectiva ética. Es vital saber que cosas están bien y que otras están mal, y de este modo poder siempre en control del desarrollo e implementación de la IAG.

Uno de los mayores miedos que rodean la implementación de la IAG en el ámbito laboral es la reducción de trabajos que la acompañará. Hablando con Pilar López, ella comparó esta situación actual con el dilema que hubo en Nueva York a finales del siglo XIX con la aparición del automóvil. Aquellos que trabajaban cuidando y criando caballos, que eran miles de personas, perdieron su trabajo ya que el uso de los caballos como medio de transporte disminuyó notablemente en la ciudad. Aún asi, estos trabajadores empezaron a dedicarse a otras cosas, muchos involucrados en la industria automovilística. Al final, el problema de la perdida de miles de trabajos no es el principal riesgo, porque con estas implementaciones tecnológicas, habra miles de trabajos nuevos también. La economía se adaptó en su momento y lo volverá a hacer en este caso, y los trabajadores se adaptarán con ella. De todas formas, para que la IAG provoque un cambio de trabajo de tal magnitud, necesitaríamos ver un aumento dramático en la productividad que actualmente no se espera, incluso con un incremento del PIB del 3-4%. Esto no se contempla actualmente por ningún experto.

Existe el riesgo de que haya desequilibrios causados por la IAG, pero no se anticipan cambios de ese calibre en el futuro próximo. Un cambio tan radical requeriría una transformación sistémica de la industria, algo que no se ve posible en un futuro cercano. Pilar proporcionó un ejemplo relevante de la industria sanitaria, donde la IAG se está implementando sistemáticamente para mejorar la sostenibilidad del sistema. En países como los Estados Unidos, donde el costo del cuidado de la salud es muy alto, o en Europa, donde los costos están cubiertos por el sistema de seguridad social, la IA tiene el potencial de alargar la vida humana, reducir los costos de los recursos médicos y mejorar la calidad de vida (López, 2023). Pero los efectos negativos que se verían si la IA se aplicara de golpe en todas las industrias serían más bien anecdóticos y no realistas, y no hay

consenso generalizado sobre que la IAG pueda aumentar el PIB en varios puntos porcentuales de manera inmediata.

“Nunca regular la tecnología en sí, es decir, no prohibir usar determinadas o desarrollar determinadas tecnologías, si regular los casos de uso. Por ejemplo, aunque los Modelos de Lenguaje Multi-Modal Generativos a Gran Escala (GLLMM) están permitidos y se fomenta su desarrollo, no se debe permitir su uso para registrar a las personas sin su consentimiento explícito. Esto protege tanto los derechos de privacidad como el desarrollo y la innovación en el ámbito de la IAG.

Prohibir tecnologías enteras podría tener un impacto negativo en la innovación y detener el progreso en campos que tienen un gran potencial para el bienestar humano y el desarrollo de la sociedad. Por tanto, el desafío radica en identificar y regular los casos de uso específicos que podrían amenazar la dignidad y los derechos humanos, manteniendo a la vez un entorno favorable para la innovación tecnológica.

Establecer consenso sobre estos casos de uso perjudicial y cómo deben gestionarse requiere la colaboración de las partes interesadas: desde desarrolladores de tecnología y legisladores hasta consumidores y sociedad. Este enfoque centrado en el caso de uso permite un equilibrio entre proteger los derechos de las personas y fomentar la innovación tecnológica. Es un camino que requiere una deliberación cuidadosa, transparencia y participación abierta.

El principal riesgo de la IAG es el uso malicioso de esta tecnología. Estas regulaciones y legislaciones, cuanto más transnacionales sean mejor. En gobiernos totalitarios por ejemplo es mucho más difícil que rijan estas reglas, por eso estas aplicaciones éticas de las que tanto se ha hablado deben ser consideradas por grandes grupos como el G7, G20, y la OCBE. La guerra de mañana será contra quienes mal usen la IAG. Pero no cunda el pánico, porque a la vez que se arman los malos, se arman también los buenos, y Pilar López me aseguró que están desarrollando estas tecnologías a cuotas impresionantes jamás vistas.

Para concluir, es importante mantener una perspectiva ética en el desarrollo y uso de la Inteligencia Artificial Generativa. Los juicios que se realicen y las políticas que se establezcan por los entes gubernamentales y las grandes corporaciones internacionales tendrán un impacto significativo en la dirección futura de nuestra sociedad.

Estos actores desempeñan un papel crucial en la formulación de normativas y la creación de infraestructuras que pueden moldear el entorno de la IAG. Por tanto, deben esforzarse por comprender las implicaciones éticas y sociales de estas tecnologías, considerando tanto sus posibles beneficios como sus riesgos potenciales.

El desarrollo y la adopción de la IAG deben realizarse de forma responsable, garantizando que se respeten los principios éticos fundamentales. Esto implica una consideración cuidadosa de cuestiones como la transparencia, la privacidad, la seguridad y la igualdad en relación con la IAG.

La necesidad de un enfoque centrado en la ética en la IAG no es solo un asunto de responsabilidad moral, sino también una inversión en la sostenibilidad a largo plazo de nuestras sociedades. Al definir un marco ético para la IAG, estamos sentando las bases para un futuro en el que la tecnología sirva a la humanidad de una manera respetuosa y beneficiosa, en lugar de exacerbar las desigualdades y crear nuevos riesgos.

Por lo tanto, la decisión que tomen los gobiernos y las empresas multinacionales en este sentido será crítica. No solo están definiendo la dirección de la IAG, sino también la forma en que nuestra sociedad interactuará y será moldeada por estas tecnologías en las décadas venideras.

Bibliografía

Libros

Andrews, John. (2010). *The Book of Isms*. London: Profile.

Fuchs, C., (2022). *Digital Humanism: A Philosophy for 21st Century Digital Society*. Emerald Group Publishing.

Kaplan, Andreas. (2022). *Artificial Intelligence, Business and Civilization : Our Fate Made in Machines*, Taylor & Francis Group. ProQuest Ebook Central

Luik, John C. (1998). *Humanism*. In Routledge Encyclopedia of Philosophy, edited by Edward Craig. doi:10. 4324/9780415249126-N025-1.

Law, S. (2011). *Humanism: A Beginner's Guide*. Oneworld Publications.

Mark Skilton, Felix Hovsepian. (2017). *The 4th Industrial Revolution : Responding to the Impact of Artificial Intelligence on Business*. Palgrave Macmillan

Norman, Richard. (2004). On Humanism. London: Routledge

Richardson, Kathleen. (2019) "The Business of Ethics, Robotics, and Artificial Intelligence." Cyborg Futures. Cham: Springer International Publishing. 113–126. Web.

Russell, S., & Norvig, P. (2021). Artificial intelligence: a modern approach. Pearson.

Zuboff, Shoshana. (2019). The Age of Surveillance Capitalism : the Fight for a Human Future at the New Frontier *of Power*. First edition. New York: PublicAffairs. Print.

Artículos

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F. y Rahwan, I. (2018). *The Moral Machine experiment.* Nature, 563, 59-64.

Brownlee, Jason. (2019) *A Gentle Introduction to Generative Adversarial Networks (Gans).* MachineLearningMastery.Com,

Clune, Jeff. (2020). *Ai-Gas: AI-Generating Algorithms, an Alternate Paradigm for Producing General Artificial Intelligence.* ArXiv.Org.

Collosa, Alfredo (2023). *La Ley Europea de Ia y La Transparencia En Los Derechos de Autor.* Noticias Blockchain | Observatorio Blockchain

De Jong, K. A. (2006). *Evolutionary computation: a unified approach*. MIT press.

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative Adversarial Networks, Generating" Art" by Learning About Styles and Deviating from Style Norms. arXiv preprint arXiv:1706.07068.

Gonzalo, Marilín.(2023). *Las Lagunas En Torno a ChatGPT y La Privacidad: Vacío Legal o Pesadilla de Protección de Datos.* Newtral

Goodfellow, Ian J, et al. (2014) *Generative Adversarial Nets - Papers.Nips.Cc.* Generative Adversarial Nets.

Haenlein, Michael, Ming-Hui Huang, and Andreas Kaplan. (2022) *Guest Editorial: Business Ethics in the Era of Artificial Intelligence.* Journal of business ethics 178.4: 867–869. Web.

Hickman, Eleanore, and Martin Petrin. (2021). *Trustworthy AI and Corporate Governance: The EU's Ethics Guidelines for Trustworthy Artificial Intelligence from a Company Law Perspective*. European business organization law review 593–625. Web.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press.

IBM (s.f.) *¿Qué Es La Ia Fuerte? ¿Que Es La IA Fuerte?*, www.ibm.com/es-es/topics/strong-ai.

Kumar Amara, D., Renu Chebrolu, N., Vinayakumar, R. y Soman, K. (2018). A Brief Survey on Autonomous Vehicle Possible Attacks, Exploits and Vulnerabilities.

Kingma, D. P., & Welling, M. (2013). *Auto-Encoding Variational Bayes*

Loureiro, Sandra Maria Correia, João Guerreiro, and Iis Tussyadiah. (2021) *Artificial Intelligence in Business: State of the Art and Future Research Agenda*. Journal of business research 129 : 911–926. Web.

Lubov, Deborah Castellano. (2023). *Pope: AI Ethics must safeguard the good of human family.* Vatican News.

Mantegna, Micaela. (2020). *Creatividad computacional, inteligencia artificial generativa y derechos de autor.* Universidad de Andrés.

Marina, J. A. (2019). *Historia visual de la inteligencia: De los orígenes de la humanidad a la Inteligencia Artificial*. Conecta.

Martindale, C. (1990). *The Clockwork Muse: The Predictability of Artistic Change*. Basic Books.

McCormack, J., Gifford, T., & Hutchings, P. (2019). *Autonomy, Authenticity, Authorship and Intention in Computer Generated Art*. ICCC.

MCCULLOCH, W, and W PITTS. (1990). *A Logical Calculus of the Ideas Immanent in Nervous Activity.*Bulletin of Mathematical Biology, vol. 52, no. 1–2, pp. 99–115,

Minsky, M., & Papert, S. (1988). *Perceptrons: An Introduction to Computational Geometry.* (Expanded Edition) MIT Press.

OpenAi.com

Páez Pino, Adriana. (2023). *Inteligencia artificial: Abordando los sesgos para un futuro más justo.* Linkedin.

Piccinini, Gualtiero. (2004). *The First Computational Theory of Mind and Brain: A Close Look at Mcculloch and Pitts's 'Logical Calculus of Ideas Immanent in Nervous Activity.* Synthese, vol. 141, no. 2, pp. 175–215

Pilar López, Entrevista de investigación

RenAIssance Foundation (2020, February 28). *The Call for AI Ethics*. Rome Call for AI Ethics.

Searle, John (1980). *Chinese room argument*. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2023 Edition).

Steinhardt, Jacob. (2021) *Updates and Lessons from Ai Forecasting*. Bounded Regret,

Telkamp, Jake B., and Marc H. Anderson. (2022) *The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence.* Journal of business ethics 178.4: 961–976.

Valenzuela, Carolina González. (2023). *¿Qué Es Ani y Por Qué Se Le Conoce Como Inteligencia Artificial Débil o Limitada?* Computer Hoy.

Wright, Scott A., and Ainslie E. Schultz. (2018) *The Rising Tide of Artificial Intelligence and Business Automation: Developing an Ethical Framework*. Business horizons 61.6 (2018): 823–832.